Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole

Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, Stéphane Schneider

6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (JEP-TALN-RÉCITAL) [1]

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).

Volume 1 : Journées d'Études sur la Parole

Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, Stéphane Schneider (Éds.)

Nancy, France, 08-19 juin 2020

1. https://jep-taln2020.loria.fr/

Avec le soutien de

Message des présidents de l'AFCP et de l'ATALA

En ce printemps 2020, et les circonstances exceptionnelles qui l'accompagnent, c'est avec une émotion toute particulière que nous vous convions à la 6e édition conjointe des Journées d'Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Après une première édition commune en 2002 (à Nancy, déjà !), et une expérience renouvelée avec succès en 2004, c'est désormais tous les quatre ans (Avignon 2008, Grenoble 2012) que se répète cet événement commun, attendu de pied ferme par les membres des deux communautés scientifiques voisines.

Cette édition 2020 est exceptionnelle, puisque dans le cadre des mesures sanitaires liées à la pandémie mondiale de COVID-19 (confinement strict, puis déconfinement progressif), la conférence ne peut avoir lieu à Nancy comme initialement prévu, mais se déroule à distance, sous forme virtuelle, soutenue par les technologies de l'information et de la communication. Nous remercions ici chaleureusement les organisateurs, Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni et Sylvain Pogodalla, qui ont dû faire preuve de souplesse, d'inventivité, de détermination, de puissance de travail, et de tant d'autres qualités encore, afin de maintenir la conférence dans ces circonstances, en proposant un format inédit. Grâce aux différentes solutions mises en œuvre dans un délai court, la publication des communications scientifiques est assurée, structurée, et les échanges scientifiques sont favorisés, même à distance.

Bien entendu, nous regrettons tous que cette réunion JEP-TALN-RECITAL ne permette pas, comme ses prédécesseurs, de nouer ou renforcer les liens sociaux entre les différents membres de nos communautés respectives – chercheurs, jeunes et moins jeunes, académiques et industriels, professionnels et étudiants – autour d'une passionnante discussion scientifique ou d'un mémorable événement social... Notre conviction est qu'il est indispensable de maintenir à l'avenir de tels lieux d'échanges dans le domaine francophone, afin bien sûr de permettre aux jeunes diplômés de venir présenter leurs travaux et poser leurs questions sans la barrière de la langue, mais aussi de dynamiser nos communautés, de renforcer les échanges et les collaborations, et d'ouvrir la discussion autour des enjeux d'avenir, qui questionnent plus que jamais la place de la science et des scientifiques dans notre société.

Lors de la précédente édition, nous nous interrogions sur les phénomènes et tendances liés à l'apprentissage profond et sur leurs impacts sur les domaines de la Parole et du TAL. Force est de constater que l'engouement pour ces approches dans nos domaines a permis un retour sur le devant de la scène des domaines liés à l'Intelligence Artificielle, animant parfois un débat tant philosophique que technique sur la place de la machine dans la société, notamment à travers le questionnement sur la vie privée de l'utilisateur. Ces questionnements impactent tant la Parole que le TAL, d'une part sur la place de la gestion des données, d'autre part sur les modèles eux-mêmes. Malgré ces questionnements, nous constatons que les acquis et les expertises perdurent, et les nouvelles approches liées à l'apprentissage profond ont permis un rapprochement des domaines de la Parole et du TAL, sans les dénaturer, à la manière des conférences JEP-TALN-RECITAL qui créent un espace plus grand d'échange et d'enrichissement réciproques.

Nous terminons ces quelques mots d'ouverture en remerciant l'ensemble des personnes qui ont rendu possible cet événement qui restera, nous l'espérons, riche et passionnant, malgré les circonstances. L'ATALA et l'AFCP tiennent tout d'abord à réitérer leurs remerciements aux organisateurs des JEP, de TALN et de RECITAL, qui sont parvenus à maintenir le cap à travers vents et marées. Nos remerciements vont également à l'ensemble des membres des comités de programme, dont le travail et l'implication ont permis de garantir la qualité et la cohérence du programme finalement retenu. Un grand merci aux relecteurs pour le temps et le soin qu'ils ont dédiés à ce travail anonyme et indispensable. Ils se reflètent dans la qualité des soumissions que chacun pourra découvrir sur le site de la conférence.

En conclusion, cette 6e édition conjointe JEP-TALN-RECITAL est exceptionnelle parce qu'elle se tient dans un contexte de crise généralisée — crise sanitaire, économique, voire sociale et politique. Mais nous

formons le vœu qu'elle reste également dans les annales pour la qualité des échanges scientifiques qu'elle aura suscités, et pour le message envoyé à nos communautés scientifiques et à la société dans son ensemble, un message de détermination et de confiance en l'avenir, où la science et les nouvelles technologies restent au service de l'humain.

Véronique Delvaux, présidente de l'Association Francophone de la Communication Parlée
Christophe Servan, président de l'Association pour le Traitement Automatique des LAngues

Préface

En 2002, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisèrent conjointement leurs principales conférences afin de réunir en un seul lieu, à Nancy, les communautés du traitement automatique et de la description des langues écrites, parlées et signées.
En 2020, la sixième conférence commune revient à Nancy, après Fès (2004), Avignon (2008), Grenoble (2012) et Paris (2016). Elle est organisée par le LORIA (Laboratoire lorrain de recherche en informatique et ses applications, UMR 7503), l'ATILF (Analyse et traitement informatique de la langue française, UMR 7118) et l'INIST (Institut de l'information scientifique et technique) et regroupe :

— les 33es Journées d'Études sur la Parole (JEP),
— la 27^{e} conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 22^{e} Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL).

Les circonstances particulières liées à l'épidémie de Covid-19 en France et dans le monde ont conduit à une virtualisation de la conférence. Ainsi, malgré un rassemblement physique qui n'a pu avoir lieu, diffusions, présentations (au gré des auteurs) et discussions des articles acceptés ont lieu sur le site internet de la conférence. Les tutoriels, certains ateliers, et le salon de l'innovation qui accompagnent la conférence ont cependant dû être annulés, mais les ateliers suivants sont maintenus :

— Défi Fouille de Textes (DEFT 2020),
— Éthique et TRaitemeNt Automatique des Langues (ÉTeRNAL).

La conférence accueille également des conférencières et conférenciers invités dont les exposés sont diffusés sur le site : Dirk Hovy (université de Bocconi, Milan, Italie, invité ÉTeRNAL) ainsi que Marie-Jean Meurs (Université du Québec à Montréal, UQAM, Canada) et Hugo Cyr (Faculté de science politique et droit à l'Université du Québec à Montréaln UQAM, Canada). En raison des circonstance particulières, un exposé conjoint de Christine Meunier (Laboratoire Parole et Langage LPL, CNRS, Aix-en-Provence, France) et Christophe Stécoli (police technique et scientifique française) a dû être annulé et reporté à une journée spéciale en septembre 2020.
Ces actes regroupent les articles des conférences JEP (volume 1), TALN (volume 2), RÉCITAL (volume 3), les articles décrivant les démonstrations (volume 4), et les articles des ateliers DEFT (volume 5) et ÉTeRNAL (volume 6). Pour la première fois, un appel spécifique à résumés en français d'articles parus dans une sélection de conférences internationales en 2019 était également proposé (volume 4). Un appel spécifique apprenti·e·s chercheur·euse·s destiné aux étudiants de licence, de master, ou en première année de thèse a également été proposé, pour leur proposer des présentations courtes ou sous forme de poster de leurs projets.
Pour les JEP, 87 articles ont été soumis, parmi lesquels 74 ont été sélectionnés, soit un taux de sélection de 85%.
Pour TALN, 58 articles ont été soumis, parmi lesquels 37 ont été sélectionnés, soit un taux de sélection de 63%, dont 10 comme article longs (17% des soumissions) et 27 comme article courts dont 20 en présentation orale (34% des soumissions) et 7 en présentation poster (12% des soumissions).
Pour RÉCITAL, 22 articles ont été soumis, parmi lesquels 16 ont été sélectionnés, soit un taux de sélection de 73%.
Nous souhaitons vivement remercier toutes les personnes qui ont participé à ce travail de relecture et de sélection :

— l'ensemble des relecteurs (voir page xi),
— le comité de programme des JEP (voir page viii),
— le comité de programme de TALN (voir page ix),
— le comité de programme de RÉCITAL (voir page x).

Nous souhaitons également remercier nos sociétés savantes : l'AFCP, assurant la continuité des éditions successives des JEP, et l'ATALA, dont le CPerm (comité permanent) assure la continuité des éditions

successives de TALN.
Nous remercions le comité d'organisation et les nombreuses personnes qui ont assuré le soutien administratif et technique pour que cette conférence se déroule dans les meilleures conditions, et en particulier Yannick Parmentier pour son travail pour la diffusion de ces actes sur HAL et les différents sites d'archives ouvertes (anthologie ACL et talnarchives.atala.org/).
Nous remercions enfin tous les partenaires institutionnels et industriels qui nous ont fait confiance, en particulier l'université de Lorraine, le CNRS, l'Inria, le LORIA, l'ATILF, l'INIST, le master TAL de l'Institut des Sciences du Digital Management & Cognition (IDMC), le projet OLKI de l'initiative Lorraine Université d'Excellence (LUE), la Région Grand Est, *The Evaluations and Language resources Distribution Agency* (ELDA), le projet ANR PARSEME-FR, la délégation générale à la langue française et aux langues de France (DGLFLF), l'Association des Professionnels des Industries de la Langue (APIL) et les entreprises Synapse, Yseop et Orange.
Bonne conférence à toutes et à tous !

Les présidentes et présidents JEP : David Langlois et Slim Ouni
TALN : Chloé Braud et Sylvain Pogodalla
RÉCITAL : Christophe Benzitoun et Laurine Huber

Comités

Comité de programme des JEP

Martine Adda-Decker (Laboratoire de Phonétique et Phonologie, CNRS)
Jean-Francois Bonastre (LIA, Université d'Avignon)
Fethi Bougares (LIUM, Le Mans Université) Philippe Boula De Mareüil (LIMSI, CNRS)
Hervé Bredin (LIMSI, CNRS)
Olivier Crouzet (LLING, Université de Nantes)
Elisabeth Delais-Roussarie (LLING, Université de Nantes)
Véronique Delvaux (Laboratoire de Phonétique, IRSTL, Université de Mons)
Camille Fauth (LiLPa, Université de Strasbourg)
Emmanuel Ferragne (CLILLAC-ARP, Université de Paris)
Cecile Fougeron (Laboratoire de Phonétique et Phonologie, CNRS)
Corinne Fredouille (LIA, Université d'Avignon)
Alain Ghio (LPL, CNRS)
Camille Guinaudeau (LIMSI, Université Paris Sud)
Anne Guyot Talbot (CLILLAC-ARP, Université de Paris 7)
Bernard Harmegnies (Laboratoire de Phonétique, IRSTL, Université de Mons)
Nathalie Henrich Bernardoni (Gipsa-lab, CNRS)
Bassam Jabaian (LIA, Université d'Avignon)
David Langlois (LORIA, Université de Lorraine)
Yves Laprie (LORIA, CNRS)
Anthony Larcher (LIUM, Université du Maine)
Gwénolé Lecorvé (IRISA, Université de Rennes)
Benjamin Lecouteux (LIG, Université Grenoble Alpes)
Georges Linarès (LIA, Université d'Avignon)
Damien Lolive (IRISA, Université Rennes)
Julie Mauclair (IRIT)
Yohann Meynadier (LPL, Aix-Marseille Université)
Slim Ouni (LORIA, Université de Lorraine)
Thomas Pellegrini (IRIT, Université de Toulouse)
François Portet (LIG, Grenoble INP)
Fabian Santiago (Structures Formelles du Langage, Université de Paris 8)
Christophe Savariaux (Gipsa-lab, CNRS)
Nathalie Vallee (Gipsa-lab, Université Grenoble Alpes)
Ioana Vasilescu (LIMSI, CNRS)

Comités de programme TALN

Maxime Amblard (LORIA, Université de Lorraine)
Chloé Braud (IRIT, CNRS)
Caroline Brun (Naver Labs Europe)
Nathalie Camelin (LIUM, Université du Maine)
Marie Candito (Universtité Paris 7)
Vincent Claveau (IRISA, CNRS)
Chloé Clavel (Telecom-ParisTech)
Mathieu Constant (ATILF, CNRS, Université de Lorraine)
Pascal Denis (Inria)
Cécile Fabre (Université Toulouse 2)
Thomas François (Université catholique de Louvain)
Núria Gala (LPL, CNRS, Aix-Marseille Université)
Natalia Grabar (STL, CNRS, Université Lille 3)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIIE, Université Paris-Saclay)
Emmanuel Morin (LINA, Université de Nantes)
Sylvain Pogodalla (LORIA, Inria)
Solen Quiniou (LINA, Université de Nantes)
Corentin Ribeyre (Etermind)
Tim van de Cruys (IRIT, CNRS)
Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay)

Comité de programme RÉCITAL

Jean-Yves Antoine (Université François Rabelais de Tours)
Sonia Badene (Linagora, IRIT)
Frédéric Béchet (LIF, Aix Marseille Université)
Christophe Benzitoun (ATILF, Université de Lorraine)
Maria Boritchev (LORIA, Inria)
Léo Bouscarrat (EURA NOVA, Aix-Marseille Université)
Manon Cassier (INALCO, Paris)
Kevin Deturck (Viseo Technologies)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Kim Gerdes (sorbonne nouvelle)
Nicolas Hernandez (LINA, UMR 6241, CNRS, Université de Nantes)
Lydia-Mai Ho-Dac (CLLE-ERSS, Université Toulouse Jean Jaurès)
Laurine Huber (LORIA, Université de Lorraine)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
Joël Legrand (LORIA, Inria, CNRS)
Anne-Laure Ligozat (LIMSI, CNRS, ENSIIE, Université Paris-Saclay)
Pierre Ludmann (LORIA, Université de Lorraine)
Yann Mathet (Université de Caen)
Anne-Lyse Minard (IRISA, CNRS)
Sandrine Ollinger (ATILF, UMR 7118, CNRS)
Yannick Parmentier (LORIA, Université de Lorraine)
Justine Reynaud (LORIA, Université de Lorraine)
Stella Zevio (LIPN, Université de Paris 13)

Relectrices et relecteurs

Gilles Adda (LIMSI, CNRS) Salah Ait-Mokhtar (Naver Labs Europe)
Charlotte Alazard (Université Toulouse 2 Jean Jaurès)
Alexandre Allauzen (LIMSI-CNRS, Université Paris-Sud)
Pascal Amsili (Université Paris Diderot)
Pierre André Hallé (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Régine André-Obrecht (Université Paul Sabatier Toulouse III)
Jean-Yves Antoine (Université François Rabelais de Tours)
Nicolas Audibert (Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3)
Nelly Barbot (IRISA, Université de Rennes 1)
Claude Barras (LIMSI, CNRS)
Loïc Barrault (University of Sheffield)
Katarina Bartkova (ATILF, Université de Lorraine)
Frédéric Béchet (LIF, Aix Marseille Université)
Nathalie Bedoin (DDL, Université Lyon 2)
Patrice Bellot (LSIS, CNRS, Aix-Marseille Université)
Asma Ben Abacha (National Library of Medicine, National Institutes of Health)
Delphine Bernhard (LiLPa, Université de Strasbourg)
Roxane Bertrand (LPL, CNRS, Aix-Marseille Université)
Laurent Besacier (Laboratoire d'Informatique de Grenoble)
Yves Bestgen (F.R.S-FNRS et Université Catholique de Louvain)
Frédéric Bimbot (IRISA, CNRS)
Caroline Bogliotti (MODYCO, UMR 7114, CNRS, Université Paris Nanterre)
Anne Bonneau (LORIA, CNRS)
Stéphanie Borel (Université de Tours)
Féthi Bougarès (LIUM, Le Mans Université)
Leila Boutora (Laboratoire Parole et Langage, Aix Marseille Université)
Paul Caillon (LORIA, Université de Lorraine)
Mélanie Canault (DDL, Université Lyon 2)
Thierry Charnois (LIPN, CNRS, Université de Paris 13)
Chloé Clavel (Telecom-ParisTech)
Maximin Coavoux (Université Grenoble Alpes, CNRS)
Vincent Colotte (LORIA, Université de Lorraine)
Juan Manuel Coria (LIMSI, Université Paris-Saclay Paris 13)
Benoît Crabbé (Université Paris 7)
Lise Crevier Buchman (Laboratoire de Phonétique et Phonologie, CNRS, Hôpital Foch)
Béatrice Daille (LINA, Université de Nantes)
Géraldine Damnati (Orange Labs)
Dan Dediu (Dynamique du Langage, UMR5596, Université Lumière Lyon 2)
Joseph Di Martino (LORIA, Université de Lorraine)
Gaël Dias (Université Caen Normandie)
Amazouz Djegdjiga (LPP, Université Sorbonne Nouvelle – Paris 3)
Benjamin Elie (IMSIA, ENSTA ParisTech)
Iris Eshkol-Taravella (Université d'Orléans)
Emmanuelle Esperança-Rodier (GETALP, Université Grenoble Alpes)
Yannick Estève (LIA, Université d'Avignon)
Dominique Estival (Western Sydney University)
Olivier Ferret (CEA LIST)
Lionel Fontan (Archean Labs)
Karën Fort (Sorbonne Université)
Claire Gardent (LORIA, CNRS)
Eric Gaussier (LIG, Université Grenoble Alpes)
Cédric Gendrot (LPP, Université Sorbonne Nouvelle – Paris 3)
James German (Laboratoire Parole et Langage, Aix Marseille Université)
Cyril Goutte (National Research Council Canada)
Cyril Grouin (LIMSI, CNRS, Université Paris-Saclay)
Pierre André Hallé (LPP, Université Sorbonne Nouvelle – Paris 3)
Olivier Hamon (Syllabs)
Thierry Hamon (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
Bernard Harmegnies (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Nabil Hathout (CLLE, CNRS)
Amir Hazem (LS2N, Université de Nantes)
Nicolas Hernandez (LS2N, Université de Nantes)
Fabrice Hirsch (Praxiling, Université Paul Valéry Montpellier 3)
Thomas Hueber (GIPSA-lab, CNRS)
Kathy Huet (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)

Stéphane Huet (LIA, Université d'Avignon)
Mathilde Hutin (LIMSI, Université Paris-Saclay, CNRS, Université Sorbonne Paris Nord)
Irina Illina (LORIA, Université de Lorraine)
Christine Jacquin (LS2N Université de Nantes)
Adèle Jatteau (STL, UMR 8163, Université de Lille, CNRS)
Denis Jouvet (LORIA, Inria)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre)
Takeki Kamiyama (LPP, Université Paris 8 Vincennes-Saint-Denis)
Hannah King (CLILLAC-ARP, Université Paris Diderot)
Olivier Kraif (Université Grenoble Alpes)
Matthieu Labeau (Telecom Paris)
Mathieu Lafourcade (LIRMM, Université de Montpellier)
Mohamed Lahrouchi (SFL, UMR 7023, CNRS Université Paris 8)
Muriel Lalain (LPL, CNRS, Aix-Marseille Université)
Joseph Lark (Dictanova)
Thomas Lavergne (LIMSI, CNRS, Univ. Paris Sud, Université Paris Saclay)
Guillaume Le Berre (LORIA, Université de Lorraine)
Gwénolé Lecorvé (IRISA, Université de Rennes, CNRS)
Benjamin Lecouteux (Laboratoire Informatique de Grenoble)
Claire Lemaire (Université Grenoble Alpes)
Yves Lepage (Waseda University)
Joseph Le Roux (LIPN, Université de Paris 13)
Veronika Lux (ATILF, CNRS)
Paolo Mairano (STL, UMR 8163, Université de Lille)
Anna Marczyk (LPL, CNRS, Aix-Marseille Université)
Denis Maurel (Université François Rabelais de Tours)
Christine Meunier (LPL, CNRS, Aix-Marseille Université)
Alexis Michaud (LACITO, CNRS)
Richard Moot (LIRMM, CNRS)
Véronique Moriceau (LIMSI, CNRS)
Philippe Muller (IRIT, Université de Toulouse)
Alexis Nasr (LIF, Université de la Méditerranée)
Sylvain Navarro (CLLE-ERSS, CNRS)
Luka Nerima (Université de Genève)
Aurélie Névéol (LIMSI, CNRS, Université Paris-Saclay)
Jian-Yun Nie (Université de Montreal)
Damien Nouvel (INaLCO)
Nicolas Obin (IRCAM)
Yannick Parmentier (LORIA, Université de Lorraine)
Sebastian Peña Saldarriaga (Dictanova)
Marie Philippart de Foy (Université de Mons)
Myriam Piccaluga (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Claire Pillot-Loiseau (LPP, UMR 7018, CNRS, Université Sorbonne Nouvelle – Paris 3)
Serge Pinto (LPL, CNRS, Aix-Marseille Université)
Agnès Piquard (LORIA, CNRS, Université de Lorraine)
Thierry Poibeau (LaTTiCe, CNRS)
Alain Polguère (ATILF Université de Lorraine)
Laurent Prévot (LPL, CNRS, Aix-Marseille Université)
Jean-Philippe Prost (LIRMM, Université de Montpellier)
Christian Raymond (IRISA, INSA de Rennes)
Christian Retoré (LIRMM, Université de Montpellier)
Albert Rilliard (LIMSI, CNRS, Université Paris-Saclay)
Virginie Roland (Institut de Recherche en Sciences et Technologies du Langage, Université de Mons)
Sophie Rosset (LIMSI, CNRS, Université Paris-Saclay)
Véronique Sabadell (LPC, Aix Marseille Université)
Stéphane Schneider (INIST, CNRS)
Didier Schwab (Université Grenoble Alpes)
Pascale Sébillot (IRISA, INSA de Rennes)
Djamé Seddah (Almanach, Université Paris la Sorbonne)
Gilles Serasset (LIG, Université Grenoble Alpes)
Romain Serizel (LORIA, Université de Lorraine)
Kamel Smaïli (LORIA, Université de Lorraine)
Rudolph Sock (LiLPa, Université de Strasbourg)
Ludovic Tanguy (CLLE, CNRS)
Xavier Tannier (LIMICS, Sorbonne Université, INSERM)
Andon Tchechmedjiev (IMR, Mines Alès)
Juan-Manuel Torres-Moreno (LIA, Université d'Avignon)
Nicolas Turenne (LISIS, INRA)
Béatrice Vaxelaire (LiLPa, Université de Strasbourg)

Anne Vilain (GIPSA-lab, Université de Grenoble Alpes)
Coriandre Vilain (GIPSA-lab, Université de Grenoble Alpes)
Guillaume Wisniewski (LLF, Université de Paris)
Jane Wottawa (LIUM, Le Mans Université)
Yaru Wu (LPP, MoDyCo, Université Paris Nanterre)
Kossi Seto Yibokou (LiLPa, Université de Strasbourg)
François Yvon (LIMSI, CNRS, Université Paris-Sud)

Table des matières

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

'Il était une fois' les patterns prosodiques des contes de fée

Rim Abrougui[2], Katarina Bartkova[1,2]
(1) Atilf UL, 54000 Nancy, France
(2) Université de Lorraine, 54000 Nancy, France
{rim.abrougui;katarina.bartkova}
@univ-lorraine.fr

RÉSUMÉ

Nous étudions ici la différence des patterns prosodiques entre deux styles de lecture, un que nous appelons 'lecture littéraire neutre' et un style de 'lecture des contes'. Les données appartenant au style de 'lecture de contes' comportent deux sous-ensembles, des contes destinés aux jeunes enfants (0-6 ans) et des contes destinés aux enfants plus âgés et aux adultes. Les corpus ont été manuellement annotés avec des étiquettes sémantico-prosodiques exprimant des attitudes, des émotions et d'autres styles prosodiques. Une analyse détaillée des caractéristiques prosodiques nous a permis d'identifier les traits pertinents des patterns intonatifs des différentes étiquettes et des différents styles de lecture. Une quantification vectorielle, utilisant essentiellement des informations de F0, a été utilisée pour dégager les patterns prosodiques typiques correspondant aux différentes étiquettes. Une classification automatique basée sur des paramètres prosodiques a montré une bonne identification des étiquettes quand leur fréquence était suffisamment élevée pour obtenir une modélisation robuste.

ABSTRACT

'Once upon a time' prosodic patterns of fairy tales
Here, we study the difference in prosodic patterns between two reading styles, one that we call "neutral literary reading" and one corresponding to "storytelling reading". The data concerned with the "storytelling" style include two subsets, tales intended for young children (0-6 years) and tales intended for older children and adults. The corpora have been manually annotated with semantic-prosodic labels expressing attitudes, emotions and other prosodic styles. A detailed analysis of the prosodic characteristics allowed to identify the relevant features of the intonation patterns associated to the different labels and different reading styles. A vector quantization procedure, essentially using F0 values information, was used to identify the typical prosodic patterns of the different semantic-prosodic labels. Using a tree classifier, an automatic classification based on the prosodic parameters showed a good identification of the labels when their frequency was high enough to obtain a reliable modeling.

MOTS-CLÉS : style de parole, prosodie, étiquette sémantico-prosodique
KEYWORDS: speech style, prosody, prosodico-semantic label

Volume 1 : Journées d'Études sur la Parole, pages 1–9.
HAL : hal-02798508.

1 Introduction

Le style de voix peut être identifié en grande partie par la prosodie utilisée. Si l'étude des styles est souvent réalisée sur l'écrit, elle est aussi omniprésente dans l'analyse orale. Les variations phoniques dues aux situations ou aux individus constituent les objets primordiaux dans le domaine de la phonostylistique. À travers la variation codée des paramètres prosodiques, se réalisent les différents styles de voix reflétant des informations implicites dans la communication qui restent souvent inaccessibles dans le sens explicite d'un énoncé (Ackerman, 1981). La variation de ces paramètres contribue à la perception naturelle de la parole car elle permet d'identifier les attitudes et les émotions exprimées par le locuteur. L'analyse et la modélisation de ces paramètres demeurent un défi permanent pour les chercheurs afin d'introduire le naturel dans la synthèse automatique de la parole (Gelin et al., 2010, Doukhan, 2007, Cambell, & Mokhtari, 2003).

L'analyse acoustique des styles de voix a démontré que les paramètres prosodiques varient systématiquement en fonction de la situation de communication et que ce sont ces variations qui reflètent l'expressivité (Beller, 2009). Cependant l'expressivité couvre un domaine encore plus vaste puisqu'elle peut se définir comme un indicateur vocal d'un état émotionnel, d'un style de parole ou même d'une intention implicite (Granström, & House, 2005). Ainsi, certains styles partagent les mêmes comportements prosodiques, comme par exemple la ressemblance dans l'utilisation des patterns de F0 entre les émotions 'tristesse' et 'dégout' (Scherer, 2005, Bartkova et al., 2016). Cette complexité des paramètres peut expliquer la multiplication des études sur la prosodie en relation avec l'expressivité abordée souvent dans un but de modélisation de l'expressivité pour les technologies vocales (Chella et al., 2008, Burkhardt, 2011).

Dans notre étude, nous avons choisi de nous intéresser à la narration des contes de fée. Cela nous a apparu pertinent pour l'étude des styles de voix car le conte, qui est par définition un récit narratif rattaché primordialement à l'écrit, est tout d'abord oral par tradition, ce que témoigne le style de narration des conteurs. En effet, lorsqu'un conteur prend la parole pour lire une histoire, il exploite la variation de sa voix, et celle de sa prosodie, pour attirer l'attention de son public. De nombreuses études se sont intéressées au style des contes souvent dans le but de leur utilisation dans les technologies de la parole (Doukhan et al., 2011 ; Sarkar et al., 2014) ou pour analyser d'une façon détaillée les phénomènes prosodiques pertinents pour l'expression de ce style de parole (Delais-Roussarie & Yoo, 2014).

L'autre style de parole étudié est la lecture de texte littéraire que l'on peut considérer comme un style plus neutre. La lecture 'neutre' se définit comme la parole préparée, elle est donc construite, et tend vers l'écrit (Bazillon et al., 2008). En cela, la lecture 'neutre' est comme la lecture des contes puisque les deux sont des genres narratifs. Elle est définie en tant que 'neutre', car son expressivité demeure à un degré qui avoisine '0', (Tao, 2006, Inanoglu, 2009). C'est pour cela que nous n'enregistrons pas de variations importantes de ses patterns prosodiques. Nous avons analysé la lecture 'neutre' pour la comparer avec le style de lecture des contes en mettant en évidence les points communs et les points de différence entre ces deux styles.

Le but de notre étude est de comparer les patterns prosodiques des contes destinés pour les auditeurs jeunes (narration visant un public d'enfants de moins de 6 ans) avec les contes destinés à des auditeurs plus âgés et des adultes. Notre objectif est d'étudier si une différence de style existe et si elle existe, nous voudrions vérifier si elle est portée par la majorité des patterns intonatifs ou si c'est la fréquence d'occurrence de certains patterns qui caractériserait un style de voix. La deuxième

question que nous nous posons dans cette étude, concerne la possibilité d'utilisation de modèles de patterns intonatifs pour annoter des corpus de parole. En effet, il n'est pas facile pour un annotateur humain de faire abstraction de contenu sémantique et de se focaliser exclusivement sur les caractéristiques prosodiques de la parole. Or, cela serait tout à fait envisageable si une approche automatique était mise en place. Mais pour une telle approche une bonne identification automatique des patterns prosodiques s'avère indispensable.

2 Corpus de parole

Notre corpus est constitué de 3 styles de lecture : style conte de fée destiné aux jeunes auditeurs ('CP', pour contes pour les petits), style conte de fée destiné aux auditeurs plus âgés ('CG', pour contes pour les grands) et style de lecture que nous considérerons comme un style neutre de lecture de texte littéraire ('LN'), destiné aux adultes et adolescents. La quasi-totalité de nos données sonores ont été obtenues à partir du site de la littérature audio (http://www.litteratureaudio.com/).

2.1 Traitement prosodique du corpus

Des traitements manuels, semi-automatiques et automatiques ont été appliqués sur les données. Les données de parole ont été transcrites orthographiquement (avec l'outil Transcriber) et ont été segmentées automatiquement par le logiciel Astali (http://ortolang108.inist.fr/astali/). La segmentation automatique a été manuellement vérifiée et corrigée quand cela semblait nécessaire. Des paramètres de F0 et d'énergie ont été calculés avec le logiciel Aurora (ETSI, 2005). À partir de ces données, différentes caractéristiques prosodiques (pentes et niveau de F0, durée et énergie vocalique normalisées …) ont été calculées. Ces données prosodiques ont été utilisées pour segmenter automatiquement la base de données en groupes intonatifs (GI) en évaluant des marques prosodiques sur les syllabes finales des unités lexicales (Bartkova et al., 2012). Pour cette segmentation une vérification manuelle a été également réalisée.

Les GI ont été utilisés comme fenêtre d'observation des paramètres prosodiques devenant également des unités d'étiquetage. Chaque GI a été étiqueté (voir section 2.2), et a été représenté par 6 valeurs de F0, 3 valeurs correspondant aux valeurs de F0 sur la dernière voyelle (début, milieu et fin), une valeur correspondant à la première voyelle et la valeur la plus basse et la valeur la plus haute parmi les valeurs restantes, tout en respectant l'ordre temporel de ces valeurs. Si le groupe prosodique ne contenait pas assez de voyelles (4 au minimum pour obtenir les 6 valeurs de F0), alors une interpolation a été effectuée à partir des valeurs disponibles. La décision de représenter la dernière voyelle par 3 valeurs de F0 a été prise dans un souci de représenter correctement un pattern intonatif de forme complexe (circonflexe, concave, etc.) présent surtout dans le style des contes.

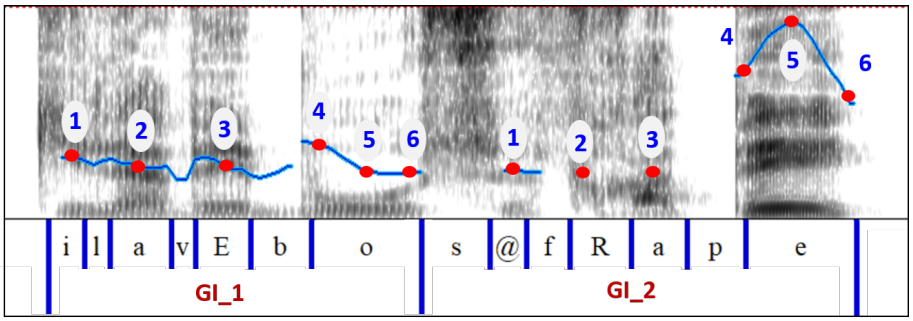


Figure 1 : Exemples de formes intonatives complexes
et des (6) valeurs de F0 retenues pour chaque GI

Le corpus CP comporte 6 contes lus par 4 locuteurs. Le corpus CG comporte 3 contes lus par 3 locuteurs et le corpus LN comporte 7 enregistrements lus par 6 locuteurs. Dans chaque corpus, les locuteurs-conteurs était féminins et masculins.

	Nb Mots	VA syl/sec	Durée (ms) V[-acc]	Durée (ms) V[+acc]	GS (nb syll)	GI (nb syll)	Pause (ms)	Étendue voc. (st)
CP	4605	4.6	79 (±33)	155 (±61)	4	2.4	627	24
CG	8379	4.9	75 (±32)	147 (±62)	5	2.6	517	26
LN	10995	5.5	68 (±23)	131 (±40)	6	3.0	639	16

Table 1 : Caractéristiques des corpus utilisés (voir explications dans le texte)

La table 1 indique la taille des corpus, ainsi que quelques informations sur l'organisation temporelle et les caractéristiques de la fréquence fondamentale de nos trois corpus. Le style le plus rapide (aussi bien par la vitesse d'articulation (VA) que par la durée des voyelles non-accentuées (V[-acc]) et accentuées (V[+acc]) est le style LN, et le plus lent est le style CP. C'est également le style CP qui a les groupes de souffles (GS) et les groupes intonatifs (GI) les plus courts (constitués en moyenne de 4 et de 2.4 syllabes respectivement). Le paramètre de l'étendue vocale du locuteur (en semi-tons dans le tableau) s'avère être très pertinent pour différencier les trois styles de parole. En effet elle est nettement plus large pour les styles de narration de contes (CP et CG) que pour le style LN. L'étendue vocale, pour un locuteur, correspond ici à la différence entre la valeur de F0 la plus élevée et la valeur la plus basse. Afin de pallier les éventuelles erreurs de détection de F0, les 2% des valeurs les plus basses et les plus élevées de F0 ont été écartées de l'estimation de l'étendue vocale.

2.2 Étiquetage sémantico-prosodique des corpus

Nos corpus ont été étiquetés manuellement par des étiquettes sémantico-prosodiques reflétant différentes attitudes et émotions ainsi que des réalisations prosodiques liées à un style donné (Akpossan-Confiac, 2007). L'étiquetage a été réalisé par 2 étiqueteurs experts qui ont attribué à chaque GI une étiquette. Les étiquettes ont été structurées sur trois niveaux (voir la figure 2) nous permettant d'introduire un lissage en cas d'étiquettes peu fréquentes dans les corpus. Les étiquettes sémantico-prosodiques représentaient des attitudes et des émotions positives et négatives (Vaudable, 2012) et des réalisations prosodiques plus neutres.

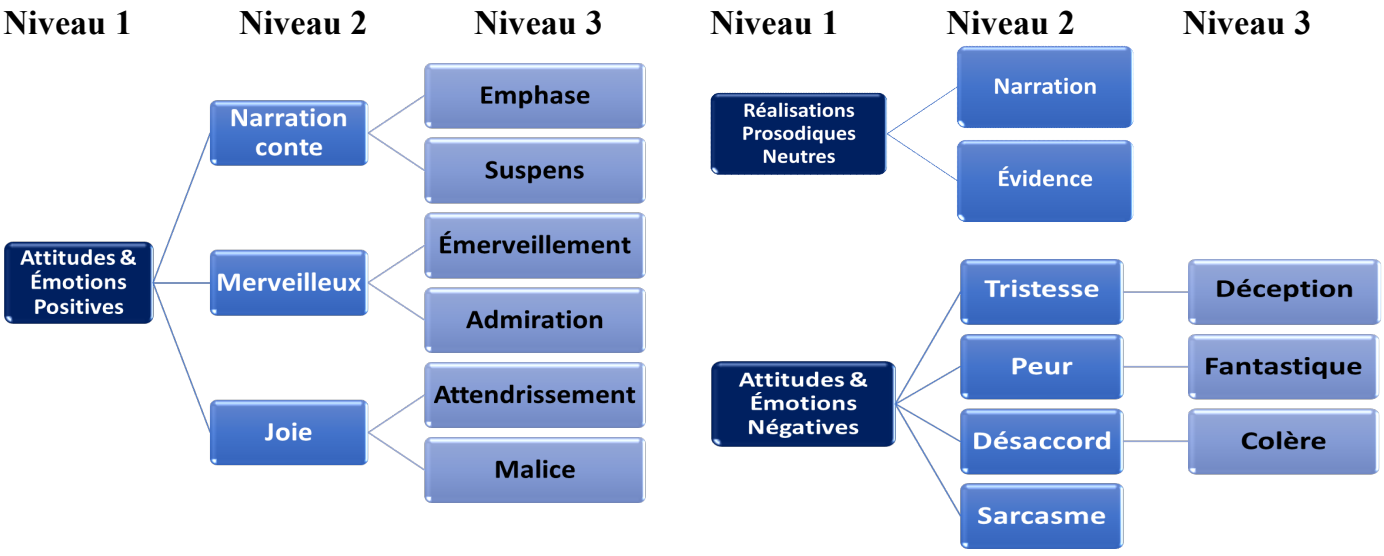


Figure 2 : Etiquettes sémantico-prosodiques

Lors de l'étiquettage sémantico-prosodique, l'attention des étiqueteurs devait se focaliser avant tout sur le pattern prosodique et non sur le contenu sémantique de l'énoncé. Les étiquettes sont distribuées d'une façon inégale dans les différents corpus. Dans le style LN, la très grande majorité (70%) des étiquettes correspondait au style 'narration'. Le nombre d'étiquettes de 'narration' avoisinait à peu près 50% dans les deux styles de contes (CP & CG). Pour les étiquettes restantes, la distribution était peu équilibrée (voir la Figure 3), les fréquences d'occurrences les plus élevées ont été observées pour l'étiquette '**suspense'** dans le style CG et '**évidence'** dans le style CP.

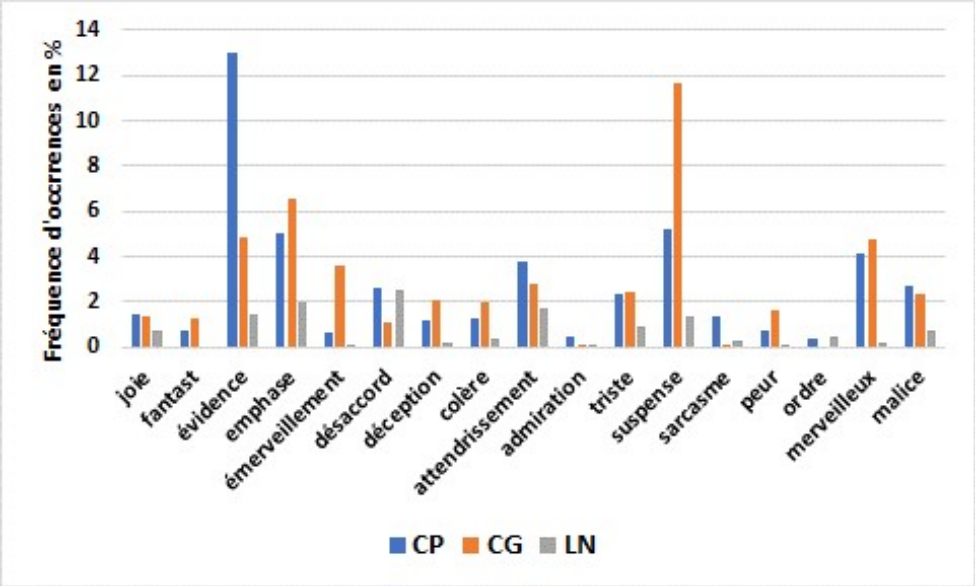


Figure 3 : Fréquence d'occurence des étiquettes sémantico-prosodiques
(l'étiquette « narration », n'est pas représentée sur la figure)

Pour évaluer le niveau de l'accord inter-annotateur, un test sur un sous-ensemble réduit de données représentant chaque style a été mené avec 6 étiqueteurs non entrainés (étudiants en linguistique du niveau master). Le test du Kappa a été réalisé à partir de ces différentes annotations et les résultats ont montré un accord inter annotateur faible (0.3) pour l'ensemble des étiquettes posées. Pour les 3 étiquettes ('*emphase', 'désaccord' et 'narration'*) l'accord Kappa était modéré (>0.4). Ce test montre un accord d'étiquetage acceptable pour un groupe non-entrainé mais spécialiste du langage et laisse penser qu'un accord plus élevé existe entre des annotateurs entraînés à la tâche ce qui était le cas lors de l'étiquetage de nos corpus.

3 Patterns intonatifs

Afin d'identifier les patterns intonatifs typiques correspondant à nos étiquettes sémantico-prosodiques, nous avons utilisé la technique de la quantification vectorielle pour classer tous les patterns intonatifs associés aux étiquettes. Avant d'appliquer la quantification vectorielle, nous avons quantifié les valeurs de F0. Pour cela nous avons estimé une étendue vocale théorique des locuteurs, comme étant supérieure d'une octave et inférieure d'une demi-octave aux valeurs médianes de F0 de chaque locuteur (De Looze, Hirst, 2014). L'étendue vocale théorique a été divisée en 10 niveaux permettant d'exprimer les valeurs de F0 par un niveau tonal de 1 à 10 à l'intérieur de l'étendue vocale théorique. Lorsqu'une valeur de F0 sort de l'étendue vocale théorique, sa valeur quantifiée (son niveau tonal) sera supérieure à 10 ou inférieure à 1.

3.1 Patterns intonatifs représentatifs

Pour représenter chaque étiquette sémantico-prosodique par son pattern intonatif, nous avons sélectionné le centroïde de la classe obtenue par la quantification vectorielle qui regroupait un grand nombre d'éléments et qui présentait un écart-type faible, c'est-à-dire, dont la variation des paramètres des éléments regroupés était faible. Ainsi nous avons obtenu un pattern typique, représentatif par étiquette.

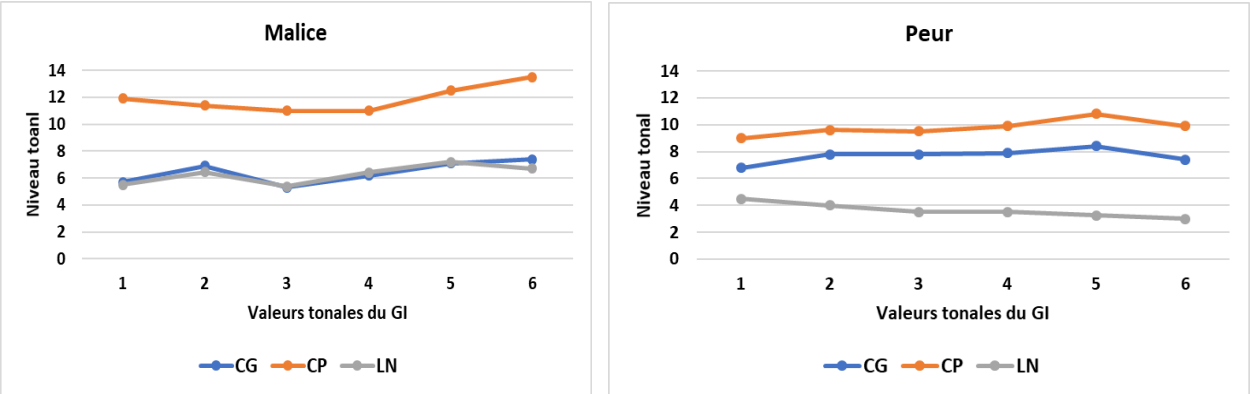


Figure 4 : Exemple de patterns intonatifs pour deux étiquettes '*Peur*' et '*Malice*' et pour les 3 styles (1 Niveau tonal = 1,8 semi-tons)

La figure 4 présente les patterns intonatifs typiques obtenus avec la quantification vectorielle pour deux étiquettes sémantico-prosodiques. Ces exemples montrent une similitude dans l'expression de la '*peur*' dans le style des contes (les deux patterns contiennent des valeurs de F0 d'un niveau élevé – valeurs comprises entre le niveau 7 et 11) alors que dans le style LN l'expression de la '*peur*' est plus 'contenue', exprimée sur un niveau tonal plus bas (entre les niveaux 3 et 4.5). Quant à la '*malice*', cette attitude est exprimée avec le même pattern intonatif dans les styles CG et LN et sur un niveau très élevé (plus haut que l'étendue vocale estimée, entre 11 et 13,5) pour le style CP.

Émotions & attitudes positives									
	Narr.	**Malice**	**Merv.**	**Admir.**	**Attendr.**	**Emphase**	**Évidence**	**Joie**	**Susp.**
CP	**2 ,7**	**11,9**	1,4	**7,3**	**1,4**	**7,0**	**3,1**	**8,5**	**1,8**
CG	4,9	6,4	1,2	5,7	1,4	5,6	1,8	6,9	7,4
LN	3,6	6,3	2,5	4,2	3,0	4,3	2,8	6,9	4,6

Émotions & attitudes négatives					
	Peur	**Tristesse**	**Colère**	**Déception**	**Désaccord**
CP	**9,8**	1,9	**7,3**	**2,1**	**8,0**
CG	7,7	1,0	6,6	5,2	2,0
LN	3,6	2,7	4,4	3,1	5,0

Table 2 : Niveau tonal moyen des étiquettes sémantico-prosodiques
(3,3 correspond à la valeur F0 médiane du locuteur)

La variation des niveaux de F0 à l'intérieur d'un même pattern (sur un même GI), n'était pas très importante – les valeurs variaient modérément d'une voyelle à l'autre, par conséquent, il est permis de considérer que les F0 des patterns peuvent être caractérisés par leur niveau intonatif moyen. Pour pouvoir comparer d'une façon plus synthétique les patterns intonatifs correspondant aux différentes étiquettes des trois corpus étudiés, nous les représentons dans la table 2, par leur valeur moyenne,

calculée sur les patterns représentatifs. Uniquement les étiquettes qui ont été observées dans les trois corpus, sont indiquées dans le tableau et seront analysées ci-après.

Les valeurs des tonalités moyennes des différents patterns représentatifs nous montrent que les patterns du style CP ont le niveau tonal le plus extrême (le plus haut – couleur rouge dans la table 2 ou le plus bas – couleur bleue dans la table 2) dans presque 86% des cas. Quant au style LN, le niveau tonal de ses patterns intonatifs se situe majoritairement (dans 80% des cas) sur un niveau bas. Le style CG reste dans la plupart des cas entre les niveaux tonals de ces deux styles, pouvant être considéré comme un style 'transitoire' entre le style CP et le style LN.

Si conformément à notre arborescence un regroupement plus global des étiquettes au niveau 1 est effectué (émotions & attitudes positives versus émotions & attitudes négatives) alors nous obtenons pour les patterns intonatifs positifs un niveau tonal moyen de 10.5 pour le style CP, un niveau de 6,2 pour le style CG et finalement un niveau tonal moyen de 4.4 pour le style LN. Pour les patterns intonatifs négatifs, nous obtenons de nouveau le niveau tonal le plus haut, 7,9, pour le style CP, le niveau tonal le plus bas, 2.3, pour le style LN, et de nouveau un niveau tonal intermédiaire de 6,7 pour le style CG.

3.2 Classification automatique

L'étiquetage sémantico-prosodique des données de parole est fastidieux et comme indiqué dans la section 2.2, l'accord inter-annotateur pour les annotateurs non-experts n'est pas très élevé car la tâche reste très complexe. Par conséquent, l'utilisation d'une annotation automatique, basée essentiellement sur les indices prosodiques, serait très utile. Pour tester la faisabilité et la fiabilité d'une telle approche, nous avons réalisé une classification automatique de nos étiquettes afin d'évaluer si une identification automatique des étiquettes à partir des patterns intonatifs et de la durée quantifiée de la voyelle accentuée (voyelle finale) était envisageable. Pour la classification automatique, nous avons utilisé l'arbre de décision J48, disponible dans le toolkit weka (https://sourceforge.net/projects/weka/). Pour les tests, nous avons, d'une part, regroupé toutes les étiquettes de nos différents corpus ('Tous corpus'), et d'autre part, nous avons également classifié les étiquettes de chacun des trois corpus séparément. Comme déjà mentionné, notre corpus LN est déséquilibré car la majorité des étiquettes de ce corpus correspond à la '*narration'*. Pour pallier au moins en partie ce biais, nous n'avons gardé qu'un cinquième de ces étiquettes dans les tests de classification. Les tests ont été réalisés pour les étiquettes se trouvant au niveau 3 (feuilles) et au niveau 2 (nœuds immédiatement supérieurs) de notre arborescence d'étiquettes (cf. figure 2).

	Tous corpus	**CP**	**CG**	**LN**
Niveau 3	45%	27%	33%	49%
Niveau 2	50%	43%	38%	55%

Table 3 : Taux d'identification correcte des étiquettes sémantico-prosodiques

Les faibles résultats du 'Niveau 3' sont imputables au nombre réduit d'exemples pour certaines étiquettes, sauf l'étiquette '*narration'*. Quand un lissage est opéré sur les étiquettes ('Niveau 2') on obtient des résultats légèrement (LN) ou significativement (CP, CG) meilleurs pour les différents corpus

Il nous a semblé intéressant de tester si le pattern intonatif de l'étiquette 'narration', très fortement présente dans nos corpus, est identifiable comme appartenant à un style de parole particulier, c'est-à-dire pouvant être identifié comme '*narration-CP'*, '*narration-CG'* ou '*narration-LN'*. Nous avons

effectué un test de détection uniquement sur cette étiquette en sélectionnant un sous-ensemble du corpus LN pour éviter le biais de la surreprésentation des étiquettes appartenant à ce corpus. L'identification de l'appartenance du pattern de *narration* à un corpus particulier a atteint 65%. Les résultats plus détaillés sont dans le tableau de confusion de la table 4.

	CP	**CG**	**LN**
CP	**40%**	**44%**	16%
CG	22%	**58%**	20%
LN	6%	14%	**80%**

Table 4 : Classification de l'étiquette '*narration'*

Comme cela apparait dans la table 4, les étiquettes '*narration'* sont fortement dépendantes des corpus. Les étiquettes '*narration-CP'* ont été souvent confondues avec les étiquettes '*narration-CG'*. Les deux corpus étant des contes, leurs différences prosodiques sont probablement exprimées par d'autres étiquettes que celle de '*narration'*. Quand des confusions existent dans l'identification de l'étiquette '*narration-CG',* elles se font aussi bien avec les étiquettes '*narration-CP'* qu'avec les étiquettes *'narration-LN'*, suggérant ainsi un positionnement intermédiaire de ces étiquettes entre ces deux styles, CP et LN. Finalement, l'étiquette '*narration-LN'* est très bien identifiée comme appartement au style de lecture et uniquement une très faible confusion (6%) existe avec les étiquettes *'narration-CP'* et une confusion très légèrement plus élevée avec les étiquettes *'narration-CG'* (14%).

4 Conclusion

Le but de cette étude était d'étudier 3 styles de parole lue, deux appartenant à la lecture de contes et un au style de lecture neutre de textes littéraires. Le corpus de contes contenait des contes destinés aux enfants jeunes et des contes destinés aux enfants moins jeunes et aux adultes. Une différence des patterns intonatifs a été observée parmi ces 3 styles, le style CP étant le plus extrême et le style LN est le plus neutre. Il apparaît de cette étude, que les styles de lecture des contes et de lecture neutre n'ont pas de socle stylistico-prosodique commun important, ainsi, par exemple, même pour l'étiquette '*narration',* très fréquente dans chaque style, les caractéristiques prosodiques sont fortement dépendantes de chaque style. Par conséquent, le style de parole doit être considéré comme un phénomène plus global que simplement l'apparition plus au moins fréquente d'étiquettes exprimant des attitudes ou des émotions.

Nous voulions également tester si l'identification automatique des patterns prosodiques représentant des étiquettes sémantico-prosodiques était suffisamment fiable, afin de les utiliser dans l'avenir pour annoter des corpus de différents styles de voix. En effet, une telle annotation bien que nécessaire, reste chronophage et fastidieuse. De plus, un annotateur humain a souvent du mal à faire abstraction du contenu sémantique de la parole et à se focaliser uniquement sur les patterns intonatifs. Cependant, pour développer une approche d'étiquetage automatique performante, il faudrait disposer de corpus plus larges et représentatifs des différentes étiquettes ; en effet, le nombre limité d'occurrences de certaines étiquettes dans nos corpus n'a pas permis d'obtenir des modèles prosodiques suffisamment fiables pour une classification robuste.

Références

ACKERMAN B.P., (1981). Young Children's Understanding of a Speaker's Intentional Use of a False Utterance, *Developmental Psychology, 17, (pp. 472-480)*

AKPOSSAN-CONFIAC, J., & DELUMEAU, F. (2007). Comment la prosodie donne du sens aux interjections?, *Interfaces discours-prosodie : actes du 2ème Symposium international IDP07 & Colloque Charles Bally, Université de Genève,* (pp. 335-347)

BARTKOVA K., DELAIS-ROUSSARIE E., & SANTIAGO VARGAS F., (2012). PROSOTRAN : Un Système d'annotation Symbolique Des Faits Prosodiques Pour Les Données Non-Standards. *In Proceedings of the Joint Conference JEP-TALN-RECITAL*, *Volume 1: JEP, Grenoble, France: ATALA/AFCP*, *(pp. 601– 608)* https://www.aclweb.org/anthology/F12-1076

BARTKOVA K., JOUVET D., DELAIS-ROUSSARIE E., (2016) Prosodic Parameters and Prosodic Structures of French Emotional Data , 8th *Speech Prosody*, Boston, United States

BAZILLON T., JOUSSE V., BÉCHET F., ESTÈVE Y., LINARÈS G. et LUZZATI D., (2008) La parole spontanée : transcription et traitement, *In Revue Traitement Automatique des Langues (TAL), volume 49, (pp. 47–67)*

BELLER, G. (2009). *Analyse et modèle génératif de l'expressivité. Application à la parole et à l'interprétation musicale*, Thèse de doctorat, Université Paris VI, IRCAM.

BURKHARDT F. (2011). An Affective Spoken Storyteller, In *In Proceedings Interspeech*, (*pp. 3305-3306) https://www.isca-speech.org/archive/interspeech_2011/i11_3305.html*

CAMPBELL, N., & MOKHTARI, P. (2003). Voice Quality: the 4th Prosodic Dimension, *15th ICPhS, (pp. 2417-2420)*

CHELLA A., BARONE R.E., PILATO G., SORBELLO R. (2008). An Emotional Storyteller Robot. *In AAAI Spring Symposium on Emotion, Personality and Social Behavior*, March 26-28, Stanford University, Stanford

DELAIS-ROUSSARIE E., YOO H., (2014). Rythme et synthèse de la parole : études comparées des patrons rythmiques de différents genres. *Nouveaux Cahiers de Linguistique Française* 31.pp. 237-247 http://www.llf.cnrs.fr/fr/node/4927

DE LOOZE C., HIRST D., (2014). The OMe (Octave-Median) scale: A natural scale for speech melody, In 7th *Speech Prosody*, 10.21437/SpeechProsody.2014-170

DOUKHAN, D. (2007). *Synthèse de parole expressive au-delà du niveau de la phrase : le cas du conte pour enfant: conception et analyse de corpus de contes pour la synthèse de parole expressive*, Thèse de doctorat, Université Paris 11.

DOUKHAN D, RILLIARD A, ROSSET S, ADDA-DECKER M, D'ALESSANDRO C., (2011). Prosodic Analysis of a Corpus of Tales, *In Proceedings INTERSPEECH*

ETSI ES 202 212 V1.1.1, STQ (2005). Distributed speech recognition; Extended advanced front-end feature extraction.

GELIN R., D'ALESSANDRO C., LE Q., DEROO O., DOUKHAN D., MARTIN J., PELACHAUD C., RILLIARD A., and ROSSET S., (2010). Towards a storytelling humanoid robot. In *AAAI Fall Symposium Series on Dialog with Robots*, *(pp 137-138)*

GRANSTRÖM, B., & HOUSE, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication, 46(3-4), (pp. 473-484)*

INANOGLU, Z., & YOUNG, S. (2009). Data-driven emotion conversion in spoken English. *Speech Communication*, *vol. 51, no. 3, (pp. 268–283)*

SARKAR P, HAQUE A, KUMAR DUTTA A, REDDY M G., HARIKRISHNA D M, PRASENJIT D., RASHMI V., NARENDRA N P, SUNIL Kr. S B, JAINATH YADAV K., RAO S., (2014). Designing Prosody Rule-set for Converting Neutral TTS Speech to storytelling style speech for Indian Languages: Bengali, Hindi and Telugu *In* 2014 Seventh International Conference on Contemporary Computing (IC3)

SCHERER. K.R. (2005) What are emotions? And how can they be measured*? Social Science Information*, vol. 44(4), *(pp 695-729)*

TAO, J., KANG, Y., & LI, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, *(pp.1145-1154)*

VAUDABLE Ch., & DEVILLERS L. (2012). Negative emotions detection as an indicator of dialogs quality in call centers, *In Proceedings 37th ICASSP*, Kyoto, Japan,

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Production de la continuation du français par des apprenants japonophones : gestion de la F0 et de la durée

Rachel Albar
Laboratoire de Linguistique Formelle, UMR 7110, Université de Paris, 5 Rue Thomas Mann, 75013, Paris, France
rachel.albar@univ-paris-diderot.fr

RÉSUMÉ

Dans cette étude, nous nous intéressons à la réalisation des contours de continuation en français, par des apprenants japonophones en parole semi-spontanée. Pour ce faire, nous avons analysé des productions obtenues à partir de description d'images en prenant en compte le niveau d'apprentissage et différentes positions prosodiques. Les résultats montrent une bonne gestion de la fréquence fondamentale par les apprenants. En effet, ils produisent des montées prosodiques aux frontières de AP et IP et sont capables de produire des montées plus importantes aux frontières de IP. Cependant, la gestion du paramètre de durée est moins homogène. Les résultats montrent que la durée n'est pas un paramètre acoustique robuste utilisé pour produire la continuation. Ces résultats suggèrent que les deux paramètres acoustiques de durée et de F0 ne peuvent pas être mis au même niveau quant à l'analyse de la continuation.

ABSTRACT

Production of French continuation by Japanese learners : F0 and duration

This study investigates the realisation of continuation contours at French prosodic boundaries by Japanese learners in semi-spontaneous speech. We thus analysed the production of image descriptions performed by Japanese learners. Our results show that learners produce rising continuation contours at both IP and AP levels and are able to produce a greater F0 rise at IP boundaries, no matter their proficiency level. However, duration is a parameter that presents lots of variability. It suggests that the two acoustic parameters calculated with duration and F0 should be considered separately when analysing continuation.

MOTS-CLÉS : Prosodie, acquisition L2, contours continuatifs, description d'images.

KEYWORDS: Prosody, L2 acquisition, continuation contours, image description.

1 Introduction

En français, la continuation joue en premier lieu un rôle important dans la structuration prosodique ; les contours de continuation sont ceux qui apparaissent à l'intérieur d'un énoncé, en contribuant ainsi à l'organisation interne d'une phrase syntaxique et établissant la cohérence de l'énoncé. Delattre (Delattre, 1966) l'oppose à la finalité. Il rajoute une différence supplémentaire en distinguant une « continuation majeure » et une « continuation mineure », indiquant ainsi qu'il peut exister différents types de continuation à l'intérieur d'une même phrase. Cependant, la continuation peut jouer également un rôle d'un point de vue sémantique ou discursif, puisqu'elle peut apparaitre en position finale

Volume 1 : Journées d'Études sur la Parole, pages 10–18.
HAL : hal-02798509.

de phrase, seulement si elle garde une valeur discursive (Portes & Bertrand, 2005). Ces dernières ont également montré que dans cette position on trouve principalement une continuation majeure lui donnant une valeur conversationnelle discursive importante. De plus, l'énumération est considérée comme une sous-catégorie de la continuation. Dans les modèles intonatifs du français, la continuation peut être réalisée aux deux (voire trois) niveaux d'analyse, les niveaux AP, IP (Jun & Fougeron, 2000) correspondant aux niveaux accentuel et intonatif, et le niveau ip, ou niveau intermédiaire (Michelas, 2011). Les attributs acoustiques sont différents à chacun de ces niveaux puisqu'au niveau intonatif, les contours de continuation seront produits avec une montée mélodique et un allongement de la dernière syllabe plus importants aux niveaux supérieurs. Bien que ce niveau prosodique reste encore controversé (Di Cristo, 2016), Michelas (Michelas, 2011) propose qu'au niveau de l'ip, certains paramètres prosodiques tels qu'une différence d'allongement et un rehaussement mélodique permettent de le distinguer des autres niveaux.

La question du nombre de niveaux prosodiques devient cruciale lorsqu'on s'intéresse à l'acquisition de la prosodie par des apprenants, puisque ces derniers doivent gérer deux, voire trois réalisations prosodiques différentes, dépendantes d'un niveau prosodique. Cette gestion devient d'autant plus difficile pour des apprenants japonais dont la structure prosodique de leur L1 est différente de celle du français. En effet, le japonais est décrit comme une langue à accent tonal qui, contrairement au français, n'associe pas des contours montant à la fin des frontières prosodiques. La structure prosodique du japonais possède deux niveaux, le niveau de syntagme accentuel portant l'accent tonal (se terminant généralement par un ton bas final) tandis que le niveau de syntagme intonatif est associé au downstep (Venditti *et al.*, 1998; Beckman & Pierrehumbert, 1986).

Cependant, si cette description prosodique est attestée en lecture, des mouvements mélodiques aux frontières prosodiques sont observés en parole spontanée. Il est donc fréquent, même si ce n'est pas obligatoire, que les japonophones produisent des contours montants (L%H%) ou montant-descendants (L%HL%) en frontière de IP. Les contours montant-descendants correspondent à une montée de F0 suivie d'une forte descente sur la dernière more, et ils sont généralement associés au maintien du tour de parole (Koiso *et al.*, 1998). La situation des contours montants, quant à elle, est plus complexe puisque selon la même étude de Koiso, ils sont plutôt associés à une interruption du tour de parole qu'à son maintien. Or, ces contours sont souvent considérés comme interchangeables : leur choix serait induit par le style de parole (le contour L%H% serait corrélé à un style plus formel, Maekawa 2009), et le contour montant L%H% pourrait être une variante "tronquée" du contour L%HL% (Igarashi, 2015). Yoneyama (Yoneyama *et al.*, 2003), dans son étude basée sur des monologues, recense un pourcentage comparable de contours montants et de montant-descendants (24,8 et 32,5%) lorsque la force de frontière de discours est faible, et donc que ces contours ont une fonction de continuation. Outre l'utilisation de contours montants pour indiquer une continuation ou le maintien du tour de parole, les japonophones produiraient également un allongement significatif du dernier phonème (Koiso *et al.*, 1998) (généralement une voyelle en japonais, car seule la nasale peut apparaître en coda) ou syllabe (Yoneyama *et al.*, 2003).

Dans cet article, nous nous proposons d'analyser la production des contours de continuation du français par des apprenants japonophones en parole semi-spontanée, dans une tâche de description d'image. Nous cherchons à analyser la gestion des contours montants des apprenants japonophones, à la fois au niveau phrastique que discursif, et aux différents niveaux d'analyse (IP et AP). Deuxièmement, nous cherchons à déterminer quels sont les critères prosodiques (F0 et durée) utilisés lors de la production de la continuation en français. Enfin, nous cherchons à voir si le niveau d'apprentissage influe sur la production de cette continuation.

2 Corpus et méthodologie

2.1 Corpus

Le corpus que nous avons utilisé pour nos enregistrements est le corpus COREIL (Delais-Roussarie & Yoo, 2011), conçu pour recueillir les productions d'apprenants de différentes L1 et permettre l'étude des systèmes d'acquisition de la prosodie chez les apprenants. Ce corpus comporte différentes tâches telles que des lectures de texte et de dialogues, de la parole spontanée et des descriptions d'images. Dans le cadre de cette étude, nous avons choisi d'analyser cette dernière tâche. Quatre images ont été sélectionnées, trois représentant des scènes de vie quotidienne et un tableau de Van Gogh.

2.2 Participants

Nous avons enregistré 17 apprenants japonophones (trois hommes et quatorze femmes) à Tokyo ainsi que 4 francophones natifs à Paris (deux hommes et deux femmes). Nous avons ensuite classé les participants en trois groupes d'apprentissage suivant les cours où ils étaient inscrits à l'université et leur durée d'apprentissage du français : un groupe de niveau débutant (A1-A2, N=2), un groupe de niveau intermédiaire (A2-B1, N=10) ainsi qu'un groupe de niveau intermédiaire plus avancé (B1-B2, N=5). Les locuteurs natifs servent de groupe contrôle.

2.3 Méthodologie

Les participants à l'expérience ont été enregistrés dans une chambre sourde à l'aide d'un micro-casque Shure WH20XLR et d'une carte son externe Roland Quad Capture. Les images ont été présentées une par une sous format papier et nous avons donné comme consigne aux locuteurs de les décrire en français aussi longuement que possible. Les apprenants ont produit en moyenne 170 mots, contre 514 pour les natifs. Les enregistrements ont été segmentés, phonétisés et alignés automatiquement en utilisant SPPAS (Bigi, 2015). Cette segmentation a ensuite été revue manuellement dans Praat (Boersma, 2002). Nous avons ensuite rajouté deux tires ; la première contient une annotation ToBI des énoncés et la deuxième contient un découpage des différents niveaux prosodiques. Nous nous sommes basés sur les critères d'identification des contours montants donnés dans Portes et al (Portes *et al.*, 2007). Nous avons donc distingué les syntagmes intonatifs (IP) terminaux et non-terminaux ainsi que les énumérations. Pour la présente analyse, nous avons choisi de regrouper deux niveaux : le niveau 2 correspondant aux frontières de AP (frontières mineures), et le niveau 3 correspondant aux frontières de IP en position finale d'énoncé (terminaux) et d'énumération. Les voyelles restantes non porteuses d'un accent H ont automatiquement été notées comme de niveau 0 soit non accentuées. Nous avons ainsi pu analyser en moyenne 24 positions AP et 20 positions IP par apprenant contre respectivement 63 et 50 pour les francophones. Les mesures acoustiques de la durée des voyelles ainsi que de F0 à différents pourcentages de la voyelle ont été extraites automatiquement. La différence de F0 entre la voyelle accentuée et la voyelle de la syllabe précédente a également été calculée (valeurs à 75% de la voyelle) et les contours ont été classifiés comme montant lorsqu'ils étaient audibles et que cet écart était supérieur à 1 demi-ton. Les hésitations produites par les locuteurs ont également été annotées ainsi que la présence de voix craquée afin de les exclure respectivement de l'analyse de la durée et de la F0. Les données ont ensuite été traitées avec le logiciel R (R Core Team, 2017) pour l'analyse statistique.

3 Résultats

3.1 Analyse des contours selon le niveau prosodique

Les résultats montrent des pourcentages de contours montants aux frontières de AP et de IP globalement très proches entre les natifs et les apprenants (Figure 1). Nous avons construit un modèle linéaire généralisé mixte (GLMM) avec comme variable dépendante le contour prosodique observé (montant ou non, codé 1 ou 0), une interaction entre le niveau de langue et la frontière prosodique comme effets fixes et le locuteur comme effet aléatoire. Ce modèle nous a confirmé que tous les groupes de langue distinguent bien les deux frontières en produisant significativement plus de montées prosodiques aux frontières de IP que AP (groupe B1-B2 : $p = .005$**, autres groupes : $p < .001$***).

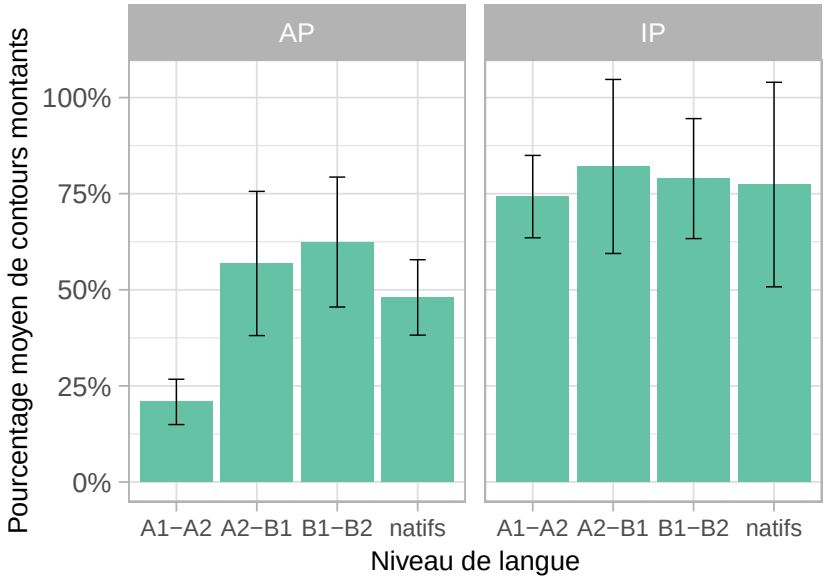


FIGURE 1 – Pourcentage moyen de contours montants aux frontières de AP et de IP selon le niveau de langue. Les barres représentent l'écart-type.

Toutefois, il est important de noter que l'on observe une grande variabilité entre les locuteurs à la fois chez les apprenants et les francophones, notamment aux frontières de IP. En effet, même si l'on observe chez trois des francophones un pourcentage de contours montants élevé allant de 89 à 92%, la quatrième locutrice (FR2) a une stratégie très différente puisqu'elle produit seulement 37,5% de montées prosodiques. Concernant les apprenants, ils produisent aussi une majorité de montées en frontières de IP, sur un intervalle plus étendu que les natifs allant de 63 à 100%, mais l'un d'entre eux (JP9, groupe A2-B1) ne produit que très peu de contours montants (25%).

3.2 Analyse des contours mélodiques montants

Les résultats de l'analyse de F0 des contours montants (Figure 2) montrent que les frontières de IP sont généralement associées à un écart de F0 plus important qu'aux frontières de AP, que ce soit chez les apprenants ou les natifs. Afin de vérifier ces observations, nous avons construit un modèle

linéaire mixte avec comme variable dépendante l'écart de F0 en demi-ton, le niveau de langue et la frontière prosodique comme effets fixes et le locuteur ainsi que la nature de la voyelle accentuée comme effets aléatoires. L'interaction entre le niveau de langue et les frontières prosodiques s'est révélée significative lors du test ANOVA ($\chi^2(3) = 27.39$, p <.001***). L'analyse des contrastes de cette interaction avec le package *emmeans* nous a montré que la différence d'écart de F0 entre les frontières AP et IP était significative pour tous les groupes de langue (p <.001***) à l'exception du groupe A1-A2 (p =.55), mais il faut noter que ce groupe étant seulement composé de deux locuteurs il est donc difficile d'interpréter ce résultat.

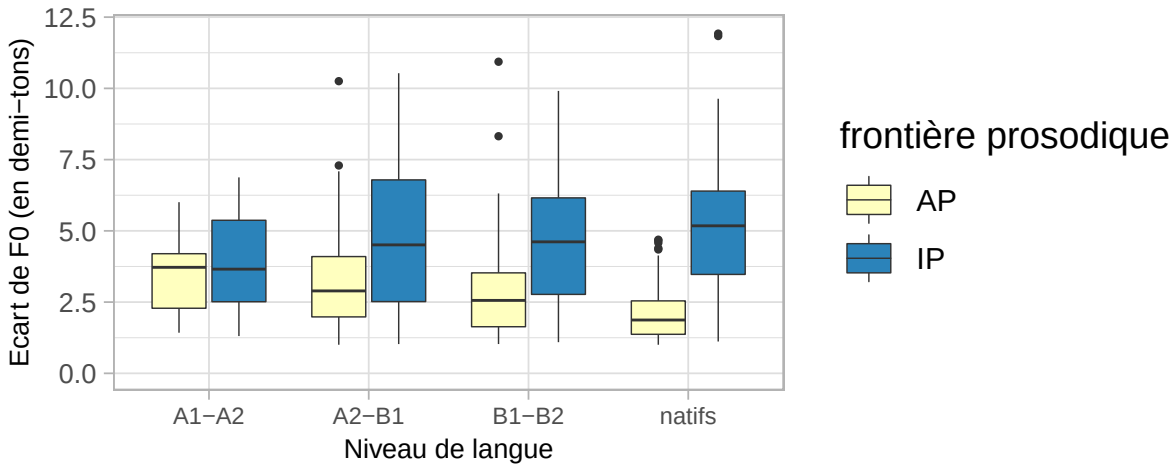


FIGURE 2 – Ecart de F0 (en demi-tons) selon la frontière prosodique et le niveau de langue

Nous n'observons que peu de variation entre les différents locuteurs : seulement 4 locuteurs montrent une tendance inverse avec des montées prosodiques légèrement moins importantes en frontière de IP. Les apprenants semblent ainsi, de façon générale, capables de produire des montées prosodiques plus importantes aux frontières de IP.

3.3 Analyse de la durée

Si les résultats de l'analyse du F0 ont montré que les apprenants distinguaient les frontières de AP des frontières de IP, il semblerait que cette distinction ne se retrouve pas dans les résultats de durée de la voyelle accentuée (voir Figure 3). En effet, les francophones produisent des voyelles nettement plus longues en frontière de IP tandis que l'on observe qu'une faible tendance chez les apprenants. Pour confirmer ces observations, nous avons construit un modèle linéaire mixte avec comme variable dépendante la durée de la voyelle en milliseconde, une interaction entre le niveau de langue et la frontière prosodique comme effets fixes et le locuteur ainsi que la nature de la voyelle accentuée comme effets aléatoires. L'analyse des contrastes de l'interaction vont dans le sens des observations faites : les francophones natifs produisent des voyelles significativement plus longues en frontière de IP que de AP (p<.001***) mais ce n'est pas le cas des deux groupes intermédiaires. Cette distinction significative est aussi présente dans le groupe de débutants (p=.002**) mais encore une fois le peu de données dans ce groupe de langue ne nous permet pas de généraliser cette observation.

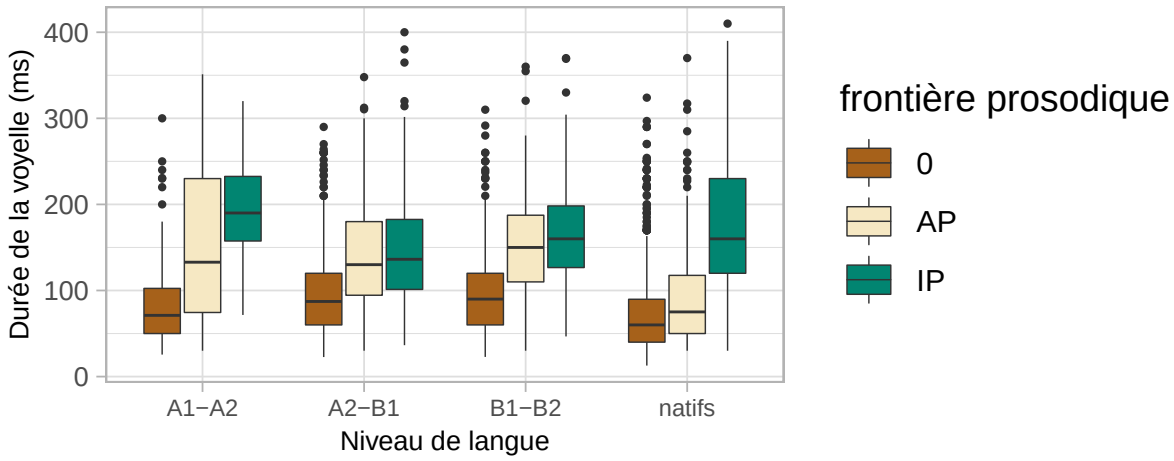


FIGURE 3 – Durée de la voyelle suivant le niveau de langue et la frontière prosodique : 0 (pas de frontière), AP ou IP

Différentes stratégies peuvent être observées chez les locuteurs (Figure 4). Les natifs produisent tous des voyelles plus longues en frontière de IP que de AP mais cette distinction est plus marquée chez deux locuteurs, en particulier FR2, où l'écart important entre les deux frontières est également associé à la suppression de la distinction voyelle non-accentuée – frontière de AP. Les quelques apprenants montrant un schéma similaire aux natifs (ici JP5 en exemple) se rapprochent plus de la production de FR1.

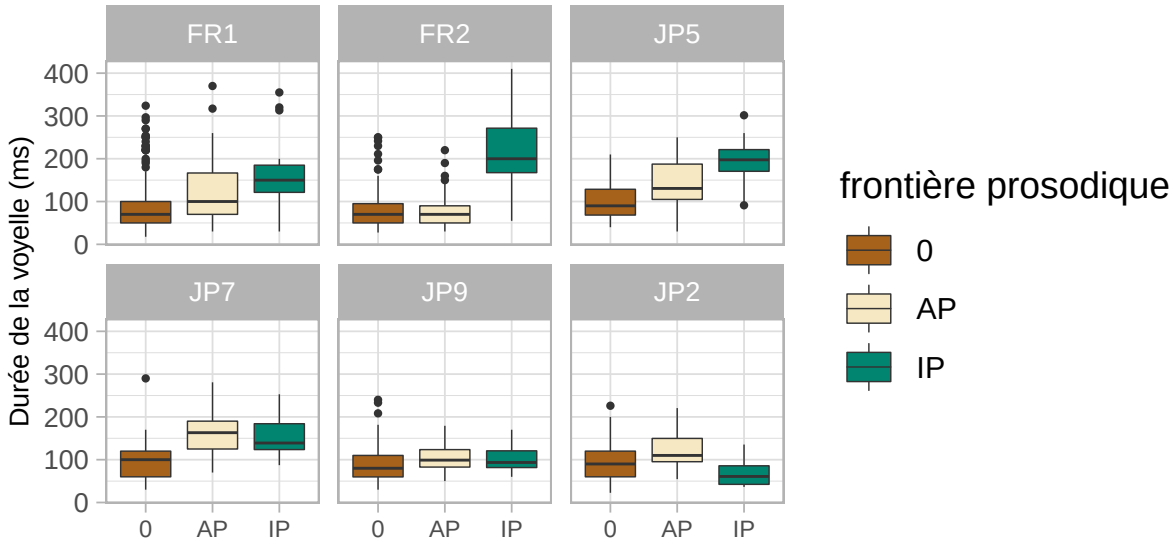


FIGURE 4 – Exemples de durées vocaliques produites par six locuteurs, deux natifs (FR1 et FR2) ainsi que quatre apprenants

Cependant, seulement 4 apprenants observent un tel schéma et le reste d'entre eux ne distinguent pas les frontières de AP de celles de IP. Trois autres schémas peuvent ainsi être observés :

— L'apprenant produit des voyelles accentuées plus longues que celle non-accentuées mais ne distingue pas les frontières de AP de celles de IP (exemple : JP7)
— L'apprenant produit des voyelles de durées équivalentes qu'elles soient accentuées ou non (exemple : JP9)
— L'apprenant produit des voyelles plus courtes en frontière de IP que de AP (exemple : JP2)

La locutrice FR2 est également celle qui produisait un faible pourcentage de contours montants aux frontières de IP (voir 3.1) parmi les natifs et il est intéressant de noter qu'en contrepartie elle y associe un allongement très important de la voyelle. Ce n'est pas le cas de l'apprenant JP9, qui produit également peu de contours montants à cette position prosodique, mais dont la durée des voyelles accentuées n'est pas allongée. Cela montre que si les deux stratégies peuvent sembler similaires (contours descendants), elles diffèrent dans leur réalisation phonétique.

3.4 Ratio d'allongement des voyelles accentuées

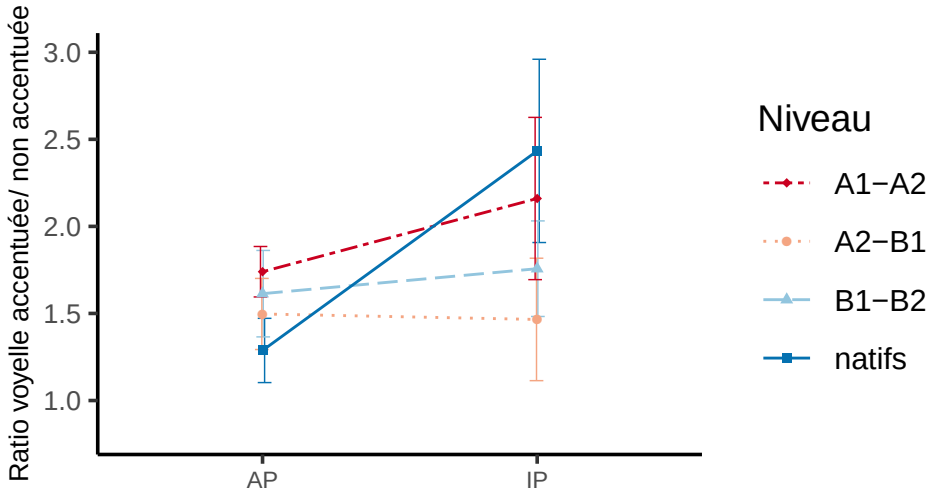


FIGURE 5 – Ratio voyelle accentuée/ non-accentuée aux frontières AP et IP selon le niveau de langue. Les barres représentent l'écart-type.

Les analyses présentées en 3.3. nous ont montré que les apprenants japonophones ne différenciaient pas les frontières prosodiques avec l'indice de durée et ne produisaient donc pas des voyelles plus longues en frontière de IP que de AP. La figure 5 présente le ratio voyelle accentuée / voyelle non-accentuée aux frontières de AP et de IP, et nous pouvons observer que ce ratio tend à être plus élevé chez les natifs en frontière de IP que pour les apprenants. En frontière de AP, il semble en revanche

plus réduit chez les natifs. Encore une fois, nous observons également beaucoup de variation dans les groupes de locuteurs comme nous le suggère l'écart-type important à l'intérieur des groupes. De manière générale, et en prenant pour référence le groupe de francophones natifs et par rapport à la durée des voyelles non accentuées, l'allongement vocalique des apprenants japonophones serait donc légèrement trop important en frontière de AP tandis qu'il serait insuffisant en frontière de IP.

4 Conclusion

Nos résultats montrent que les apprenants japonophones produisent autant de contours montants que les natifs lors des descriptions et d'images et que, comme eux, ils produisent plus de contours montants en frontière de IP que de AP. Ils arrivent également à distinguer les montées prosodiques à ces deux frontières en produisant des montées de F0 plus importantes en IP qu'en AP. En revanche, leur utilisation du paramètre de durée diffère des francophones natifs puisqu'ils produisent un allongement équivalent en frontière de AP et de IP, là où la différence est très claire dans les productions du groupe contrôle. Ainsi, les apprenants japonophones ne parviendraient pas à produire un allongement suffisant lors des continuations majeures aux frontières de IP.

La capacité des apprenants à bien gérer la fréquence fondamentale pour produire des contours montants en distinguant les frontières pourrait s'expliquer par la présence de ce même type de mouvements mélodiques montants dans leur langue maternelle en parole spontanée pouvant être produits aux frontières prosodiques (Igarashi, 2015). Ces contours, ayant une fonction discursive comparable aux continuations majeures du français, sont cependant facultatifs et cela aurait pu impliquer un nombre réduit de contours montants, que nous ne n'avons cependant pas observé. En ce qui concerne la gestion de la durée, les apprenants produisent des voyelles plus longues en position accentuée que non accentuée. Néanmoins, ce paramètre ne semble pas être un indice robuste pour exprimer la continuation discursive, contrairement aux observations sur la parole spontanée et monologue en japonais L1 (Koiso *et al.*, 1998; Yoneyama *et al.*, 2003).

En japonais, les mouvements mélodiques montants n'apparaissent qu'en parole spontanée, et les frontières prosodiques sont donc associées à un ton bas (L%) en parole lue. Nos résultats peuvent être mis en parallèle avec ceux obtenus par l'étude de lectures de textes des mêmes apprenants (Albar & Yoo, 2020). Cette étude a montré que les apprenants ne distinguent pas les frontières prosodiques (notamment AP et IP) en parole lue dans leur production de la F0 et de la durée. Si ces résultats sont très proches de ceux obtenus dans la présente étude, la principale différence réside dans la gestion de la F0. En effet, nous avons observé que les apprenants sont ici capables de produire des montées plus importantes en frontière de IP que de AP. Ce résultat spécifique à la parole spontanée confirmerait ainsi l'hypothèse d'un transfert positif des mouvements mélodiques montants et de leur fonction discursive, facilitant la gestion de la fréquence fondamentale. En revanche, la facilité à produire les montées prosodiques à la fois en parole lue (où les contours montants sont absents en japonais) et en parole semi-spontanée ne peut s'expliquer par un potentiel transfert prosodique.

Remerciements

Cette étude a été financée par l'ANR-10-LABX-0083. Je souhaite également remercier Hiyon Yoo pour sa supervision.

Références

ALBAR R. & YOO H. (2020). The production of French continuation contours at different prosodic boundaries by Japanese learners. In *Actes de conférences de Speech Prosody, à paraître*.

BECKMAN M. E. & PIERREHUMBERT J. B. (1986). Intonational structure in Japanese and English. *Phonology*, **3**, 255–309.

BIGI B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician-International Society of Phonetic Sciences*, (111-112), 54–69.

BOERSMA P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, **5**.

DELAIS-ROUSSARIE E. & YOO H.-Y. (2011). Learner corpora and prosody : From the coreil corpus to principles on data collection and corpus design. *Poznań Studies in Contemporary Linguistics PSiCL*, **47**, 26.

DELATTRE P. (1966). Les dix intonations de base du français. *French review*, p. 1–14.

DI CRISTO A. (2016). *Les musiques du français parlé : Essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain*, volume 1. Walter de Gruyter GmbH & Co KG.

IGARASHI Y. (2015). Intonation. *Handbook of Japanese phonetics and phonology*, p. 525–568.

JUN S.-A. & FOUGERON C. (2000). A phonological model of French intonation. In *Intonation*, p. 209–242. Springer.

KOISO H., HORIUCHI Y., TUTIYA S., ICHIKAWA A. & DEN Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech*, **41**(3-4), 295–321.

MICHELAS A. (2011). *Caractérisation phonétique et phonologique du syntagme intermédiaire en français : de la production à la perception.* Thèse de doctorat.

PORTES C. & BERTRAND R. (2005). De la valeur interactionnelle du contour "continuatif" en français. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, **24**, 139–157. HAL : hal-00241553.

PORTES C., BERTRAND R. & ESPESSER R. (2007). Contribution to a grammar of intonation in French. form and function of three rising patterns. *Nouveaux cahiers de linguistique française*, **28**, 155–162.

R CORE TEAM (2017). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

VENDITTI J. J., MAEDA K. & SANTEN J. P. V. (1998). Modeling Japanese boundary pitch movements for speech synthesis. In *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*, p. 317–322.

YONEYAMA K., FON J. & KOISO H. (2003). Durational and prosodic patterning at discourse boundaries in Japanese spontaneous monologs. In *Proc. 15th International Congress of Phonetic Sciences*, p. 2637–2640.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La pause chez les personnes âgées – une étude exploratoire

Betty Appavoo, Camille Fauth, Béatrice Vaxelaire et Rudolph Sock
Université de Strasbourg, LiLPa UR1339, F-67000 Strasbourg

bappavoo@unistra.fr

RESUME

La production de la parole chez la personne âgée a fait l'objet de nombreuses études qui portaient essentiellement mais pas exclusivement sur les aspects vocaux. Dans ce travail exploratoire, nous cherchons à évaluer les effets du vieillissement sur l'organisation de la lecture. La distribution des pauses et des groupes rythmiques, ainsi que leurs durées respectives ont été quantifiés, de même que la vitesse d'élocution et la vitesse d'articulation pour un groupe de 10 locuteurs âgés (60 à 80 ans) et un groupe de 10 locuteurs témoins (40 à 55 ans). Les résultats indiquent des différences significatives pour les durées des groupes rythmiques et des pauses et pour la vitesse d'élocution ; les locuteurs âgés ayant un débit plus lent que les locuteurs plus jeunes. Ces différences nous semblent intéressantes à poursuivre afin d'étudier plus précisément les différentes stratégies de réorganisation que peuvent mettre en place les locuteurs en fonction de leur âge.

ABSTRACT

Pauses in Older Speakers reading task - An Exploratory Study.
The production of speech in the elderly has been the subject of many studies that have focused primarily, but not exclusively, on the vocal aspects. In this exploratory work, we seek to evaluate the effects of aging on the organization of reading. The distribution of pauses and rhythmic groups and their respective durations were quantified, as well as the speed of speech and the speed of articulation for a group of 10 elderly speakers (60 to 80 years old) and a group of 10 control speakers (40 to 55 years old). The results indicate significant differences for the duration of the rhythm groups and pauses and for the speaking rate; older speakers had slower speaking rate than younger speakers. These differences seem interesting to us to pursue in order to study more precisely the different strategies that speakers can implement according to their age.

MOTS-CLES : Production de la parole, lecture, pauses, vieillissement
KEYWORDS: Speech production, reading, pauses, aging,

Volume 1 : Journées d'Études sur la Parole, pages 19–26.
HAL : hal-02798510.

1 Introduction

A l'heure où la population française continue de croitre et où le vieillissement démographique se poursuit, il faut considérer les spécificités de la production de la parole chez les personnes âgées dans sa globalité. Les études sur le vieillissement de la voix sont relativement anciennes (par ex Linville, 1987, 1996 et Schötz, 2007) et se sont souvent attachées à décrire les marqueurs acoustiques du vieillissement vocal cherchant parfois à y trouver des similitudes avec la dysphonie (Amman, 1999). Avec l'âge, le temps maximum phonatoire diminue (Remacle, 2006) ce qui est notamment le reflet d'une réduction de l'efficacité pulmonaire (Abitbol, 2013), des facteurs de béance glottique (Remacle, 2006) et d'une modification générale du larynx (Bianco, 2017). Ces études regroupent les modifications de la voix chez la personne âgée sous le terme de presbyphonie (Dehesdin, 1992, Estienne 1998, ou Thomas, 2012 par exemple).

Par ailleurs, nous savons que l'organisation des cycles respiratoires lors de la production de la parole diffère selon différents paramètres tels que les stratégies individuelles (Teston & Autesserre, 1987), la complexité de la planification (Mitchell et al., 1996), la nature de la tâche de production (Wang, et al., 2010), la condition physique (Fauth et al, 2018) mais également selon l'âge du locuteur (Sperry & Klich, 1992). Selon Huber et *al.* (2012), les adultes âgés produisent un plus grand nombre de respirations aux frontières syntaxiques mineures et que très peu à des endroits non syntaxiques sans pour autant affecter l'intelligibilité. Rousier-Vercruyssen et *al.* (2018) ont étudié les pauses silencieuses en relation avec la saillance référentielle et avec l'habilité de prise en compte de l'autre lors de narrations d'images séquentielles. Ils ont uniquement considéré les pauses d'une durée supérieure à 200ms pour calculer la durée moyenne des pauses (DMP) relevant d'une activité cognitive. Ils ont trouvé que « le passage d'une étape de discours à l'autre [comme un changement référentiel] semble plus complexe pour les séniors » et que ceux-ci présentent également « une difficulté de prise en compte de l'autre [qui] est corrélée à une augmentation de la DMP chez les séniors » (Rousier-Vercruyssen et *al*., 2018). Rousier-Vercruyssen et *al.* (2014) notent cependant que l'âge n'a pas d'effet sur la fréquence du nombre de pauses par seconde (Rousier-Vercruyssen et *al*., 2014). Pour eux, « la planification discursive des locuteurs âgés étudiée sous la loupe de variations phonétiques serait moins efficiente que celle de participants jeunes ».

Le vieillissement peut donc provoquer diverses perturbations au niveau de la phonation (et de la cognition) qui vont entrainer des changements au niveau acoustique et au niveau de l'organisation temporelle de la parole. L'objectif de ce travail est de comparer les stratégies de lecture de locuteurs âgés (60 et 80 ans) et des locuteurs plus jeunes (40 et 55 ans). La plupart des études comparent les productions de locuteurs ayant de plus grandes différences d'âge, nous cherchons au contraire à observer s'il est possible d'établir un continuum. Notre comparaison se fonde sur la

réalisation des pauses, la vitesse d'élocution et la vitesse d'articulation en situation de lecture d'un texte court.

2 Méthodologie

2.1 Corpus et participants

Les sujets ont été enregistrés dans la chambre anéchoïque de l'Institut de Phonétique de Strasbourg ou à leur domicile dans une pièce calme, en position assise à l'aide d'un microphone cardioïde de type micro-cravate à pinces, relié à un enregistreur numérique. Les locuteurs avaient pour tâche de lire à une hauteur et à une vitesse confortables une version modifiée de l'histoire des Trois petits cochons, soit 13 phrases. Pour cette étude exploratoire, 20 locuteurs ont été enregistrés. 5 hommes (67,4 ans en moyenne, E.T. 7,7) et 5 femmes (71,6 ans en moyenne, E.T. 6,7) âgés de 60 à 80 ans et 10 personnes appariées en genre âgées de 51,5 ans en moyenne (groupe témoin). Ils avaient tous pour langue maternelle le français, ne présentaient aucune pathologie vocale, respiratoire ou auditive et été non-fumeurs. Le niveau d'éducation a également été contrôlé.

2.2 Mesures

Les données acoustiques ont été analysées à l'aide de Praat. Elles ont été segmentées de façon semi-automatique grâce à EasyAlign (Goldman, 2011). La détection des pauses s'est faite à partir d'indices perceptivo-visuels, On a considéré que les pauses silencieuses correspondaient à un silence perceptible, accompagné d'une rupture d'activité acoustique visible sur le signal de parole. Aucun seuil de durée n'a été appliqué hormis pour la tenue des occlusives non voisées, leur tenue a été fixée à 50ms dans ces cas. Les pauses initiales et finales de l'énoncé n'ont pas été prises en compte.

La durée et le nombre des syllabes, des pauses silencieuses syntaxiques et des pauses non syntaxiques ont été relevés automatiquement. Nous avons considéré qu'une pause silencieuse syntaxique est une pause silencieuse dont l'emplacement respecte les règles de syntaxe. Elle doit être placée entre deux groupes syntaxiques. Nous avons considéré les pauses silencieuses placées à l'intérieur d'un groupe syntaxique minimal, ne respectant pas les règles de syntaxe comme non syntaxiques. Etant donné que nous nous intéressons à l'organisation temporelle de la parole, les hésitations, faux-départs et allongements ont été considérés comme faisant partie du signal de parole et non comme des pauses remplies. De plus ont été calculées : la vitesse d'élocution (VE), c'est-à-dire le nombre de syllabes par seconde en prenant en compte la durée des pauses, et la vitesse d'articulation (VA) ou le nombre de syllabes par secondes sans tenir compte de la durée des pauses.

2.3 Hypothèses

Étant donné que le vieillissement touche également le système de production de la parole, il devrait y avoir des différences objectives entre les stratégies de lecture des locuteurs âgés et les plus jeunes, ce qui devrait se traduire par une durée totale d'élocution plus longue chez les locuteurs âgés par rapport aux locuteurs témoin, un nombre de pauses et donc un nombre de groupes rythmiques plus important chez les locuteurs âgés par rapport aux locuteurs témoin, une durée de pauses plus longue chez les locuteurs âgés par rapport aux locuteurs témoin, une vitesse d'élocution et une vitesse d'articulation plus lente chez les locuteurs âgés par rapport aux locuteurs témoin.

3 Résultats

Les valeurs suivant une loi normale, nous avons appliqué un test t de Student pour échantillon appariés ou non appariés en fonction des données sélectionnées. Le seuil de significativité a été considéré comme suit : $p < 0,05$.

3.1 Durées totales d'énonciation

La durée totale de lecture (voir Figure 1), qui prend en compte la durée d'articulation et la durée des pauses silencieuses, est en moyenne plus longue chez les locuteurs âgés que chez les locuteurs témoins, soit 79 129ms (E.T. 12 678) et 71 367ms (E.T. 11 491). Autrement dit, les locuteurs âgés ont produit en moyenne une énonciation totale 11% plus longue que les locuteurs témoin. Cette différence n'est toutefois pas significative ($p > 0.05$). Ce sont notamment les femmes âgées qui ont la durée totale d'énonciation la plus longue (19% plus importante que celles des locutrices témoins). La différence est moins notable pour les locuteurs masculins. Les écarts-types témoignent de la grande variabilité de stratégie pour ces deux populations.

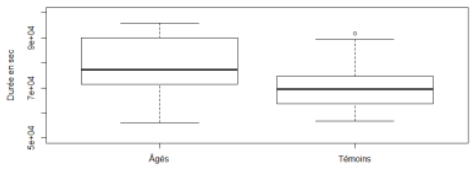
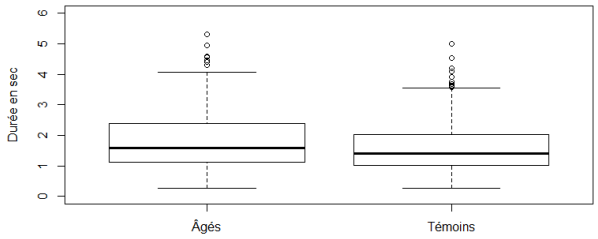


Figure 1 : durées moyennes totales pour les locuteurs âgés (à gauche) et pour les locuteurs témoins (à droite)

3.2 Groupes rythmiques et pauses silencieuses

Le nombre de groupes rythmiques dépend du nombre de pauses. Il est égal au nombre de pauses +1. En comparant les nombres moyens de groupes rythmiques et donc de pauses entre le groupe âgé et le groupe témoin, on peut observer que si le nombre moyen de groupes rythmiques et de pauses syntaxiques dans le groupe âgé est légèrement moins important que dans le groupe témoin, la différence n'est pas significative (32 groupes rythmiques en moyenne et 31 pauses syntaxiques en moyenne pour le groupe âgé ; 34 et 32 respectivement pour le groupe témoin). Comme attendu, les locuteurs produisent très peu de pauses non syntaxiques (maximum 1 par locuteur), ceci s'explique naturellement par la tâche de production et par sa simplicité, nous avons donc considéré les pauses dans leur totalité qu'elles soient syntaxiques ou non syntaxiques du moment qu'elles étaient silencieuses.

Si l'on s'intéresse aux durées des groupes rythmiques (figure 2) et des pauses silencieuses (figure 3) pour chacun des groupes, nous observons des différences significatives ($p < 0,05$) de durées.



.

Figure 2 : Durée moyenne en seconde des groupes rythmiques pour les locuteurs âgés (à gauche) et pour les locuteurs témoins (à droite)

La durée moyenne des groupes rythmiques est de 1 818ms pour le groupe des locuteurs âgés alors qu'elle n'est que de 1 610ms pour les locuteurs plus jeunes, soit une différence significative ($p < 0.008$). Les locuteurs âgés produisent donc des groupes rythmiques plus longs que les locuteurs témoins. Les écarts types sont très faibles pour les deux populations (997ms et 866ms respectivement) et témoignent d'une faible variabilité pour cette mesure. Toutefois, si l'on étudie les groupes un peu plus en détail, nous observons que ce sont les femmes âgées qui produisent les groupes rythmiques les plus longs. Ceci était partiellement prévisible puisque ce sont également elles qui avaient la durée d'énonciation totale la plus importante.

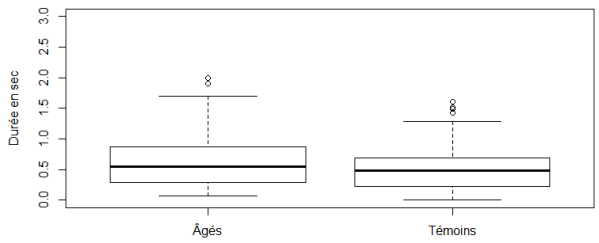


Figure 3 : Durée moyenne en seconde pauses pour les locuteurs âgés (à gauche) et pour les locuteurs témoins (à droite)

Enfin, en ce qui concerne la durée des pauses, leur durée moyenne dans le groupe âgé est de 641ms et de 518ms dans le groupe de témoin. Cette différence est significative ($p < 0.001$). Concernant les pauses, les écarts-type par groupe et sous-groupe sont relativement importants, cela signifie qu'il existe une forte variabilité interindividuelle. Ici également, ce sont les femmes âgées qui produisent les pauses les plus longues, ce qui s'explique partiellement compte tenu du fait que c'était également elles qui avaient la durée totale d'énonciation la plus importante. Il apparait donc qu'elles ont les groupes rythmiques les plus longs mais également les pauses les plus longues.

3.3 Vitesse d'élocution et d'articulation

Enfin en dernier lieu, nous avons cherché à quantifier les vitesses d'élocution et d'articulation en syllabes par seconde (en prenant en compte les pauses ou sans prendre en compte les pauses, respectivement). En ce qui concerne ces deux mesures, elles sont effectivement modifiées en fonction du groupe mais seule la vitesse d'élocution est significativement modifiée en fonction de l'âge du locuteur ($p < 0.03$). La vitesse d'articulation est, en moyenne, de 4,8 syllabes par seconde (E.T. 0,64) pour les locuteurs âgés et de 5,2 syllabes par seconde (E.T. 0,48) pour les locuteurs témoins. La vitesse d'élocution, elle, est en moyenne de 3,5 syllabes par seconde (E.T. 0,73) pour les locuteurs âgés et de 4,1 syllabes par seconde (E.T. 0,64) pour les locuteurs témoins (E.T. 0,60). C'est donc bien la durée des pauses qui produit le plus grand changement dans l'organisation temporelle de la parole chez les locuteurs âgés, ceci étant particulièrement marqué dans les productions des locutrices âgées.

4 Conclusions

Il convient à présent de vérifier si nos hypothèses initiales. Nous pouvons confirmer notre première hypothèse, qui stipulait que les locuteurs âgés présenteraient une durée totale d'énonciation plus longue que les locuteurs témoins. Cette différence n'est

toutefois pas significative. Contrairement à notre seconde hypothèse, les locuteurs âgés ne produisent pas nécessairement plus de pauses syntaxiques et donc plus de groupes rythmiques que les locuteurs témoin. Il n'existe en effet pas de différence remarquable au niveau du nombre de pauses syntaxiques et des groupes rythmiques. Quant aux pauses non syntaxiques, elles sont présentes dans les deux groupes mais en très petit nombre. Il n'est donc pas pertinent de les étudier dans une comparaison de locuteurs âgés et moins âgés dans une tâche de lecture. Les locuteurs âgés ont, en effet, tendance à produire des pauses significativement plus longues que les locuteurs témoin. En ce qui concerne la vitesse d'articulation et la vitesse d'élocution, il existe certes une différence mais celle-ci est essentiellement due à une durée plus longue des pauses. Les locuteurs âgés présentent une vitesse d'articulation moyenne moins importante que celle des locuteurs témoin. Il ne s'agit cependant que d'une petite différence. Nous pouvons confirmer ce que la littérature postule quant à l'existence de différence entre la production de la parole féminine âgée et la production de la parole masculine âgée. Ces différences n'existent cependant qu'au niveau de la vitesse d'articulation qui est plus rapide chez les hommes si on considère les paramètres pris en compte dans notre étude. L'augmentation de la durée des pauses confirment les résultats attestés dans la littérature notamment dans des tâches de parole spontanée. Les auteurs (Rousier et al, 2018 par exemple) font le lien entre la durée des pauses et les capacités cognitives de planification des séniors. Si nos investigations n'ont pas mesuré cette variable, il apparait cependant que l'âge a effectivement un effet, non pas sur le nombre, mais bien sur la durée moyenne des pauses. Il convient donc de prendre en considération toutes les étapes des processus de production de parole qui devraient être examinées dans un contexte conceptuel élargi en prenant en compte les modèles psycholinguistiques de la production de la parole, notamment.

5 Perspectives

Notre étude exploratoire présentait une cohorte relativement étroite, il aurait été intéressant de travailler avec des locuteurs plus nombreux et peut-être considérer des intervalles d'âge plus jeunes (également appariés en sexe). Par ailleurs, il nous semble intéressant de proposer la lecture d'un texte plus long qui pourrait faire émerger de façon plus saillante encore des stratégies de lecture différentes notamment en raison de la fatigue vocale. De plus, le vieillissement a également un effet sur la loquacité des sujets, une tâche de parole semi-spontanée pourra donc également être ajoutée (Singh et al., 2001). Enfin, cette étude acoustique nous semble être un bon point de départ pour conduire des investigations sur la gestion des flux respiratoires. Ces données pourront être acquises à l'aide d'un Respitrace et ainsi permettre des comparaisons avec d'autres études. Notons enfin que la tâche de lecture n'est peut- être pas aussi aisée pour certains sujets que pour d'autres. Ceci peut être dû à l'évidence à des problèmes sensoriels visuels. Ce peut être aussi le résultat d'une mauvaise maitrise du processus de lecture, voire des processus d'oralisation de la lecture performée.

6 Références

ABITBOL J. (2005). *l'Odyssée de la voix* (Robert Laffont). Paris.

DEHESDIN D (1992). *Presbyphonie*. In, FRACHET B., MORGON A., LEGENT F., Eds. *Pratique phoniatrique en ORL* Masson : Paris.

DUEZ D. (2003). Le pouvoir du silence et le silence du pouvoir : comment interpréter le discours politique, *MediaMorphoses*, (8), 77-82. http://hdl.handle.net/2042/23254.

ESTIENNE F. (1998). *Voix parlée, voix chantée : examen et thérapie* Masson : Paris.

FAUTH C., DUCHEMIN A., VAXELAIRE B., SOCK R. (2018). Perturbation de l'organisation temporelle de la parole suite à un effort physique. *Actes XXXIIe Journées d'Études sur la Parole*, 240-248. DOI : 10.21437/JEP.2018-28.

GOLDMAN J-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Proceeding of Interspeech 2011*. Florence, p. 3233-3236.

HUBER J., DARLING M., FRANCIS E., ZHANG D (2012) Impact of typical aging and Parkinson's disease on the relationship among breath pausing, syntax and poncutation, American Journal of Speech Language Pathology, 21-4, p 368-379. DOI 10.1044/1058-0360(2012/11-0059)

LINVILLE S E. (1996). The sound of senescence, *Journal of Voice*, 10, p. 190–200. DOI : 10.1016/S0892-1997(96)80046-4.

MITCHELL H L., HOIT J D., WATSON P J. (1996). Cognitive-linguistic demands and speech breathing. *Journal of Speech and Hearing Research,* 39(1), p. 93-104. DOI : 10.1044/jshr.3901.93.

ROUSIER-VERCRUYSSEN L.-D., LACHERET-DUJOUR A & FOSSARD M. (2014). Pauses silencieuses, planification discursive et vieillissement langagier, Pauses silencieuses, planification discursive et vieillissement langagier. *Nouveaux Cahiers de Linguistique Française*, Université de Genève, 2014, SWIP 3 - Swiss Workshop In Prosody, 31, p. 197-203. halshs-01086450.

REMACLE E et al. (2006). La presbyphonie. Le vieillissement de la voix. In *La voix parlée et chantée, Klein-Dallant* (Klein-Dallant, pp. 141–147). Ville-d'Avray.

ROUSIER-VERCRUYSSEN L.-D., LACHERET-DUJOUR A & FOSSARD M (2018). Que révèle la pause silencieuse sur l'accessibilité cognitive d'un référent et le vieillissement langagier ? *Langages* 2018/3, 211, p. 97 - 109. DOI : 10.3917/lang.211.0097.

SCHÖTZ S. (2007). Acoustic analysis of adult speaker age, in C. Müller (Ed.) *Speaker classification I, Lecture notes in Computer Science*, 2007, 1, p. 88-107. DOI : 10.1007/978-3-540-742.

SINGH, S., BUCKS, R., & CUERDEN, J. (2001). An evaluation of an objective technique for analyzing temporal variables in DAT spontaneous speech. Aphasiology, 15 (6), p. 571-583.

SPERRY E E., KLICH R J. (1992). Speech breathing in senescent and younger women during oral reading. *Journal of Speech and Hearing Research*, *35*(6), p. 1246-1255. DOI : 10.1044/jshr.3506.1246.

TESTON B., AUTESSERRE D. (1987). L'aérodynamique du souffle phonatoire utilisé dans la lecture d'un texte en français. *Proceedings of XIth Congress of Phonetic Sciences (ICPhS)*, Tallin, Estonie, p. 33-36.

THOMAS L (2012). *The aging voice from clinical symptoms to biological realities.* Thèse de doctorat,. University of Kentucky.

WANG Y T., GREEN J R., NIP I S B., KENT R D., KENT J F. (2010). Breath Group Analysis for Reading and Spontaneous Speech in Healthy Adults. *Folia Phoniatrica et Logopaedica*, *62*(6), p. 297-302. DOI : 10.1159/000316976.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Reconnaissance automatique de la parole : génération des prononciations non natives pour l'enrichissement du lexique

Ismaël Bada, Dominique Fohr, Irina Illina

Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France
{ismael.bada,dominique.fohr,irina.illina}@loria.fr

RÉSUMÉ

Dans cet article nous proposons une méthode d'adaptation du lexique, destinée à améliorer les systèmes de la reconnaissance automatique de la parole (SRAP) des locuteurs non natifs. En effet, la reconnaissance automatique souffre d'une chute significative de ses performances quand elle est utilisée pour reconnaître la parole des locuteurs non natifs, car les phonèmes de la langue étrangère sont fréquemment mal prononcés par ces locuteurs. Pour prendre en compte ce problème de prononciations erronées, notre approche propose d'intégrer les prononciations non natives dans le lexique et par la suite d'utiliser ce lexique enrichi pour la reconnaissance. Pour réaliser notre approche nous avons besoin d'un petit corpus de parole non native et de sa transcription. Pour générer les prononciations non natives, nous proposons de tenir compte des correspondances graphèmes-phonèmes en vue de générer de manière automatique des règles de création de nouvelles prononciations. Ces nouvelles prononciations seront ajoutées au lexique. Nous présentons une évaluation de notre méthode sur un corpus de locuteurs non natifs français s'exprimant en anglais.

ABSTRACT

In this study we propose a method for lexicon adaptation in order to improve the automatic speech recognition (ASR) of non-native speakers. ASR suffers from a significant drop in performance when used to recognize the speech of non-native speakers, since the phonemes of the foreign language are frequently poorly pronounced by these speakers. To take into account this problem of erroneous pronunciations, we integrate non-native pronunciations in the lexicon and subsequently we use this augmented lexicon for speech recognition of non-natives speakers. To realize our approach we need a small corpus of non-native speech and its transcription. To generate non-native pronunciations, we take into account relationships graphemes-phonemes in the analyzed pronunciations, with a view to automatically generating rules for creating new pronunciations, which will be added to the lexicon. We present an evaluation of our method on a corpus of non-native French speakers, pronouncing sentences in english.

MOTS-CLÉS : reconnaissance automatique de la parole, locuteurs non natifs, lexique
KEYWORDS : automatic speech recognition, non-native speech, lexicon

1 Introduction

Les systèmes de reconnaissance automatique de la parole (SRAP) ont fait des progrès continus au fil des années, grâce à l'utilisation des réseaux de neurones artificiels dans l'élaboration des modèles acoustiques et à la puissance de calcul toujours plus élevée, entraînant une baisse notable du taux d'erreur de reconnaissance. Mais le traitement de la parole non native reste encore un défi à relever.

Volume 1 : Journées d'Études sur la Parole, pages 27–35.
HAL : hal-02798511.

Des études phonologiques (Park et Culnan, 2019) ont mis en évidence les problèmes posés par les locuteurs non natifs : les prononciations des phonèmes de la langue parlée, influencées par celles de la langue maternelle, des intonations différentes, des prononciations hachées, des hésitations et des auto-corrections. Plusieurs approches ont été tentées pour rendre les SRAP tolérants à la parole non native. Elles sont de deux types : l'une utilise l'**augmentation automatique du lexique** avec des prononciations alternatives, et l'autre consiste à **modifier les modèles acoustiques** pour les rendre compatibles avec les phonèmes de la langue parlée. Cela peut être effectué en utilisant un corpus de parole non native (Duan *et al.*, 2017 ; Li *et al.*, 2016) ou seulement la parole native (Lee et Glass, 2015 ; Das et Hasegawa-Johnson, 2015). Une combinaison de ces deux approches est également envisageable (Tan, 2008 ; Goronzy *et al.*, 2004). Une bonne reconnaissance des noms propres dans un contexte multilingue est obtenu en utilisant un modèle acoustique multilingue et de transcriptions nativisées émergeant de G2P (*grapheme-to-phoneme*) de ces langues (Reveil *et al.* 2010).

L'enrichissement du lexique par génération automatique de prononciations alternatives peut se faire *via* un jeu de règles pré-établies de substitution de phonèmes (Schaden, 2004). Ces règles sont issues d'études phonologiques entre les différences de prononciations de phonèmes entre la langue parlée et la langue native. Une autre variante automatisée consiste à utiliser une base de données de parole non native afin de générer une matrice de confusion entre les phonèmes de la langue parlée et les phonèmes de la langue native (Livescu *et al.*, 2000).

L'adaptation des modèles acoustiques en vue de rendre tolérant le SRAP à la parole non native peut consister à utiliser deux ensembles de modèles acoustiques, l'un adapté à la langue maternelle du locuteur, et l'autre à la langue non native. Cela permet de faire une reconnaissance phonétique et un alignement sur la parole non native, dans le but d'extraire les différences de prononciations de phonèmes et de les intégrer dans les modèles acoustiques (Morgan J., 2004). Enfin une variante consiste à intégrer la graphie des mots (Bouselmi G. et al., 2006) dans l'étude des différences de prononciations entre la parole native et celle non native, afin de générer des règles de création de prononciations, règles qui seront ensuite utilisées pour modifier les modèles acoustiques pour les rendre tolérant à une parole non native.

Dans notre article nous nous intéressons à l**'augmentation du lexique.** Notre méthode pour traiter le problème de la parole non native emprunte certaines idées de deux approches : l'enrichissement du lexique canonique par génération de règles, et la prise en compte des correspondances graphèmes-phonèmes dans l'élaboration de ces règles. Nous ne modifions pas le modèle acoustique, c'est donc une approche rapide à mettre en œuvre. Pour réaliser notre approche nous avons besoin d'un petit corpus de parole non native et sa transcription qui nous permet de générer ces nouvelles règles de prononciations. Par rapport aux méthodes de l'état de l'art, notre approche permet d'utiliser simultanément des phonèmes de la langue cible et de la langue maternelle du locuteur. De plus, seul le lexique de l'application est modifié.

2 Méthodologie proposée

Notre approche de prise en compte de la parole non native dans un système de reconnaissance de la parole s'appuie sur l'hypothèse que les locuteurs non natifs **peuvent être influencés par la graphie des mots étrangers à prononcer.** Dans ce cas, les locuteurs non natifs peuvent prononcer certains phonèmes d'un mot en utilisant des règles de phonétisation issues de leur langue maternelle. Par exemple, le graphème « e » du mot anglais *zero* doit se prononcer /ɪ/ (zero /z ɪ r oʊ/). Mais un

français risque de le prononcer /e/ car le graphème « e » ne se prononce pratiquement jamais */i/* en français.

Pour prendre en compte ce phénomène, nous proposons d'enrichir le lexique du système de reconnaissance en y ajoutant les prononciations non natives en tenant compte de la graphie des mots. Nous proposons effectuer seulement les **substitutions** de certains phonèmes natifs par des phonèmes non natifs en fonction de la graphie. Le lexique augmenté sera utilisé pour effectuer la reconnaissance.

Voici en quelques lignes le résumé de notre méthode. Dans une première étape, nous extrayons les correspondances phonème-graphème à partir d'un *lexique phonétique canonique* (lexique de départ du SRAP) : nous effectuons un lien entre chaque phonème d'un mot du lexique et les graphèmes composant ce mot. Puis, à l'aide d'un corpus audio de parole non native, nous identifions les associations entre les phonèmes attendus et les phonèmes réellement prononcés (en utilisant un alignement forcé et une reconnaissance phonétique). A partir de cela, nous créons des règles de substitution des phonèmes en tenant compte des couples graphème-phonème. Ces règles seront appliquées au lexique destiné à la reconnaissance de la parole des locuteurs non natifs, afin de l'enrichir en nouvelles prononciations. La Figure 1 illustre ces étapes. Dans les sections suivantes nous présenterons ces étapes plus en détails et avec des exemples.

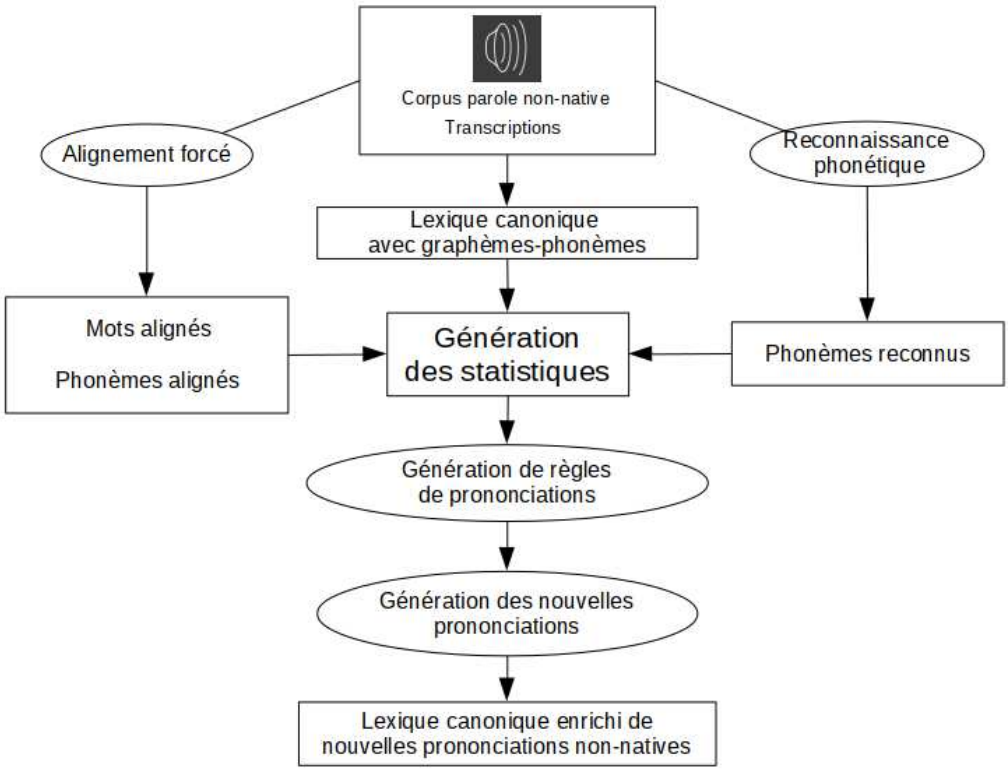


Figure 1*:* Les étapes de la génération des nouvelles prononciations non natives.

2.1 Création d'un lexique avec les associations graphème-phonème

Dans un système de reconnaissance automatique de la parole, **un lexique** phonétisé est une liste de mots avec leurs prononciations (suite de phonèmes). Dans le cadre de notre étude, nous avons également besoin d'un second type de lexique, qui contient ces mêmes mots, mais avec des associations graphème-phonème pour chaque mot. Ce lexique sera généré automatiquement de la manière suivante : en utilisant un algorithme de programmation dynamique, à chaque phonème (ou

groupe de phonèmes) d'un mot du lexique canonique (CMU dictionary), nous affectons le graphème (ou le groupe des graphèmes) qui lui correspond. Voici quelques exemples d'associations graphème-phonème générées :

nogo	n : n	o : əʊ	g : g	o : əʊ
weather	w : w	ea : e	th : ð	er : ɚ
taxi	t : t	a : æ	x : k=s	i : i
zero	z : z	e : ɪ	r : r	o : oʊ

On peut noter que ces associations sont parfois du type « un graphème associé à un phonème » (par exemple e:ɪ) ou parfois du type plusieurs graphèmes associés à un phonème (par exemple ea:e) ou du type « un graphème associé à plusieurs phonèmes » (x:k=s).

2.2 Alignement forcé et reconnaissance phonétique

Pour générer les variantes de prononciations non natives, nous utilisons un corpus audio de parole non native pour lequel nous avons la transcription. Tout d'abord, nous effectuons un alignement forcé de corpus non natif en utilisant le signal audio, les transcriptions et un lexique canonique. Le résultat est un alignement au niveau des phonèmes et au niveau des mots. Ensuite, nous effectuons une reconnaissance phonétique du même corpus, trame par trame (sans lexique ni modèle de langage). Par exemple, pour le mot *nogo* à l'issu de l'alignement forcé nous obtenons :

```
1.530 0.130 n
1.660 0.100 əʊ
1.760 0.070 g
1.830 0.680 əʊ
```

A l'issu de la reconnaissance phonétique trame par trame, pour le segment de parole correspondant à la prononciation du phonème əʊ du mot *nogo*, débutant à la trame 166 et se terminant à la trame 175, nous obtenons les probabilités suivantes :

```
166   əʊ 0.43   n 0.21   ɑ 0.13   u 0.11   ɔ 0.05
167   əʊ 0.88   u 0.04   ɑ 0.04   ɔ 0.02
168   əʊ 0.73   ɔ 0.13   u 0.10   ɑ 0.02
169   əʊ 0.65   ɔ 0.26   u 0.03   ɑ 0.02
170   əʊ 0.62   ɔ 0.27   u 0.03   ɑ 0.02   ʌ 0.01
171   əʊ 0.52   ɔ 0.36   u 0.03   ʌ 0.02   r 0.02   # 0.01
172   əʊ 0.67   ɔ 0.21   r 0.04   l 0.02   ʌ 0.02   ɑ 0.02
173   əʊ 0.78   ɔ 0.10   r 0.04   u 0.02   ʌ 0.01
174   r 0.31   əʊ 0.22   ɔ 0.11   u 0.11   m 0.07   ŋ 0.04
175   r 0.66   əʊ 0.11   g 0.04   m 0.04   ɔ 0.02   ɜː 0.02
```

On peut noter que la somme des probabilités pour chaque trame somme à 1 (les phonèmes qui ont une probabilité inférieure à 0.01 ne sont pas affichés).

2.3 Création de statistiques

Pour créer des statistiques, nous allons utiliser les résultats de l'alignement forcé et de la reconnaissance phonétique. Pour chaque association graphème-phonème, pour chaque phonème reconnu, et pour chaque trame correspondant à ce phonème, nous cumulons les probabilités obtenues. De plus, nous calculons le nombre de trames affecté à chaque association (graphème - phonème, phonème reconnu). Ainsi, nous obtenons des statistiques où pour chaque couple

graphème-phonème, nous avons la distribution en pourcentage des phonèmes reconnus, et leur nombre d'apparitions. Nous appelons ces paramètres α et β. Par exemple, pour le graphème-phonème *e:*ɪ nous avons obtenu :

Association	Phonèmes reconnus			
e: ɪ	ɪ *(40%) (3200)*	e *(35%) (2800)*	ʌ *(23%) (1840)*	i: *(2%) (160)*

Ce tableau montre que le phonème *ɪ* lorsqu'il est associé à la graphie « e », est dans 40 % des cas correctement reconnu comme ɪ, mais dans 35 % des cas il est reconnu comme e.

2.4 Création de règles de prononciations alternatives et génération de nouvelles prononciations

A l'aide de ces statistiques, nous créons des règles de prononciations pour les associations graphème-phonème selon deux critères : nous utilisons seulement les associations graphème-phonème qui ont un pourcentage suffisamment élevé et un nombre d'apparitions est suffisamment grand. On obtient alors un ensemble de règles qui serviront à générer de nouvelles prononciations. Par exemple, en utilisant un seuil de 20 % de taux de phonèmes reconnus et un nombre d'apparitions de 1500, pour notre corpus de parole non native nous obtenons les règles suivantes :

e : ɪ → e
e : ɪ → ʌ

Ces règles seront appliquées à l'ensemble des mots du lexique qui contient ces associations graphème-phonème. Cela permet de créer un nouveau lexique enrichi de prononciations alternatives non natives. Par exemple, pour le mot anglais *zero* du lexique avec les associations graphème-phonème suivantes :

zero z : z e : ɪ r : ɾ o : əʊ

et deux règles trouvées e:ɪ → e, e:ɪ → ʌ, après application de ces règles nous ajouterons dans notre lexique les prononciations suivantes :

zero z e ɾ əʊ
zero z ʌ ɾ əʊ

3 Expériences

Cette partie décrit les corpus utilisés et le protocole expérimental.

3.1 Les corpus utilisés

Pour valider l'approche proposée dans cet article, nous avons utilisé trois corpus de parole non natives. Ces trois corpus ont été prononcés par des français qui s'expriment en anglais :

- ***Nombres*** **:** Il est constitué d'enregistrements audio de locuteurs français non natifs lisant des nombres en anglais. Dans ce corpus, il y a 15 locuteurs, chacun a prononcé 66 commandes comprenant des nombres et de chiffres. La durée totale du corpus est de 78 minutes, le nombre des mots différents est 70. Le corpus contient de la parole lue.
- ***HIWIRE*** : Ce corpus est constitué d'enregistrements audios de locuteurs français non natifs lisant des commandes simples d'aéronautique en anglais. Ce corpus a été enregistré dans le cadre du projet *HIWIRE* (Bouselmi, 2008). Dans ce corpus, 31 locuteurs français ont chacun prononcé 100 commandes. Des données audios ont été enregistrées dans

différents bureaux à l'aide de micro-casque (*close-talking microphone*). Au total il y a 128 minutes de parole enregistrée. La taille du vocabulaire est de 130 mots différents. Le corpus correspond à la parole lue.

- ***Aéro* :** ce corpus contient des commandes aéronautiques en anglais prononcées par 2 locuteurs français. Les enregistrements correspondent à des commandes aéronautiques complexes. A la différence des deux corpus précédents qui étaient de la parole lue, ce corpus est un enregistrement de parole spontanée et de nombreuses hésitations et reprises sont présentes. La durée d'enregistrement est de 27 minutes. Le vocabulaire est de 900 mots.

La Table 1 résume ces corpus. Tous les corpus ont été transcrits manuellement. Nous utilisons le corpus *Aero* et une partie du corpus *HIWIRE* pour valider l'approche proposée et les **autres corpus** pour la **génération automatique des règles** de prononciations non natives.

Corpus	*Taille de vocabulaire (nbr. de mots)*	*Nbr. de locuteurs*	*Durée (minutes)*
Nombres	70	15	78
HIWIRE	130	31	128
Aero	900	2	27

Table 1 : Statistiques des corpus.

3.2 Système de reconnaissance

Notre système de reconnaissance est fondé sur la boîte à outils de reconnaissance vocale *Kaldi* (Povey et al., 2011). Nous utilisons les modèles acoustiques senones de type TDNN (*Time Delay Neural Network*).

Nous avons développé deux systèmes de reconnaissance.

- **SRAP_anglais**. Il est appris sur un corpus de conférences en **anglais** (TED-LIUM). L'apprentissage des modèles acoustiques anglais TDNN a été réalisé en utilisant les 452 heures du corpus d'apprentissage. 39 phonèmes anglais plus un modèle de silence et un modèle de bruit sont utilisés.
- **SRAP_anglais-français**. Il est appris en utilisant le même corpus que précédemment (les 452 heures de TED-LIUM) auquel on a ajouté un corpus radiophonique **français** (247 heures des corpus ESTER et ETAPE (Gravier *et al.*, 2012). Les phonèmes français et anglais sont appris simultanément : 30 phonèmes français, 39 phonèmes anglais, un modèle de respiration, un modèle de bruit et un modèle de silence. Ce système permet de reconnaître une personne s'exprimant en français ou en anglais et même de changer de langue au cours d'une même phrase.

Le lexique du système de reconnaissance **SRAP_anglais** utilise uniquement des phonème anglais. Le lexique du système **SRAP_anglais-français** contient les mêmes mots avec les mêmes prononciations (anglaises) mais certains mots ont également des prononciations qui utilisent les phonèmes français. Par exemple, pour les noms de villes françaises (comme, par exemple, *Marseille*), nous avons ajouté les prononciations canoniques françaises (avec des phonèmes français). C'est aussi le cas pour les mots transparents, c'est-à-dire les mots qui existent en français et en anglais, comme *ok* ou *possible.*

Les modèles de langage utilisés sont propres aux deux corpus de test : *HIWIRE* et *Aero*. Pour le corpus *HIWIRE*, le modèle de langage est une boucle de mots. Pour le corpus *Aéro*, nous avons appris un modèle tri-gramme sur un corpus textuel de commandes aéronautiques.
Les résultats de reconnaissance seront présentés en terme de taux d'erreur de mots (*Word Error Rate, WER*).

3.3 Protocole expérimental

Les nouvelles règles de prononciations non natives sont générées en utilisant soit le corpus *Nombres* seul, soit le corpus *Nombres* et le corpus *HIWIRE*. Les règles générées sont ensuite appliquées au lexique de départ pour générer des nouvelles prononciations non natives. Puis ce lexique augmenté est utilisé pour effectuer la reconnaissance des phrases du corpus de développement et de test. Pour chaque configuration, nous choisissons les valeurs qui maximisent le taux de reconnaissance sur le corpus de développement pour les paramètres suivants :

- le seuil de la distribution en pourcentage des phonèmes : paramètre α ;
- le seuil de nombre minimal de trames d'apparitions des phonèmes : paramètre β.

4. Résultats

Nous étudions l'impact du lexique de prononciation avec variantes non natives et l'impact des modèles acoustiques. Nous nous intéressons également à la génération des règles en utilisant le corpus de *Nombres* ou le corpus de *Nombres* et le corpus *HIWIRE*. α et β étant optimisés sur un corpus de développement.

Les résultats des expériences sont présentés dans les tables 2 et 3. Dans ces tables « Système de base » désigne le SRAP avec le lexique canonique avant l'enrichissement du lexique. La colonne *Phonèmes* désigne quel ensemble de phonèmes est utilisé pour représenter les phonèmes du lexique et les modèles acoustiques. Notons que le corpus *Aéro* contient la parole de seulement 2 locuteurs, donc il est délicat d'évaluer la variabilité des prononciations générées par notre approche. En revanche ce corpus correspond à de la parole spontanée et son vocabulaire est de 900 mots. De l'autre coté, le corpus *HIWIRE* présente une plus variabilité en termes de locuteurs (31). Mais le vocabulaire de ce corpus est plus petit (130 mots) et la parole est lue. Pour générer les règles de prononciations, pour la Table 2 nous utilisons le corpus *HIWIRE* entier, pour la Table 3 nous utilisons une partie de *HIWIRE* (cf. les explications ci-dessous).

	Phonèmes	*Corpus pour générer les règles*	*Nbr. de nouvelles prononciations*	*Corpus Aero WER (%)*
Système de base	anglais			13,84
	anglais+français			12,43
Approche proposée	anglais	*Nombres*	290	13,6
		Nombres et HIWIRE	290	13,6
	anglais+français	*Nombres*	219	12,38
		Nombres et HIWIRE	1148	**12,09**

Table 2. Résultats de reconnaissance en terme de WER (%). Corpus de test : *Aero*. α et β sont optimisés sur le corpus de développement.

La Table 2 montre que pour le corpus Aéro, utiliser juste les phonèmes anglais ne semble pas améliorer le taux d'erreurs (ligne 3 et 4 de Table 2). En revanche, en utilisant les phonèmes anglais et français notre approche de génération de nouvelles prononciations améliore les performances du SRAP. En générant les règles à partir des corpus *Nombres et HIWIRE,* nous obtenons 12,09 % WER par rapport à 12,43 % obtenu par le système de base.

La Table 3 montre les résultats de reconnaissance sur le corpus *HIWIRE* dans les mêmes configurations que les expériences présentées dans la Table 2. Pour *HIWIRE* nous avons utilisé la technique de validation croisée : nous utilisons la parole de 10 locuteurs pour générer les règles, puis 10 locuteurs comme corpus de développement et puis 11 locuteurs restant pour le test. Puis nous répétons l'opération en sélectionnant une autre répartition de locuteurs. Nous itérons 3 fois.

Dans la Table 3 nous observons que le nombre de règles générées est plus petit que celles données dans la Table 2, car le vocabulaire de *HIWIRE* contient seulement 130 mots par rapport au vocabulaire de 900 mots du corpus *Aero*. Nous observons une légère amélioration de taux d'erreurs de reconnaissance avec notre approche en utilisant les phonèmes anglais : de 9,7 % WER nous passons à 9,5 %. L'utilisation des phonèmes anglais+français n'améliore pas les résultats. Nous allons faire des investigations pour comprendre pourquoi le meilleur résultat n'est pas obtenu avec les phonèmes anglais+français et comment il est possible de mieux sélectionner les règles de prononciations non natives.

	Phonèmes	*Corpus pour générer les règles*	*Nbr. de nouvelles prononciations*	*HIWIRE WER (%)*
Système de base	anglais			9,7
	anglais+français			10,0
Approche proposée	anglais	*Nombres*	19	**9,5**
		Nombres et HIWIRE	35	**9,5**
	anglais+français	*Nombres*	20	9,9
		Nombres et HIWIRE	15	10,0

Table 3. Résultats de reconnaissance en terme de WER (%). Corpus de développement et de test : *HIWIRE (*validation croisée*)*. α et β sont optimisés sur le corpus de développement.

5. Conclusion

Dans cette étude nous avons présenté notre approche de génération des nouvelles prononciations pour l'enrichissement du lexique. Cette approche est conçue pour l'adaptation à la parole de locuteurs non natifs. Les nouvelles prononciations sont générées à partir des règles de prononciation déduites en utilisant un petit corpus représentatif de parole non native. Les règles prennent en compte les graphies des mots et permettent de substituer certains phonèmes par des phonèmes plus appropriés pour la parole non native. L'avantage de notre méthode est que seul le lexique du SRAP est modifié, les modèles acoustiques et le modèle de langage restent inchangés. Les expériences montrent que cette approche est pertinente pour enrichir le lexique de SRAP. Une validation sur un corpus de parole non native plus important serait intéressante. Nous travaillons sur la prise en compte des insertions et des suppressions de phonèmes effectuées par les locuteurs non natifs.

Remerciements

Les auteurs remercient la DGA (Direction Générale de l'Armement), Thales AVS et Dassault Aviation qui soutiennent le financement de cette étude et du programme scientifique «Man-Machine Teaming» dans lequel se déroule ce projet de recherche.

Références

BOUSELMI G. *(2008)*. Contributions à la Reconnaissance Automatique de la Parole Non Native. Thèse Université Lorraine.

BOUSELMI G., FOHR D., ILLINA I., AND HATON J.-P. (*2006)*. Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration and Graphemic Constraints. *In Proc. ICASSP*, France.

DAS A. AND HASEGAWA-JOHNSON M. *(2015)*. Cross-lingual Transfert Learning During Supervised Training in Low Resource Scenarions. *In Proc. of Interspeech.* pp. *3531–3535, 2015.*

DUAN R., KAWAHARA T., DANTSUJI M., AND ZHANG J. *(2017)*. Articulatory Modeling for Pronunciation Error Detection without Non-Native Training Data based on DNN Transfer Learning. *IEICE Transactions on Information and Systems, vol. E100D, no. 9, pp. 2174–2182.*

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A., GALIBERT O. (2012). The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language , *LREC.*

GORONZY S., RAPP S., KOMPE R. (2004). Generating non-native pronunciation variants for lexicon adaptation. *In Journal Speech Communication.*

LEE A. AND GLASS J. *(2015)*. Mispronunciation Detection Without Non-Native Training Data. *in Proc. of Interspeech, pp. 643–647.*

LI W., SINISCALCHI S. M., CHEN N. F., AND LEE C. *(2016)*. Improving Non-Native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-based Speech Attribute Modeling. *in Proc. of ICASSP, pp. 6135–6139.*

LIVESCU K. AND GLASS J. *(2000),* Lexical Modeling of Non-Native Speech for Automatic Speech Recognition, *In Proc. ICASSP.*

MATASSONI M., GRETTER R., FALAVIGNA D., GIULIANI D. (2018). Non-Native Children Speech Recognition Through Transfer Learning. *ArXiv:1809.09658.*

MORGAN J. (2004), Making a Speech Recognizer Tolerate Non-Native Speech Through Gaussian Mixture Merging, *In Proc. InSTIL/ICALL.*

PARK S. AND CULNAN JOHN (*2019)*. A comparison between native and non-native speech for automatic speech recognition.*The Journal of the Acoustical Society of America* 145, 1827.

POVEY D., GHOSHAL A., BOULIANNE G.I, BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G., VESELY K. (*2011)*. The Kaldi Speech Recognition Toolkit, *IEEE 2011 ASRU Workshop.*

PRÉVEIL B., MARTENS J.P., VAN DEN HEUVEL H. (2010). Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon. *In Proc. LREC.*

SCHADEN S. *(2004),* Generating Non-Native Pronunciation Lexicons by Phonological Rule, *InProc ICSLP2004.*

TAN T.-P. (2008), Automatic Speech Recognition for Non-Native Speakers, *thèse de l'université Joseph Fourrier,* 2008.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La phonotaxe du russe dans la typologie des langues : focus sur la palatalisation

Ekaterina Biteeva Lecocq Nathalie Vallée Denis Faure-Vincent
Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
*Institute of Engineering Univ. Grenoble Alpes
ekaterina.lecocq@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr,
denis.faure-vincent@gipsa-lab.fr

RÉSUMÉ

Cet article présente un travail de description phonotactique du russe basé sur une analyse de 15 000 lemmes transcrits phonologiquement et syllabés. Un ensemble de données quantitatives relatives aux structures syllabiques a été examiné dans une perspective typologique. À partir d'une analyse distributionnelle des segments consonantiques ±PAL, des probabilités phonotactiques ont été estimées. Les résultats montrent que le russe suit globalement les tendances générales observées dans les langues de la base de données G-ULSID (Vallée, Rousset & Rossato, 2009) et mettent en évidence des asymétries de distribution des consonnes ±PAL à l'intérieur de la syllabe. Le fait que le système consonantique du russe présente une distinctivité ±PAL étendue à tous les lieux d'articulation, semble contraindre les coccurrences entre consonne et voyelle d'une même syllabe prédites par la théorie Frame/Content (MacNeilage, 1998) et trouvées dans de nombreuses langues.

ABSTRACT

This paper presents a phonotactic description of Russian based on an analysis of 15,000 phonologically transcribed and syllabified lemmas. A set of quantitative data relating to the syllabic structures of Russian has been examined in a typological perspective. From a distributional analysis of ±PAL consonant segments, phonotactic probabilities were estimated. Our results show that Russian broadly follows the general trends observed in the languages of the G-ULSID database (Vallée, Rousset & Rossato, 2009) and highlight asymmetries in the distribution of ±PAL consonants within-syllable units. The fact that Russian presents ±PAL distinctiveness extended to all its consonant places of articulation seems to constrain tautosyllabic consonant/vowel cooccurrences predicted by the Frame/Content Theory (MacNeilage, 1998) and overrepresented in lot of languages.

MOTS-CLÉS : russe, phonotaxe, syllabe, palatalisation, tendances universelles, G-ULSID
KEYWORDS: Russian phonotactics, syllable, palatalization, universal trends, G-ULSID

1 Introduction

La prise en compte des données phonotactiques dans la caractérisation des langues n'est plus à démontrer. Au niveau segmental, les langues ne se distinguent pas seulement dans l'inventaire de leurs systèmes phonologiques, les régularités distributionnelles des segments à l'intérieur des mots et des syllabes participent elles aussi à définir le processus structurel de formation des séquences phonologiques. Ainsi les langues possèdent des propriétés phonotactiques qui règlent les combinaisons de segments dans les syllabes et dans les enchainements de syllabes qui forment les

Volume 1 : Journées d'Études sur la Parole, pages 36–44.
HAL : hal-02798512.

séquences sonores. Les contraintes de positions syllabiques et celles sur la formation des mots, ainsi que les probabilités de cooccurrences entre segments déterminent la phonotaxe des langues (Jusczyk et al. 1994). Nombreuses sont les études qui ont montré un effet de la composante phonotactique de la langue sur le traitement du langage : la phonotaxe, avec ses patrons de régularité, conditionne la perception de la parole (Segui et al., 2002 ; Dehaene-Lambertz et al., 2000), fournit des indices de segmentation du flux de parole (Mattys & Jusczyk, 2001), agit sur les performances de rappel mnésique (Gathercole et al., 1999) et influence la production de la parole (Goldrick & Larson, 2008). La sensibilité aux contraintes phonotactiques de la langue maternelle se mettrait en place aux alentours de 9 mois (Jusczyk & Luce, 1994) et l'effet de probabilité sur le traitement lexical entre 7 et 10 ans (Storkel & Rogers, 2000). Les contraintes de la phonotaxe influencent aussi la phonologisation des emprunts (Kenstowicz, 2010) et on les retrouve également dans les patrons structurels des erreurs de production (Warker & Dell, 2006). Ces études montrent la nécessité de disposer de données de phonotaxe. D'autres travaux montrent également l'intérêt d'utiliser les indices phonotactiques dans le domaine de l'identification (ex. Najafian et al. 2016 ; Srivastava et al. 2017) ou du traitement automatique des langues (ex. Zhu & Adda-Decker, 2006) ou encore de l'apprentissage automatique (Eychenne, 2015). Ce dernier propose d'intégrer au cadre théorique formel Maximum Entropy de Hayes et Wilson (2008) une modélisation des phénomènes phonotactiques basée sur des contraintes pondérées de bonne formation afin de modéliser la grammaire sous forme probabiliste. On comprend alors pourquoi la recherche des régularités phonotactiques s'impose dans la description des langues.

Nous proposons ici une étude de la phonotaxe du russe basée sur une analyse de 15 000 lemmes phonologisés et syllabés de première main. Elle nous permet de présenter un ensemble de généralisations pour le russe que nous abordons dans une perspective typologique en adressant aux contraintes phonotactiques la question de la marque. Notre étude propose également un focus sur les consonnes palatalisées pour lesquelles des désaccords persistent entre les deux écoles phonologiques de Saint-Pétersbourg (ex. Zinder, 1979) et de Moscou (ex. Avanesov, 1956). En russe, contrairement à d'autres langues slaves (Iskarous & Kavitskaya, 2018), la palatalisation connaît un fort rendement phonologique puisque la plupart des consonnes possèdent leur équivalent palatalisé. La palatalisation peut être également l'output d'un processus d'assimilation régressive lorsque $C_{[-PAL]} \rightarrow C_{[+PAL]} / \_ C_{[+PAL]}$. La phonologisation de la palatalisation devant les voyelles coexiste ainsi avec la palatalisation comme résultat d'une assimilation régressive lorsque les consonnes acquièrent une articulation palatale secondaire au contact de $C_{[+PAL]}$. Ce processus allophonique ainsi que les six phonèmes palatals /ʧ ʃ ʃʲ: ʒ (ʒʲ:) j/ rendent le système phonologique du russe très intéressant pour observer, grâce à l'approche typologique, des régularités distributionnelles des consonnes palatalisées considérées comme articulatoirement complexes. À notre connaissance, les travaux antérieurs de description phonotactique du russe ont été établis à partir de deux études (Peshkovskii, 1925 ; Pirogova, 2018) basées sur des types de textes de divers genre et longueur. Au-delà de la recherche d'indices phonotactiques, la comparaison de nos résultats avec les données de ces études nous permet de discuter la corrélation entre les régularités observées dans les entrées lemmatiques d'un lexique référant à une prononciation standardisée et la fréquence des mots en corpus.

2 Matériel et méthode

Le lexique du russe sur lequel s'appuie notre étude contient 15 000 lemmes extraits d'un dictionnaire d'environ 35 000 mots. Ce dictionnaire de fréquences du russe était un des projets de l'Institut russe de recherche en intelligence artificielle conduit d'abord par Sharov (http://www.artint.ru/projects/frqlist.php) puis poursuivi par Lyashevskaya & Sharov (2009). Il a été élaboré sur une collection de textes du corpus national de la langue russe

(http://www.ruscorpora.ru/new/), représentant la langue de la période 1950-2007. Une version électronique du dictionnaire est publiée sur le site de l'Institut de la Langue Russe de V.V. Vinogradov de l'Académie des Sciences (http://dict.ruslang.ru). La liste des items lexicaux est représentative du russe moderne. Il comprend une sélection de prose, de mémoires politiques, de journaux et de littérature scientifique populaire (environ 40 millions de mots). Tous les textes du corpus ont été écrits en russe entre 1970 et 2002.

Pour notre étude, les traitements effectués sur le lexique relèvent du protocole mis en place pour le projet G-ULSID (*Grenoble & UCLA – Lexical and Syllabic Inventory Database*). Ce projet vise à constituer une base de données relationnelles (MYSQL) de lexiques phonologisés et syllabés pour la recherche de régularités dans la phonotaxe des langues du monde qui prennent en compte la structure de la syllabe et les niveaux infra-et supra-syllabiques (Vallée et al., 2009). Chaque entrée lexicale est phonologiquement transcrite et découpée en syllabe(s), et chaque syllabe est décomposée en sous constituants (attaque et rime décomposée en noyau et coda). Dans la continuité de Maddieson & Precoda (1992), seuls les lemmes sont pris en compte et les emprunts récents sont écartés. Par conséquent, les prénoms et patronymes, noms propres, abréviations, interjections et emprunts récents n'ont pas été pris en compte pour le russe. La transcription ne tient pas compte des allophonies et n'ont donc pas été retenus : (1) les cas de palatalisation par anticipation de la première consonne dans les clusters si suivie d'une consonne palatalisée ; (2) le dévoisement final ; (3) les processus de réduction vocalique ; (4) la simplification des groupes de consonnes systématiquement observée à l'oral dans les combinaisons -stn, -zdn, -stl, -ntsk, -stsk, -vstv, -lnts.

La base de donnée est consultable depuis des pages web (via le serveur APACHE et les langages de programmation PHP, HTML et Javascript). Le lexique du russe a été intégré d'abord par une conversion des graphèmes cyrilliques vers les symboles de l'API, traitement effectué par un programme PHP et des requêtes SQL. Une autre application web a permis à un locuteur natif de contrôler le lexique et d'effectuer la syllabation, vérifiée ensuite par deux autres locuteurs natifs. Enfin, la plupart des statistiques et des graphiques de cet article sont issus de l'application web qui analyse les données du russe provenant de la base de données MYSQL.

3 Résultats

Dans une première partie est présenté un ensemble de données quantitatives relatives aux structures syllabiques et lexicales du russe, et replacé dans le contexte de l'étude typologique de Vallée & Rousset (2004) réalisée à partir d'un corpus constitué d'une quinzaine de langues de G-ULSID. Une seconde partie est consacrée à une analyse distributionnelle des segments consonantiques palatalisés (C^j). Elle contient une analyse des occurrences et cooccurrences des consonnes et des voyelles dans les syllabes de types CV, CVC, C^jV, C^jVC, VC, CVC, VC^j et CVC^j à partir desquelles sont estimées des probabilités phonotactiques.

3.1 Données typologiques

Les lemmes de 3 syllabes sont trouvés majoritaires en russe (près de 37 % des entrées), les unités dissyllabiques arrivant au deuxième rang avec 27.36 %. Avec une moyenne de 3.06 syllabes par lemme, le russe se place dans le Type 4 de la typologie des langues de G-ULSID proposée par Vallée & Rousset (2004), à savoir, parmi les langues présentant dans le même temps peu de lemmes monosyllabiques et de plus de 3 syllabes (Figure 1). À noter qu'aucune langue de la base de données ne comporte un mode statistique supérieur à 3.

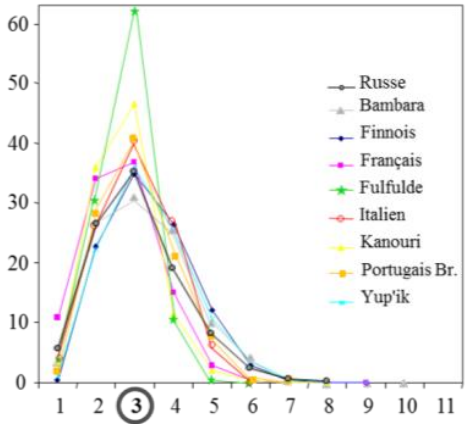


FIGURE 1: Répartition des unités lexicales en fonction du nombre de syllabes qu'elles contiennent. Le russe est ici représenté avec les autres langues de g-ulsid qui présentent un mode de distribution à 3 syllabes. Il s'agit du Type 4 regroupant les langues ayant une majorité d'unités lexicales trisyllabiques (adapté à partir de Vallée, 2017).

La syllabe en russe est majoritairement constituée de 2 segments comme c'est le cas pour la plupart des langues de la base de données, suivie de près par les syllabes à 3 segments (47 % et 42 % respectivement). Les syllabes de plus de 3 phonèmes sont sous-représentées (moins de 8 % du nombre total) ce qui correspond aux tendances générales remarquées pour l'ensemble des langues. Quant au nombre moyen de phonèmes par lemme, celui-ci s'élève à 7.82 (min=5.09, max=7.91), valeurs cohérentes avec celles présentées par les langues des types 3 et 4 de G-ULSID (le type 3 regroupe les langues dont le mode statistique de la distribution des unités lexicales par nombre de syllabes est 2). La loi de Menzerath (1954) décrit le lien qui existe entre le nombre de phonèmes par syllabe et le nombre de syllabes par unité lexicale : le premier a tendance à diminuer lorsque le deuxième augmente. Cette tendance ne se vérifie pas en russe où le nombre de phonèmes par syllabe reste stable et oscille autour de 2,5 quelle que soit la longueur de l'unité lexicale. Le rendement syllabique calculé pour un lexique (ratio entre le nombre total de syllabes et le nombre de syllabes différentes) permet d'évaluer l'efficacité d'une langue dans la combinaison de ses syllabes. Pour le russe, le ratio calculé est égal à 12.81. Cette valeur est cohérente avec celles des langues du type 3 qui ont un rendement entre 5 et 14 (Vallée & Rousset, 2004) soutenant l'idée que plus la proportion d'unités lexicales longues est importante, plus les langues réutilisent des syllabes limitant ainsi la taille de leur inventaire syllabique. Les gabarits lexicaux les plus fréquents en russe sont CV.CVC et CV.CV.CVC avec 828 et 814 entrées. En revanche, le gabarit dissyllabique le plus fréquent CV.CV qui arrive au 1er rang pour les langues du type 3 ou 4 ne représente en russe que 1.45 % des lemmes (sur 15 000).

Les structures syllabiques CV et CVC totalisent respectivement 44 % et 36 % des syllabes et sont de loin les plus fréquentes en russe. Ces deux structures sont également dominantes dans la plupart des langues de G-ULSID. Le russe comporte plus de syllabes à attaque pleine et satisfait ainsi au principe du MOP *Maximal Onset Principal*. En revanche, les structures à attaque et/ou coda branchante(s) sont très peu fréquentes (6 % pour CCV et < 4 % pour les autres). Cela est cohérent avec la relation inverse universelle entre la fréquence d'une structure syllabique donnée et son degré de complexité.

3.2 Consonnes palatalisées et phonotaxe

Dans le lexique analysé, le nombre d'occurrences des consonnes palatalisées (+PAL) est inférieur à celui des non palatalisées (-PAL), modes et lieux d'articulation confondus. Les consonnes +PAL totalisent 31 % (18 271) d'occurrences sur l'ensemble des consonnes -PAL et +PAL et sont moitié moins représentées que les -PAL avec 69 % (41 395) du nombre global d'occurrences (59 666). Plus précisément, l'analyse de leur distribution dans la syllabe indique que les -PAL apparaissent plus souvent dans des attaques et codas simples (respectivement AS et CS) ou complexes (respectivement AC et CC) que leurs contreparties +PAL (Table 2). On note cependant que dans les structures syllabiques avec attaque et/ou coda complexe(s) la proportion de consonnes -PAL et +PAL diminue quels que soient les lieu et mode d'articulation : ex. la différence est de 12 511 occurrences entre AS et AC et de 7 330 entre CS et CC pour les -PAL. Pour les +PAL, la différence est égale à 9 321 entre AS et AC et à 3 818 entre CS et CC.

	AS	AC	CS	CC
-PAL	64 % (20 499)	79 % (7 988)	70 % (10 119)	86 % (2 789)
+PAL	36 % (11 441)	21 % (2 120)	30 % (4 264)	14 % (446)
Total	31 940	10 108	14 383	3 235

TABLE 2 : Distribution des C ±PAL en fonction de la position dans la syllabe

La distribution des consonnes -PAL *vs* +PAL varie également en fonction du lieu articulatoire et une interaction lieu*position à l'intérieur de la syllabe est relevée. Premièrement, les +PAL les plus favorisées sont labiales ou coronales : le ratio -PAL / +PAL est de 3 pour les bilabiales *Bi* et les labiodentales *LDe*, 2 pour les coronales *Co* et 6 pour les vélaires *Ve*. Parmi les +PAL la coronale /tʲ/ est surreprésentée et totalise presque 28 % (5 152) d'occurrences. Au 2ème rang on trouve /lʲ/ suivie des deux autres sonantes coronales /nʲ rʲ/ avec respectivement 14 % (2 544), 12 % (2 238) et 12 % (2 227) du nombre global de +PAL (18 271). Les pourcentages suivants sont basés sur le nombre total d'occurrences pour chacune des catégories : 8 448 occurrences pour *Bi*, 5 254 pour *LDe*, 39 240 pour *Co* et enfin, 6 724 occurrences relevées pour *Ve*.

Bi	*Bi Pal*	*LDe*	*LDe Pal*	*Co*	*Co Pal*	*Ve*	*Ve Pal*
75,44 %	*24,55 %*	*74,34 %*	*25,65 %*	*64,35 %*	*35,64 %*	*87,2 %*	*12,79 %*

Deuxièmement, les labiales *La* (*Bi*+*LDe*) et les *Co* présentent un certain nombre d'asymétries distributionnelles. La Table 3 présente la répartition des consonnes ±PAL en russe en fonction du lieu d'articulation et de la position dans la syllabe.

		AS (%)	AC (%)	CS (%)	CC (%)	Total
Bi	-PAL	41,35	20,71	12,11	1,28	8 448
	+PAL	22,49	1,69	0,36	0,01	
LDe	-PAL	50,00	12,30	10,93	1,12	5 254
	+PAL	21,53	3,86	0,23	0,04	
Co	-PAL	28,52	11,63	18,02	6,18	39 240
	+PAL	19,64	4,12	10,76	1,13	
Ve	-PAL	47,41	15,27	21,58	2,94	6 724
	+PAL	10,47	2,32	0	0	

TABLE 3 : Répartition des C ±PAL selon leur lieu d'articulation et leur position dans la syllabe

Les *La* +PAL occupent assez rarement les positions de CS ou CC contrairement à leurs équivalents -PAL : on retrouve seulement une occurrence de /bʲ/ en CC et une trentaine de *Bi* +PAL en CS sur 8 448 occurrences de bilabiales relevées dans le lexique. Les labiales rencontrées le plus souvent en attaque, simple ou complexe, sont –PAL : 3 493 *vs* 1 900 pour +PAL en AS ; cet écart est plus important encore en AC où les +PAL totalisent seulement 143 occurrences contre 1 750 pour les -PAL. Les *LDe* suivent exactement les mêmes tendances. On note une très faible présence de /f/ et /fʲ/ dans le lexique avec un rendement nul pour /fʲ/ en CS et CC. Les *Co* +PAL sont généralement moins représentées dans le lexique que leurs contreparties -PAL mais on relève cependant quelques exceptions : /tʲ/ est surreprésentée en CS par rapport à /t/: 3 242 contre 941 ; le nombre d'occurrences de /lʲ/ est supérieur à /l/ sauf en AC. Les *Ve* +PAL sont souvent limitées à la position initiale de syllabe devant voyelle et sont plus rares en AC : 153 occurrences sur 602 pour /kʲ/ et seulement 2 (sur 190) et 1 (sur 68) respectivement pour /gʲ/ et /xʲ/. Enfin, nous n'avons pas relevé de *Ve* +PAL en coda, simple ou complexe, dans le lexique. Nous avons calculé les matrices de cooccurrences entre consonnes et voyelles tautosyllabiques pour les structures CV, CVC et VC afin d'estimer la probabilité des segments en séquence en fonction de la présence ou non d'une consonne palatalisée. Ainsi les matrices ont été calculées pour les 7 types de syllabes suivants : CV, CʲV, CVC, CʲVC, CVCʲ, VC et VC ʲ. Des études antérieures ayant montré l'absence de tendance au niveau des modes articulatoires dans

ces types de syllabe (Rousset, 2004 ; Vallée et al., 2009), seuls les lieux d'articulation ont retenu notre intérêt. Les préférences phonotactiques ont donc été observées en regroupant les C ±PAL par lieu d'articulation (La /p b m f v/, Co /t d s z l n r/ ou Ve /k g x/). Ce regroupement a permis de comparer les résultats obtenus pour le russe, à ceux trouvés pour d'autres langues dans des études antérieures. Les résultats présentés Table 4 regroupent les syllabes CVC avec CV, et C^{j}VC avec C^{j}V, car aucun effet de la coda n'a été relevé sur les préférences entre consonne en attaque et noyau vocalique.

	I	A	U	E	O
La	0.98	0.66	0.92	0	1.29
Co	1.29	1.07	0.99	0	0.84
Ve	0	1.35	1.17	0	1.07

	I	A	U	E	O
La	0.71	0.93	0.27	1.35	0.46
Co	1.01	1.15	1.41	0.94	1.34
Ve	1.91	0	0.11	0.27	0

TABLE 4 : Ratio R entre syllabes attestées et attendues : cvc avec cv (à gauche) et c^{j}vc avec c^{j}v (à droite). Les attaques sont en colonnes. R < 1 signifie que les combinaisons ne sont pas favorisées ; R = 1 signifie qu'il n'est pas possible de faire une prédiction et R > 1 indique que la combinaison est favorisée. Les cellules en grisé correspondent aux cooccurrences favorisées en russe.

Un nombre élevé d'occurrences de /s^{j}/ en position d'attaque simple et de syllabe /s^{j}a/ a été relevé dans le lexique. Ce phénomène est lié à la présence des verbes réfléchis formés en ajoutant le postfixe *-s^{j}a* à la base verbale. Le rendement important de cette syllabe (1 241 occurrences) renvoyant à une fonction grammaticale est susceptible d'apporter un biais dans les préférences phonotactiques. Nous avons choisi de l'extraire des calculs sans l'évacuer de notre discussion. Les résultats présentés tiennent compte de cette soustraction. Les résultats sur les préférences entre voyelle et consonne en attaque dans les syllabes de structure CV et CVC montrent que la voyelle centrale est moins favorisée après les attaques labiales qui privilégient d'une part un noyau postérieur et d'autre part /o/ à /u/. Les consonnes *Ve* sont plutôt suivies des voyelles postérieures dans CVC et des postérieures et centrales dans CV. Les combinaisons entre *Co* et voyelles centrales sont privilégiées. Cette tendance est à relier à la centralisation de /i/ réalisé [ɨ] en russe après les consonnes -PAL. En revanche, la tendance coronale-antérieure est portée essentiellement par les syllabes de type C^{j}V et C^{j}VC suggérant que l'assimilation palatale dans un système avec une distinctivité ±PAL étendue à la plupart des unités consonantiques qui le constituent, efface la tendance observée dans les langues du monde. En effet, la matrice montre que lorsque l'attaque est une consonne +PAL, le noyau antérieur est retrouvé dans les combinaisons quel que soit le lieu d'articulation consonantique. L'examen des cooccurrences entre constituants de la rime dans VC, CVC, VC^{j} et CVC^{j}, montre que le nombre de combinaisons possibles entre voyelle et consonne est plus restreint dans une structure VC^{j} que dans CVC^{j}. Cependant, les combinaisons préférées pour les deux structures sont /a/+/l^{j}/ et /i/+/t^{j}/ alors que CVC^{j} favorise /a/+/t^{j}/ et /u/+/t^{j}/. Les rimes avec /a/, /i/ ou /u/ suivi de la coda *Co* +PAL /t^{j}/ relèvent pour l'essentiel des formes verbales infinitives. La structure VC présente quatre séquences saillantes /i/+/s/, /i/+/z/, /o/+/b/, /o/+/t/ attribuables aux suffixes perfectifs.

4 Discussion

Cet article présente des données sur les tendances dans les structures syllabiques et lexicales du russe analysées dans une perspective typologique. Les résultats montrent que le russe suit globalement les tendances générales observées dans les langues de la base de données G-ULSID par Vallée & Rousset (2004), Rousset (2004). Parmi celles-ci citons les syllabes constituées de 2 segments, les unités lexicales formées de 3 syllabes, cependant sans suivre la relation tendancielle entre le nombre de syllabes dans un mot et le nombre moyen de segments qui la compose. Rousset

(2004) confronte la typologie basée sur la répartition du nombre de syllabes par unité lexicale au nombre de phonèmes par syllabe et obtient une deuxième typologie présentant trois patrons différents d'organisation lexicale A, B et C. D'après cette typologie, les langues semblent favoriser les syllabes constituées de 2 segments (type B, le plus représenté) de la même manière qu'elles favorisent les unités lexicales de 2 syllabes. Avec deux langues dissyllabiques, le sora et le suédois, et une langue trissyllabique, le yup'ik, le russe appartient au type C qui regroupe les langues dont le mode statistique de la distribution des syllabes par nombre de phonème est 2, avec une proportion très proche des syllabes de 3 segments, les deux totalisant près de 90 % des syllabes relevées sur l'ensemble du lexique. Le russe est une langue à structure syllabique majoritairement CV (structure canonique universelle), et CVC au 2e rang des fréquences. Les structures se raréfient dans le lexique en fonction du degré de complexité des attaques et/ou des codas (complexité estimée par le nombre de segments qui les composent). Le cadre ou gabarit lexical le plus fréquent en russe, bien que dissyllabique, diffère du gabarit CV.CV le plus fréquemment trouvé dans les langues de G-ULSID par le fait que la deuxième syllabe est composée avec une coda simple. Les résultats obtenus mettent en évidence des asymétries de distribution des consonnes ±PAL à l'intérieur de la syllabe. Ils montrent que quels que soient le lieu d'articulation et la position occupée par ces consonnes dans une structure syllabique donnée, la proportion des +PAL est toujours inférieure à celle des -PAL avec quelques exceptions au niveau des coronales /tʲ lʲ nʲ rʲ/, et des occurrences sous-représentées voire nulles dans le cas des vélaires et des labiales. Le classement de ces dernières par ordre de fréquence décroissant dans le lexique est /k/>/g/>/x/ et /v/>/p/>/m/>/b/> /f/ pour les ±PAL. Il est identique à celui proposé par Pirogova (2018) et à celui de Peshkovskij (1925). Notons un faible nombre d'occurrences voire une absence de /gʲ/ et /xʲ/ et de /fʲ/ en attaque complexe et en coda. La sous-représentation de /f/ et /fʲ/ dans le lexique a une raison historique. D'abord présentes dans les emprunts, elles apparaissent en vieux slave au XIe siècle lorsque /v/, après la chute des deux voyelles faibles antérieure et postérieure, se retrouve en fin de mot ou devant une consonne sourde. Dans les langues slaves et dans certains dialectes du russe, il existe toujours une tendance à éviter [f] en le remplaçant par [p] ou [xv] (Remneva, 2012). La distribution des vélaires trouvée dans notre lexique est aussi un héritage du proto-slave qui n'autorisait pas aux consonnes k, g et x de se trouver devant les voyelles antérieures. Cela reste le cas en russe qui présente pour CV les combinaisons favorisées $Ve_{[+PAL]}$-An et $Ve_{[-PAL]}$-Po. Selon l'école phonologique de Moscou (Avanesov, 1956), [kʲ], [gʲ] et [xʲ] n'ont pas de statut phonémique car elles ne se trouvent jamais en position finale de mot alors que [k], [g] et [x] y sont trouvées. Les oppositions /k/~/kʲ/, /g/ ~/gʲ/, /x/~/xʲ/ sont également marginales devant voyelle (Remneva, 2012). Elles correspondent effectivement dans notre étude aux occurrences les plus faibles en position d'attaque syllabique. Nos résultats confirment ceux de Kochetov (2002) selon lesquels les palatalisations labiales et coronales sont plus fréquentes dans les langues slaves que les palatalisations vélaires. Quant aux coronales, l'ordre de fréquences des segments ±PAL relevé par Pirogova et Peshkovskij diffère de celui trouvé dans notre lexique. Pirogova propose /n/>/t/>/s/>/l/>/r/>/d/>/z/ et idem pour les +PAL. Nous avons obtenu pour les -PAL /n/>/s/>/r/>/t/>/l/>/d/>/z/ et pour les +PAL /tʲ/>/lʲ/>/nʲ/>/rʲ/>/sʲ/>/dʲ/>/zʲ/. La surreprésentation de /tʲ/ dans nos données s'explique par la présence des verbes à l'infinitif en /tʲ/ car notre lexique est composé de lemmes alors que Pirogova et Peshkovskij ont travaillé à partir de textes et donc de formes verbales fléchies. La forte présence de /lʲ/ a également une origine morphologique puisque cette consonne est un élément des suffixes *-/lʲiv/* et *-/alʲ/* qui servent à former des adjectifs et des adverbes. L'explication d'un fort rendement des coronales +PAL se trouve peut-être dans l'évolution du proto-slave. Selon Shevelov (1965), la palatalisation des consonnes peut être reconstruite vers le proto-slave du Ve au VIIIe siècle après JC. À cette époque s'est produite la coalescence des consonnes avec -j. Les combinaisons La+[j] sont devenues La+[lʲ] par ajout d'une liquide épenthétique : pj>plʲ, vj>vlʲ etc. Les autres séquences C+[j] ont fusionné en consonnes palatalisées Cj>Cʲ. Ainsi, les contrastes /l/~/lʲ/, /n/~/nʲ/, /r/~/rʲ/, /t/~/tʲ/ étaient déjà présents dans tous les dialectes du proto-slave. D'autre part, /t/ et /tʲ/ en finale de mot jouent le rôle d'un indice morphologique, en

marquant soit les formes verbales du participe passé passif ou de l'infinitif ; /tʲ/ est aussi un des marqueurs des noms féminins de la 3e déclinaison. Les matrices de cooccurrences obtenues révèlent que les *La* et les *Co* +PAL présentent un certain nombre d'asymétries distributionnelles, tandis que les *Ve* +PAL semblent assez limitées à la position prévocalique de début de syllabe. Selon Kochetov (2002), cette dernière observation est valable pour les autres langues slaves. Le fait que le système consonantique du russe présente une distinctivité ±PAL, pourrait contraindre d'autres cooccurrences entre consonne et voyelle d'une même syllabe que celles prédites par la théorie Frame/Content (MacNeilage, 1998) et trouvées dans des inventaires de langues par MacNeilage & Davis, 2000 ; Vallée et al. 2009) : *La-Ce*, *Co-An* et *Ve-Po*. Ainsi, en russe, nous trouvons /e i/ privilégiées après les consonnes +PAL indépendamment de leur lieu d'articulation. Les *La* et les *Ve* -PAL favorisent les combinaisons avec les noyaux postérieurs. Ce résultat confirme ceux de Kochetov (2002) et Pirogova (2018). Les *Co* ±PAL sont trouvées favorisées dans plusieurs combinaisons avec des noyaux vocaliques comme ce que pointait Pirogova (2018).

Notre corpus de lemmes renvoie aux tendances des autres études basées sur d'autres types de corpus qui permettent de considérer les mots en contexte, ce qui indique que les propriétés phonotactiques du lexique sont proches de celles des mots en contexte. Une transcription qui intègrerait des phénomènes d'allophonie fréquents pourrait sans doute permettre d'affiner les tendances observées à ce jour. Une étude plus complète est en cours avec la recherche de régularités phonotactiques dans les clusters et entre syllabes consécutives. Au-delà d'une description de la phonotaxe du russe, notre étude s'inscrit dans une perspective typologique des propriétés phonotactiques universelles. Dans le cadre de la linguistique russe, les résultats obtenus sont utilisables dans des paradigmes de phonotactique prédictive pour le TAL ou de tests psycholinguistiques.

Remerciements

Ce travail a bénéficié du soutien financier de l'IRS IDEX ComUE UGA projet PALGEST.

Références

AVANESOV R. (1956). *Fonetika sovremennogo russkogo literaturnogo jazyka*. Moscow: MUP.

AVANESOV R. (1972). *Russkoe literaturnoe proiznoshenie*. Moscow: Prosveshchenie.

DEHAENE-LAMBERTZ G., DUPOUX E. & GOUT A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, 12(4), 635-647.

EYCHENNE J. (2015). De l'émergence des contraintes phonotactiques en français. *Langages*, 2, 73-90.

GATHERCOLE S. E., FRANKISH C. R., PICKERING S. J. & PEAKER S. (1999). Phonotactic influences on short-term memory. *J. of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 84-95.

GOLDRICK M. & LARSON M. (2008). Phonotactic probability influences speech production. *Cognition*, 107(3), 1155-1164.

HAYES B. & WILSON C. (2008), "A maximum entropy model of phonotactics and phonotactic learning", *Linguistic Inquiry* 39, 379-440.

ISKAROUS K. & KAVITSKAYA D. (2018). Sound change and the structure of synchronic variability: Phonetic and phonological factors in Slavic palatalization. *Language*, 94, 1-41.

JUSCZYK P. W., LUCE P. A. & CHARLES-LUCE J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630–645.

KENSTOWICZ M. (2010). Loan phonology and Enhancement. In Kang Y.-S. et al., Eds., *Proceedings of the Seoul International Conference on Linguistics, Universal Grammar and Particular*

Languages, p. 104–112, Seoul, South Korea: Linguistic Society of Korea.

KOCHETOV A. (2002). *Production, perception and emergent phonotactic patterns: A case of contrastive palatalization*. New York: Routledge.

LIASHEVSKAIA O. N. & SHAROV S. A. (2009). *Tchastotnyi slovar sovremennogo russkogo iazyka (na materialakh Natsionalnogo korpusa russkogo yazika)*. Moscow: Azbukovnik.

MACNEILAGE P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(04), 499-511.

MACNEILAGE P. & DAVIS B. (2000). On the origin of internal structure of word forms. *Sciences*, 288, 527-531.

MADDIESON I. & PRECODA K. (1992). Syllable structure and phonetic models. *Phonology*, 9, 45-60.

MATTYS S. L. & JUSCZYK P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91–121.

NAJAFIAN M., SAFAVI S., WEBER P. & RUSSELL M. J. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey 2016,* p. 132-139, Bilbao, Spain.

PESHKOVSKII A. M. (1925). Sbornik statei: Metodika rodnogo iazyka, lingvistika, stilistika, poetika. Moscow: Gosizdat.

PIROGOVA N. K. (2018). *Konsonantizm russkogo jazyka*. Moscow: MAKS Press.

REMNEVA M. L. (2012). *Staroslavianskii iazyk*. Moscow: Izdatelstvo Moskovskogo universiteta.

ROUSSET I. (2004). *Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles*. Thèse de doctorat, Université Grenoble III - Stendhal, Grenoble.

SHEVELOV G. Y. (1965). *A prehistory of Slavic: The historical phonology of Common Slavic*. New York: Columbia University Press.

SEGUI J., FRAUENFELDER U. H. & HALLE P. (2002). Les contraintes phonotactiques conditionnent la perception de la parole : implications pour les traitements lexicaux et sous-lexicaux. In E. Dupoux, *Les langages du cerveau*, p. 199-211, Paris : O. Jacob.

SRIVASTAVA B. M. L., VYDANA H., VUPPALA A. K. & SHRIVASTAVA M. (2017). Significance of neural phonotactic models for large-scale spoken language identification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, p. 2144-2151, Anchorage, Alaska, USA: IEEE.

STORKEL H. L. & ROGERS M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. C*linical linguistics & phonetics*, 14(6), 407-425.

VALLEE N. & ROUSSET I. (2004). Indices en typologie des structures lexicales et syllabiques pour la discrimination et l'identification des langues. In *Actes du colloque Identification des langues et des variétés dialectales par les humains et par les machines*, p. 37-42, Paris : ENST.

VALLÉE N., ROSSATO S. & ROUSSET I. (2009). Favoured syllabic patterns in the world's languages and sensorimotor constraints. In Pellegrino F. et al., Eds., *Approaches to Phonological Complexity*, p. 111-139, Berlin : Mouton de Gruyter.

VALLEE N. (2017). *Phonologie et capacités sensorimotrices : de la syllabe au phonème*. Habilitation à diriger des recherches, Université Lumières Lyon 2, Lyon.

WARKER J. A. & DELL G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 387–398.

ZHU D. & ADDA-DECKER M. (2006). Language identification using lattice-based phonotactic and syllabotactic approaches. In *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, p. 1-4, San Juan (Puerto Rico): IEEE.

ZINDER L. R. (1979). *Obshchaia fonetika*. Moscow: Vysshaia shkola.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Débit et réduction vocalique : effets de la tâche de parole et du locuteur.

Angélina Bourbon, *Daria D'Alessandro*, *Cécile Fougeron*
Laboratoire de Phonétique et Phonologie, 19, rue des Bernardins, 75005 Paris, France
angelina.bourbon@sorbonne-nouvelle.fr, daria.dalessandro@sorbonne-nouvelle.fr, cecile.fougeron@sorbonne-nouvelle.fr

RESUME

Dans cette étude nous examinons, sur un groupe varié de 29 locuteurs, les différences de réponses entre locuteur à une demande explicite de modification du débit tout d'abord dans une tâche de répétition rapide, puis entre une tâche de lecture et une tâche de répétition confortable. Ces réponses sont évaluées en termes de débit articulatoire et de réduction vocalique (temporelle et/ou spectrale). Les résultats montrent différents profils de réponses dans la tâche de répétition rapide par rapport à la même tâche sans contrainte temporelle, et on voit que le débit peut être augmenté avec ou sans réduction spectrale. On montre également une forte variation dans les réponses des locuteurs à une tâche de répétition confortable par rapport à de la lecture, avec pour certains locuteurs des différences nettes d'organisation spectro-temporelle. Dans cette tâche assez artificielle de répétition, sans instruction précise, davantage de différences individuelles émergent.

ABSTRACT

Rate and vowel reduction : effects of speech task and speaker.
We investigated on a wide variety of speakers, speaker-dependent responses to an explicit demand of speech rate increase in a fast repetition task, and between a reading and a self-paced repetition task. Responses are tested in terms of articulation rate and temporal and/or spectral vowel reduction. Results show different patterns of response in the fast repetition task compared to the same task without a temporal constraint, and we observe that rate can be increased with or without spectral reduction. Inter-speakers variation in responses to a self-paced repetition task compared to reading is equally showed, with clear-cut differences in the spectro-temporal organization between the two tasks for some speakers. In the absence of precise instruction, more individual differences emerge in this quite artificial task.

MOTS-CLES : variations interlocuteurs ; phonétique acoustique ; débit articulatoire ; réduction vocalique ;
KEYWORDS: interspeaker variation ; acoustic phonetics ; articulation rate ; vowel reduction ;

Volume 1 : Journées d'Études sur la Parole, pages 45–53.
HAL : hal-02798513.

1 Introduction

Le débit de la parole peut être vu comme un aspect variant de la parole, mais aussi comme un facteur de variation, puisqu'il influe sur la précision articulatoire, c'est-à-dire la façon de gérer ses mouvements articulatoires dans le temps et dans l'espace. Nous savons que le débit de la parole varie intrinsèquement entre les individus, les locuteurs adoptent un débit qui leur est propre (Tsao & Weismer, 2006) et celui-ci peut véhiculer des informations sur le locuteur, comme son âge ou son sexe (Jacewicz, 2010) ou sur la présence de certains troubles moteurs de la parole (Weismer, 1995). Nous savons aussi que le débit varie en fonction de la tâche de parole, le débit de parole étant, par exemple, moins rapide dans une tâche contrôlée comme celle de lecture qu'en parole spontanée (Crystal & House, 1988 ; Hirose & Kawanami, 2002). La littérature a montré qu'il peut exister différentes stratégies pour augmenter le débit de la parole et qu'elles peuvent varier en fonction des locuteurs et/ou des tâches ou style de parole (Klatt, 1975; Gay, 1978; Berry, 2011; Rosen, 2011). L'un des effets du débit le plus documenté est la réduction temporelle impliquant une réduction des mouvements, avec un possible *undershoot* des cibles articulatoires et acoustiques (Lindblom, 1963; Moon & Lindblom, 1989; Fourakis, 1991), mais il peut être aussi possible d'augmenter la vélocité des gestes articulatoires pour parvenir à atteindre les cibles articulatoires et/ou acoustiques dans un temps réduit (Kelso, 1987).

Cependant il n'est pas acquis que ces observations dans des tâches de parole ou de 'pseudo-parole' variées, employées dans les différentes études en phonétique tout comme en clinique, soient comparables car pour faire varier la parole des locuteurs en situation expérimentale divers moyens peuvent être utilisés : instructions explicites (ex. demander de parler plus vite/lentement/clairement), de contenu (ex. une phrase, une syllabe, des non-mots) et de complexité (ex. structure syllabique, nombre de répétitions). Dans le cas de la tâche de répétition maximale/rapide de parole, elle s'inspire généralement de la tâche clinique de diadococinésie, impliquant la répétition rapide de mouvements alternants (de la main par exemple). L'adaptation à la parole se fait en demandant au locuteur de répéter le plus rapidement possible une syllabe (ex. /pa/) ou une séquence de syllabes (ex. /pataka/) pendant un laps de temps défini. Outre l'instruction explicite d'augmenter son débit, on cherche à s'assurer que les gestes articulatoires ne soient pas réduits en demandant de préserver une parole intelligible ou une bonne précision articulatoire (par ex. Duffy, 2005). Du fait des nombreux débats existant quant à l'utilisation de ces tâches comme moyen d'évaluation de la 'parole', en raison de leur caractère peu naturel, on préfère les qualifier de tâches de 'performance maximale' (Ziegler, 2003 ; Maas, 2017). Les réponses à cette tâche dans les troubles moteurs de la parole sont variées: ralentissement, compensation sur d'autres dimensions, réduction de l'intelligibilité ou encore désorganisation du débit (Hustad, 2003; Yorkston, 2007; Blanchet & Snyder, 2010). On peut donc se demander si parler et 'performer' en réponse à une instruction est un acte similaire et/ou si tous les locuteurs mettent en place des stratégies similaires face à ce type de tâche expérimentale peu naturelle.

Ainsi, pour mieux comprendre ces variations, cette étude cherche à appréhender les différences dans la gestion spectro-temporelle de la parole en fonction des locuteurs, mais aussi en fonction des tâches. Pour ce faire, nous analysons les réponses d'un groupe varié de locuteurs sur trois tâches de paroles, à savoir la lecture de phrases et la répétition continue d'une phrase, avec ou sans instruction explicite d'augmentation du débit. On regardera comment les locuteurs adaptent

leur débit mais aussi leur 'précision articulatoire' en termes de réduction ou non des caractéristiques spectro-temporelles d'une cible vocalique définie.

2 Méthode

Notre objectif étant de capturer de la variabilité dans les réponses à différentes tâches de parole, nous avons sélectionné une population variée en termes de débit intrinsèque car elle inclue : des locuteurs d'âges très différents, des systèmes vocaliques différents puisqu'elle mélange des hommes et des femmes, pas de contrôle de la variété régionale et probablement une façon différente de réagir à un paradigme expérimental puisque certains locuteurs sont des étudiants aguerris à la parole de laboratoire et d'autres sont des retraités n'ayant jamais fréquenté l'université. Cette population se compose donc de **29 locuteurs francophones natifs**, 10 hommes et 19 femmes, âgés entre 23 et 90 ans (*moy*. = 58,4 (±23,7)). Ils ont été enregistrés sur **trois tâches de parole** dans le cadre d'une collecte de données plus large. Une tâche de **Lecture** où les participants doivent lire 8 phrases différentes constituant une histoire et contenant toutes au moins une fois le mot '*Papa*'. Chaque phrase se répète 3 fois dans un ordre pseudo-aléatoire, soit 24 phrases au total. Deux tâches de répétition où les participants doivent répéter une phrase ('*Papi et Papa papotaient constamment*') de façon continue durant 15 secondes : tout d'abord, aussi vite que possible en restant le plus précis possible (**RepetMax**), puis à un débit confortable et naturel déterminé par le participant (**RepetConfo**).

Les réponses des locuteurs à la tâche sont évaluées sur la production de la voyelle /a/ dans la première syllabe du mot '*p**a**pa*', en termes de variations pouvant être liées à une réduction de la cible vocalique : réduction de durée acoustique et/ou réduction spectrale. La durée totale de la voyelle est mesurée en ms à partir d'une segmentation manuelle (**durée V**) et le **F1** est estimé à partir de la moyenne des valeurs (en Hz) prises sur la portion centrale de la voyelle (au 40-50-60% de la durée totale). Il y a 13 occurrences de /a/ dans la tâche de Lecture et respectivement 4 à 10 et 4 à 13 occurrences dans les tâches RepetConfo et RepetMax, en fonction du nombre de phrases produites par chaque locuteur durant les 15 secondes de répétition demandées. Les productions des locuteurs dans les trois tâches sont aussi comparées en terme de **débit articulatoire**, calculé par le nombre de syllabes produites sur la durée de chaque phrase (pauses exclues). Deux analyses ont été effectuées : l'une comparant les productions des locuteurs entre les tâches RepetConfo et RepetMax pour tester les effets d'une instruction explicite d'augmentation du débit, et l'autre, entre les tâches de Lecture et de RepetConfo pour comparer la répétition (sans contrainte temporelle) à de la parole lue. Les effets de la TACHE (RepetConfo vs RepetMax, puis Lecture vs RepetConfo) et du LOCUTEUR et l'interaction entre TACHE et LOCUTEUR ont été testés avec une série de modèles linéaires mixtes, avec une pente aléatoire par ITEM, sur les variables DEBIT, DUREE V, F1 de /a/. L'effet de chaque facteur et des interactions a été testé en comparant le modèle incluant un certain facteur avec un modèle excluant ce même facteur en effectuant manuellement le Likelihood ratio test (fonction anova()). Des corrélations de pearson entre DEBIT, DUREE V, et F1, ont également été effectuées.

3 Résultats

Un résumé des effets principaux et des interactions est présenté en Table 1. Dans notre première analyse, sur les deux tâches de répétition, nous observons tout d'abord que le débit articulatoire varie en fonction du LOCUTEUR, comme attendu, et aussi en fonction de la TACHE. Toutefois, l'interaction entre les deux facteurs montre que l'augmentation globale de débit dans la tâche RepetMax, illustrée sur la Figure 1, dépend du locuteur. Deuxièmement, nous testons la variabilité des indicateurs de réduction (temporelle et spectrale) de la voyelle /a/ en fonction de la tâche et du locuteur. Sur F1 il y a un effet du LOCUTEUR et une interaction significative entre LOCUTEUR et TACHE. Pour DUREE V, il y a un effet de la TACHE et du LOCUTEUR, ainsi qu'une interaction entre les deux facteurs. Globalement, en RepetMax il y a une tendance à une réduction des voyelles, avec des /a/ plus courts et avec des F1 réduits, mais ces modifications dépendent du locuteur. D'autre part, la réduction sur F1 et la réduction de durée sont modérément corrélées (r = .5) comme montré Figure 2. La réduction spectrale de F1 n'est donc qu'en partie liée à la réduction de la durée de la voyelle. De plus, on remarque une corrélation partielle (r = -0.7) entre la durée des /a/ et le débit articulatoire global sur les phrases. Ceci suggère que les variations de débit ne se manifestent qu'en partie par des variations sur la durée du /a/.

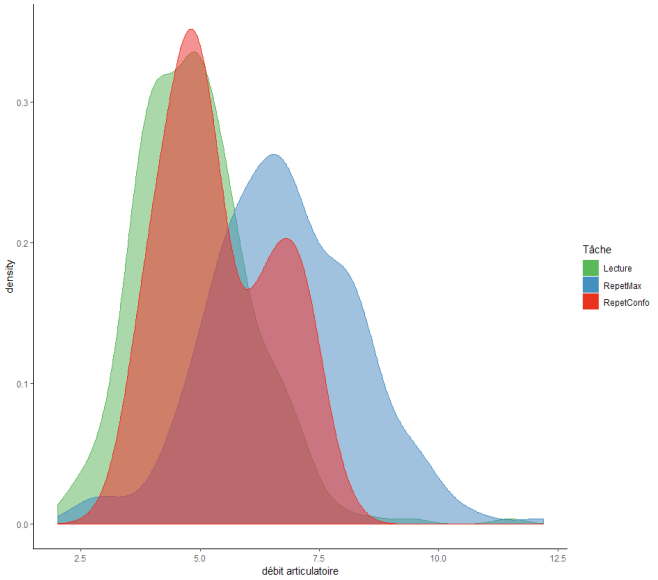


FIG.1 – Distribution du débit articulatoire dans les trois tâches : RepetConfo (rouge), RepetMax (bleue), Lecture (vert).

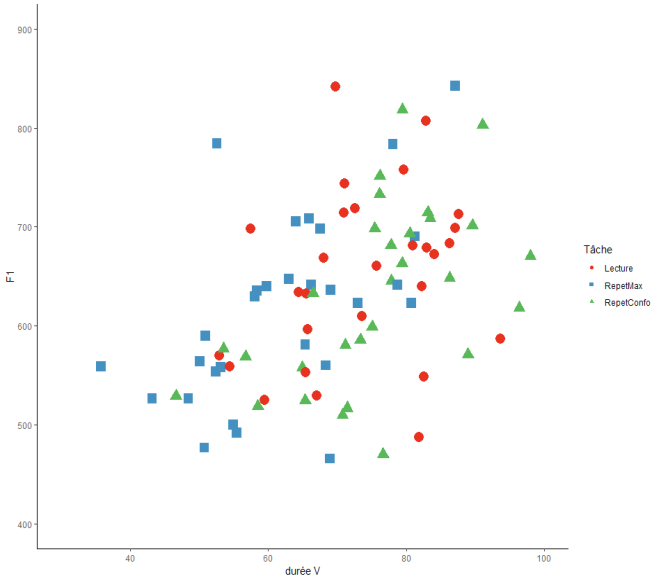


FIG.2 – Représentation par locuteur de la relation entre F1 (y) et durée (x) (moyens par loc.) dans les trois tâches : RepetConfo (ronds), RepetMax (carrés) et Lecture (triangles).

Pris dans leur ensemble, ces résultats montrent donc que les locuteurs réagissent différemment à la demande de modification de la vitesse de la parole (avec maintien de la précision) dans la tâche RepetMax. L'examen des profils des 29 locuteurs fait ressortir trois patterns majeurs, illustrés en Fig.3 avec des locuteurs types. Le premier en (1), 11 locuteurs, présente une

augmentation de débit en RepetMax provoquant une réduction des /a/ au niveau de leur durée, mais pas de réduction spectrale. Le second en (2), 9 locuteurs, présente une augmentation du débit entraînant une réduction temporelle des /a/ et une réduction spectrale. Le troisième en (3), 9 locuteurs, montre une augmentation du débit articulatoire en RepetMax, mais qui n'affecte pas la voyelle /a/, ni dans sa durée, ni sur F1.

TABLE 1: Résumé des effets statistiques. (**=p<.0001, *=p<.01)

Analyse I : RepetConfo vs RepetMax			
	F1	DUREE V	DEBIT
TACHE	ns	$\chi^2(1)=26.192$ **	$\chi^2(1)=34.019$ **
LOCUTEUR	$\chi^2(28)=684.25$ **	$\chi^2(28)=356.9$ **	($\chi^2(28)=503.2$ **
TACHE: LOCUTEUR	$\chi^2(28)=78.21$ **	($\chi^2(28)=148.9$ **	($\chi^2(28)=141.72$ **
Analyse II : Lecture vs RepetConfo			
	F1	DUREE V	DEBIT
TACHE	$\chi^2(1)=17.86$ **	ns	$\chi^2(1)=6.7$ *
LOCUTEUR	$\chi^2(28)=785.99$ **	$\chi^2(28)=459.02$ **	$\chi^2(28)=730.85$ *
TACHE: LOCUTEUR	$\chi^2(28)=107.22$ **	$\chi^2(28)=99.105$ **	$\chi^2(25)=279.39$ *

Dans la deuxième analyse, nous comparons la tâche de répétition à débit confortable (RepetConfo) à la tâche de lecture. La tâche de lecture pouvant être considérée plus proche de la parole naturelle, elle est considérée comme une sorte de *baseline*. Pour F1, les résultats montrent un effet de la TACHE, avec des voyelles globalement réduites dans RepetConfo, mais l'interaction avec LOCUTEUR nous montre que cette réduction ne se réalise pas pour tous les locuteurs. Alors que plusieurs locuteurs présentent des F1 plus bas en RepetConfo qu'en Lecture, d'autres ne montrent pas de différence entre les tâches et certains ont un F1 plus haut. Quant à la durée des voyelles, on voit qu'elle dépend du LOCUTEUR, en interaction avec la TACHE. Selon le locuteur, la durée de la voyelle en RepetConfo se présente comparable, plus courte ou plus longue qu'en lecture. Une corrélation modérée (r = .4) entre réduction spectrale et temporelle nous montre que la durée des voyelles n'explique que partiellement la baisse de leur F1. Enfin, on observe une variation de débit entre les deux tâches, avec une tendance à une augmentation du débit dans RepetConfo qui dépend du locuteur.

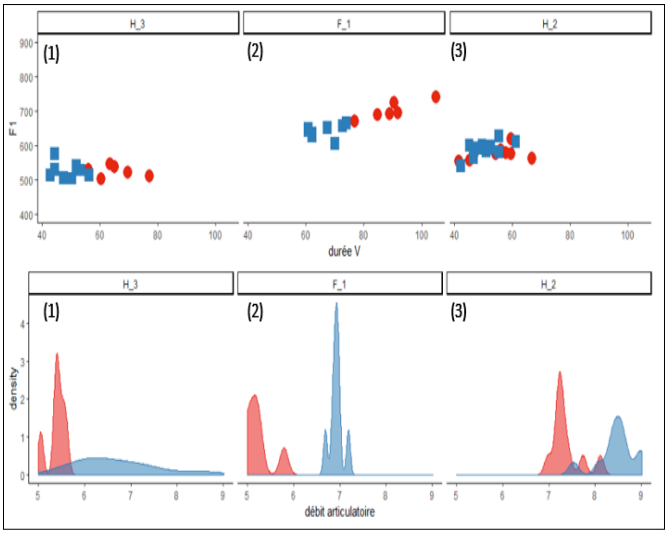


FIG.3 – Illustration sur trois locuteurs types (H_3, F_1 et H_2) des profils de modifications entre RepetConfo (rouge) et RepetMax (bleu). La relation entre F1 et durée de la voyelle /a/ est présentée en haut, et la distribution du débit selon les phrases en bas.

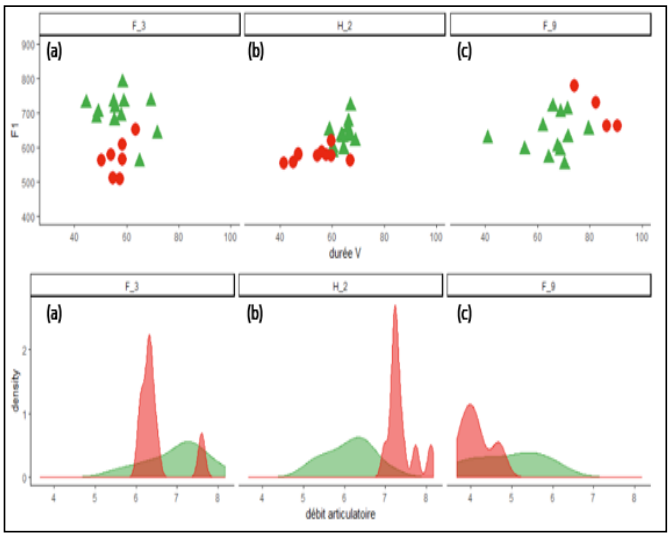


FIG.4 – Illustration sur trois locuteurs types (H_3, H_2 et F_9) des profils de modifications entre les tâches RepetConfo (rouge) et Lecture (vert). La relation entre F1 et durée de la voyelle /a/ est présentée en haut, et la distribution du débit selon les phrases en bas.

La variabilité observée dans les différences de débits adoptés, de durée et de F1 des /a/ entre RepetConfo et Lecture, rend plus difficile l'identification de patterns communs à plusieurs locuteurs. Toutefois, si on considère globalement les résultats sur ces trois variables, on peut dégager trois tendances, illustrées par des exemples individuels en Figure 4 : une tendance à une réduction spectrale de la voyelle, sans changement de durée ou débit en RepetConfo ((a) pour 9 locuteurs) ; une tendance à la réduction spectrale accompagnée d'une réduction temporelle et une augmentation de débit en RepetConfo ((b) pour 5 locuteurs) ; enfin, la tendance majoritaire, une cible vocalique inchangée ou une tendance à l'hyperarticulation avec des voyelles plus ouvertes (F1 plus haut) et plus longues, et un débit inchangé ou plus lent ((c) pour 15 locuteurs).

Discussion et conclusion

Dans cette étude, selon la tâche et le locuteur nous avons observé des changements de débit et de caractéristiques temporelles et spectrales des voyelles. Globalement, les résultats nous montrent que les locuteurs peuvent adopter des stratégies différentes pour répondre aux instructions d'une tâche.

Les résultats montrent que la relation entre réduction temporelle et spectrale pour augmenter le débit de parole dans la tache de répétition rapide est très dépendante du locuteur comme montré par d'autres (Kuehn & Moll, 1976; Gay 1978; Van Son, 1992). Dans nos données, 9 locuteurs sur 29 suivent un pattern répondant à la description de l'*undershoot* de Lindblom (1963) où l'accélération du débit de parole se traduit par un raccourcissement de la voyelle s'accompagnant d'une réduction de F1. Avec un F1 plus bas, les /a/ sont moins ouverts et donc plus réduits. En

revanche, pour 11 locuteurs sur 29, le raccourcissement de la voyelle ne s'accompagne pas d'une réduction spectrale de la voyelle : les voyelles sont courtes mais leur cible acoustique, avec un F1 haut, est atteinte avec précision. Des analyses plus poussées nous permettrons de voir en quoi ces deux profils de locuteurs diffèrent sur d'autres aspects de leur parole. Toutefois, il sera difficile de démontrer si ces deux groupes diffèrent dans leur gestion du compromis entre dimension temporelle et spatiale (ici spectrale) des cibles de parole, ou s'ils ne prennent pas en compte de la même façon l'instruction de la tâche, à savoir : augmenter leur débit tout en préservant la précision de leur articulation. Le troisième pattern observé est très intéressant car il ne correspond pas aux attentes : c'est celui des locuteurs qui augmentent leur débit dans la tâche de répétition rapide, mais qui ne modifient ni la durée, ni le F1 du /a/. Pour ce groupe de 9 locuteurs, il s'agira alors de voir comment l'augmentation de débit est portée par des réductions temporelles (voire spectrales) sur d'autres sons dans la phrase (voyelles et/ou consonnes).

Un second résultat important de cette étude est la différence entre une tâche de répétition à un débit confortable et une tâche de lecture. Le comportement différent dans ces deux tâches va dans le sens des discussions autour du caractère 'naturel' de la parole (Ziegler, 2013 ; Maas 2017). Bien qu'il s'agisse d'une phrase porteuse de sens, le comportement adopté dans la tâche de répétition indique probablement déjà un changement : le type de parole sollicité par le locuteur est moins 'naturel' que lors de la tâche de lecture. Un autre aspect à retenir est le contenu de la phrase à répéter qui comporte une succession de mouvements alternants /pV/. Il est possible que pour optimiser la répétition de ces séquences, même sans instruction de rapidité, les locuteurs bloquent leur mâchoire en position haute pour faciliter les occlusions bilabiales successives. Au contraire, dans la tâche de lecture, les phrases ont un contenu segmental plus varié et une modulation prosodique plus riche, nécessitant des mouvements articulatoires moins contraints. C'est donc une stratégie de 'performance' pour optimiser la répétition qui pourrait expliquer pourquoi dans la tâche RepetConfo, sans changement de durée, 9 locuteurs ont des /a/ spectralement réduits (profil (a)). L'augmentation du débit dans le profil (b) serait alors la conséquence indirecte de cette stratégie : les mouvements articulatoires /pV/ étant plus réduits, les locuteurs parlent plus vite.

Ces patterns pourraient aussi s'expliquer par l'ordre dans lequel les tâches de répétitions sont effectuées : les locuteurs répètent la phrase à débit confortable après la tâche de répétition maximale. Ils pourraient donc avoir reporté leur stratégie de performance mise en place dans RepetMax (augmentation du débit et/ou réduction des cibles) à la tâche de répétition sans contrainte temporelle, comme un *aftereffect*. De la même façon, l'hyperarticulation observée dans le profil (c) pourrait être un *aftereffect* de la consigne visant à un maintien de la précision articulatoire de la tâche RepetMax. En outre, nous pensons que la plus grande difficulté à trouver des patterns communs à plusieurs locuteurs dans la répetition confortable est aussi liée à l'absence de contrainte imposée par la consigne : sans instructions précises de débit et de précision, dans une tâche assez artificielle, les locuteurs adoptent des types de production plus variées que celles qu'ils ont pour la production d'une tâche de répétition avec instruction de débit et précision. Augmenter le nombre de locuteur pourrait apporter une définition plus claire des patterns. Enfin, ces différences dans les résultats montrent l'intérêt de continuer à étudier des tâches sur un continuum de production de la parole plus ou moins *speech-like*, en nuançant l'effet de la tâche par son contenu (phrases, mots, etc) et/ou par sa modalité (narration, répétition, vitesse).

7

Remerciements

Cette étude a été financée par le projet MoSpeeDi - CRSII5_173711/1 du Fond National Suisse de la Recherche Scientifique, par le programme "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL) et par le projet Speech N' Co (ID-RCB: 2019-A02553-54).

Références

BERRY, J. (2011). Speaking rate effects on normal aspects of articulation: Outcomes and issues. *Perspectives on Speech Science and Orofacial Disorders*, 21(1), 15-26. DOI : 10.1044/ssod21.1.15

BLANCHET, P. G., & SNYDER, G. J. (2010). Speech rate treatments for individuals with dysarthria: A tutorial. *Perceptual and motor skills*, 110(3), 965-982. DOI : 10.2466/pms.110.3.965-982

CRYSTAL, T. H., & HOUSE, A. S. (1988). A note on the variability of timing control. *Journal of Speech and Hearing Research,* 31, 497-502. DOI : 10.1044/jshr.3103.497

DUFFY, J. R. (2005). *Motor Speech Disorders: Substrates. Differential Diagnosis, and Management*, 3rd Ed. Mosby, St Louis, MO.

FOURAKIS, M. (1991). Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, 90, 1816-1827. DOI : 10.1121/1.401662

GAY, T. (1978). Effect of speaking rate on vowel formant movements. *The journal of the Acoustical society of America*, 63(1), 223-230. DOI : 10.1121/1.381717

HIROSE, K., & KAWANAMI, H. (2002). Temporal rate change of dialogue speech in prosodic units as compared to read speech. *Speech Communication*, 36(1-2), 97-111. DOI : 10.1016/S0167-6393(01)00028-0

HUSTAD, K.C., JONES, T., DAILEY, S. (2003). Implementing speech supplementation strategies: effects on intelligibility and speech rate of individuals with chronic severe dysarthria. *Journal of Speech, Language, and Hearing Research*, 46(2), 462-74. DOI : 10.1044/1092-4388(2003/er02)

JACEWICZ, E., FOX, R. A., & WEI, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2), 839-850. DOI : 10.1121/1.3459842

KELSO, J. A. S., & TULLER, B. (1987). Intrinsic time in speech production: Theory, methodology, and preliminary observations. In KELLER E. & GOPNIK M. Éds., *Motor and sensory processes of language*, chapitre 8, p. 203-222. Erlbaum. DOI : 10.4324/9780203767702

KUEHN, D. P., & MOLL, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of phonetics*, 4(4), 303-320. DOI : 10.1016/S0095-4470(19)31257-4

KLATT, D. H., & COOPER, W. E. (1975). Perception of segment duration in sentence contexts. In COHEN A., NOOTEBOOM S. G. Éds., *Structure and process in speech perception (pp. 69-89)*. Chapitre 2, p. 69-89, Springer, Berlin, Heidelberg. DOI : 10.1007/978-3-642-81000-8

LINDBLOM, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical society of America*, *35*(11), 1773-1781. DOI : 10.1121/1.1918816

MAAS, E. (2017). Speech and nonspeech: What are we talking about? *International Journal of Speech-Language Pathology,* 19(4), 345-359. DOI : 10.1080/17549507.2016.1221995

MOON, S. J., & LINDBLOM, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical society of America*, 96(1). DOI : 40-55. 10.1121/1.410492

ROSEN, K. M., FOLKER, J. E., MURDOCH, B. E., VOGEL, A. P., CAHILL, L. M., DELATYCKI, M. B., & CORBEN, L. A. (2011). Measures of spectral change and their application to habitual, slow, and clear speaking modes. *International Journal of Speech-Language Pathology*, 13, 165-173. DOI : 10.3109/17549507.2011.529939

TSAO, Y. C., & WEISMER, G. (2006). Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech, Language, and Hearing Research*, 49, 1156-1164. DOI : 10.1044/1092-4388(2006/083)

VAN SON, R.J.J.H. & POLS, L.C.W. (1992). Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 92, 121-127. DOI : 10.1121/1.404277

WEISMER, G., TJADEN, K., & KENT, R. D. (1995). Can articulatory behavior in motor speech disorders be accounted for by theories of normal speech production? *Journal of Phonetics*, 23(1-2), 149-164. DOI : 10.1016/S0095-4470(95)80039-5

YORKSTON, K. M., HAKEL, M., BEUKELMAN, D. R., & FAGER, S. (2007). Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review. *Journal of Medical Speech-Language Pathology*, 15(2), 19-36.

ZIEGLER, W. (2003). Speech motor control is task-specific: evidence form dysarthria and apraxia of speech. *Aphasiology*, 17, 3-36. DOI : 10.1080/729254892

ZIEGLER, W., & ACKERMANN, H. (2013). Neuromotor speech impairment: It's all in the talking. *Folia Phoniatrica et Logopaedica*, 65(2), 55-67. DOI : 10.1159/000353855

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Voice Onset Time en code-switching anglais-français : une étude des occlusives sourdes en début de mot

Marguerite Cameron[1,]
(1) Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins, 75005 Paris, France
marguerite.cameron@sorbonne-nouvelle.fr

Résumé

Le "code-switching" ou l'alternance codique - l'alternance entre plusieurs langues dans une seule interaction - offre une occasion unique d'observer comment les locuteurs multilingues utilisent leurs langues. Récemment, des études phonétiques sur les qualités acoustiques des énoncés code-switchés, telles que le VOT, ont examiné comment les locuteurs équilibrent plusieurs systèmes phonologiques. La présente étude examine les effets du code-switching sur le VOT des occlusives sourdes /p t k/ produites par les locuteurs bilingues anglais-français (L1 anglais et L1 français), d'une acquisition tardive de leur L2. Les données ont été recueillies à partir des enregistrements de discours conversationnels, entre des binômes de participants. Pour les participants francophones (L1 français), le VOT du /p/ des mots anglais était plus long lors d'un code-switch (du français, vers l'anglais) que dans un énoncé tiré d'une conversation monolingue anglais, et le VOT du /t/ était plus court. Aucun effet de contexte (le fait qu'une occlusive mesurée vienne d'un code-switch vers l'anglais ou lors d'une conversation monolingue anglais) n'a été observé pour les anglophones (L1 anglais).

Abstract

Voice Onset Time in English-French code-switching : a study of word-initial voiceless stop consonants
Code-switching – alternating between multiple languages during a single interaction – offers a unique occasion to observe how multilingual speakers use their languages. Recently, phonetic studies on the acoustic qualities of code-switched utterances, such as the VOT, have examined how these speakers balance multiple phonological systems. The present study examines the effects of code-switching on the VOT of the voiceless stop consonants /p t k/, as produced by late-acquisition French-English bilingual speakers (L1 French and L1 English). The data were collected from conversation recordings, between pairs of participants. For the francophone participants (L1 French), the VOT of /p/ in English words was longer during a code-switched utterance (from French, to English) than in an utterance pulled from a monolingual English conversation, and the VOT was shorter. No such effect was observed for the anglophone participants (L1 English).

Mots-clés : code-switching, bilinguisme, VOT
Keywords: code-switching, bilingualism, VOT

1 Introduction

Le code-switching - l'utilisation de plusieurs langues dans un seul énoncé - est un phénomène courant chez les locuteurs multilingues (Bullock & Turibio 2009). Dans les domaines de la syntaxe

Volume 1 : Journées d'Études sur la Parole, pages 54–63.
HAL : hal-02798515.

et de la sociolinguistique, le code-switching a fait l'objet d'études approfondies pendant de nombreuses années et a permis de mieux comprendre la manière dont les locuteurs multilingues utilisent et apprennent leurs langues (Poplack 1980).

Ce n'est que plus récemment, que le code-switching a été étudié sous l'angle de la phonétique et de la phonologie. Les résultats sont, jusqu'à présent, mitigés. La présente étude vise à contribuer à ce domaine de recherche naissant.

Les propriétés acoustiques des énoncés code-switch ont été comparées aux conversations monolingues en mesurant le Voice Onset Time (désormais VOT) – le délai entre le relâchement du conduit vocale lors d'une occlusive sourde et le début du voisement de la voyelle suivante – des consonnes /p t k/ en position initiale de mot. Le VOT est une mesure relativement fiable pour fournir un aperçu sur la parole des bilingues. Les gammes de VOT pour chaqu'une de ces occlusives sont propres à la phonologie de leur langue. C'est à dire que les consonnes homologues entre l'anglais et le français ne manifestent pas les mêmes gammes de VOT. Les mesures de VOT sont à la base d'une forte proportion des études déjà présentes dans la littérature. Ainsi, cette mesure permet d'y relier la présente étude.

1.1 État de l'art

Grosjean & Miller 1994 et Lopez 2012 ont testé des bilingues tardifs et ont tous deux conclu que les bilingues sont capables d'accéder aux deux systèmes phonologiques pendant un code-switch. Grosjean & Miller 1994 ont considéré qu'il n'y avait aucune interférence entre les deux systèmes phonologiques lors d'un code-switch. Lopez 2012 a conclu que même si les divers lieux d'articulation peuvent être plus difficiles à maîtriser, les bilingues forment et utilisent toujours deux catégories phonétiques distinctes pour les phonèmes homologues dans leurs deux langues. D'autres études ont montré des différences de VOT selon le contexte de l'énoncé pour une des langues mais pas l'autre. Les bilingues grec-anglais (grec L1) d'Antoniou et al 2011 ont démontré un VOT en anglais plus court lors des énoncés code-switchs par rapport aux énoncés monolingues. L'effet correspondant n'a pas été observé dans le VOT grec lors des code-switchs. Balukas et Koops 2014 ont également constaté de tels effets unilatéraux chez les personnes bilingues espagnol-anglais. Leur VOT anglais était plus court près d'un code-switch, mais aucun effet sur le VOT espagnol n'a été observé.

Certaines études ont observé des effets dans les deux langues, en fonction de la population. Bullock et al 2006 ont observé chez les bilingues espagnols et anglais que les hispanophones (L1 espagnol) produisaient un VOT plus court en anglais lors des code-switchs (de l'espagnol vers l'anglais), mais ne montraient aucun changement dans leur VOT de code-switch en espagnol (de l'anglais vers l'espagnole). Les anglophones (L1 anglais) ont raccourci leur VOT dans un énoncé code-switch à la fois vers l'anglais et vers l'espagnol. En observant les bilingues espagnol-anglais (L1 espagnol), Olsen 2013 a observé un impact sur la langue dominante lors d'un code-switch.

Piccinini & Arvaniti 2015 a été la principale source d'inspiration pour le protocole utilisé dans la présente expérience. Leur étude de 2015 cherche à résoudre certaines failles des protocoles précédents en examinant la parole spontanée et en contrôlant soigneusement le profil linguistique des locuteurs étudiés. Leur étude examine les effets du code-switching sur le VOT des /p t k/ dans la parole spontanée d'un groupe homogène de personnes bilingues espagnol-anglais, pour lesquels leur L2 (l'anglais) est la langue dominante. Piccinini & Arvaniti 2015 a observé que les occurrences des ces consonnes lors d'un code-switch étaient significativement différents des occurrences d'un

contexte monolingue. Ces locuteurs avaient quand-même maintenu des gammes de VOT distinctes en anglais et en espagnol.

2 La présent étude

2.1 Hypothèses et prédictions

La présente analyse de VOT est basée sur la parole spontanée de binômes de bilingues anglais-français. Les participants sont des bilingues dits « tardifs », c'est-à-dire qu'ils ont acquis leur L2 après la période critique. Néanmoins, ils maîtrisent leur L2 et s'en servent quotidiennement. De manière importante, ces personnes sont de profils linguistiques comparables (âge, âge d'apprentissage de la L2, dialecte d'anglais et de français, niveau d'étude).

Il est prévu que les gammes de VOT des occlusives anglais soient globalement plus longues que les gammes de VOT des occlusives français dans les deux contextes de langue – aux contextes monolingues ainsi qu'aux instances de code- switch – et que cela soit le cas pour les anglophones et pour les francophones. Ceci est conforme aux différences de gammes de VOT dans les deux langues pour les occlusives homologues /p t k/.

En raison de l'acquisition tardive de la L2 par les participants ils gardent un léger accent dans leur L2, et les moyennes du VOT pour les productions en L2 devraient dévier d'une production native. On prévoit que les anglophones auront un VOT plus long en français que les francophones. On s'attend à ce que les francophones aient un VOT plus court en anglais que les anglophones.

L'hypothèse est qu'il y aura effectivement un effet du contexte linguistique. On s'attend à ce que les durées de VOT des occlusives d'une langue ou l'autre soient différentes dans un contexte monolingue ou dans un contexte de code-switch. On s'attend à ce qu'il y ait des interférences phonologiques de la langue matrice sur les énoncés de langue cible, et que les valeurs de VOT pour ces derniers s'écartent des occurrences monolingues de la même langue.

Il est en outre supposé que les différences de moyennes de VOT tirés d'un contexte code-switch montreront un effet de la L1 du locuteur. On s'attend à ce que, lors des code-switchs vers la L2 d'un locuteur depuis une matrice de sa L1, il y aura un écart plus important entre les VOT moyens des occlusives des code-switchs vers la L2 et celles des énoncés monolingues L2 , qu'il y aura entre les VOT moyens des occlusives des code-switchs vers le L1 et celles des énoncés monolingues L1. Autrement dit, on s'attend à ce que les gammes de VOT pour les /p t k/ en L1 soient plus stables dans tout contexte que celles de la L2. Si cet effet se produit, il devrait être observable pour les anglophones et les francophones, étant donné la dominance de leur L1.

2.2 Méthodologie

Le protocole de la présente expérience a été conçu dans le but de susciter un grand nombre d'occurrences de /p t k/ dans des énoncés monolingues et code-switch, dans la L1 et la L2 des locuteurs, à partir de parole spontanée. Les participants connaissaient déjà l'expérimentateur. La langue utilisée pour leur expliquer les tâches a été adaptée à la paire de participants donnée – l'expérimentateur leur parlait dans la langue qui leur était la moins habituelle lors de leurs contacts sociaux, afin d'encourager un cadre bilingue. Ceux ayant l'habitude de parler à l'expérimentateur

en anglais ont été instruits en français et vice versa. L'expérimentateur n'était directement présent que pour expliquer le déroulement de chaque activité.

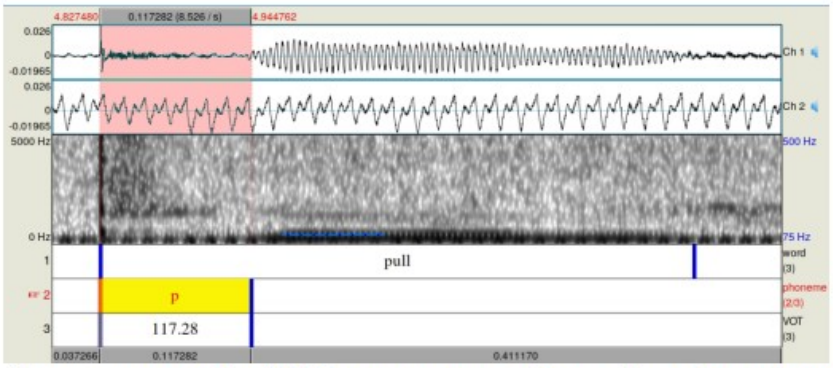
Il y avait deux tâches de durée libre, à faire en binôme : 1. Questions de discussion et 2. Jeu de société. Après avoir compris le déroulement des activités, les participants les effectuaient dans une chambre sourde et signalait à l'expérimentateur quand ils considéraient que l'activité avait suivi son cours.

Des questions de discussion liées au thème du quotidien des bilingues ont été présentées sous forme écrite. Elles sollicitaient des anecdotes liés au bilinguisme et des exemples de code-switchs dont les locuteurs se servent. Les participants ont été invités à utiliser ces questions comme point de départ pour une discussion sur leurs propres expériences. Leurs réponses rapportent des occurrences de /p t k/ exploitables ainsi que plus d'informations sur la population ciblée.

Le jeu de cartes, "Cards Against Humanity" (Cards Against Humanity LLC, 2011), a été présenté dans sa version anglaise et française, l'une après l'autre. Ce jeu contient des cartes de sujet/question auxquelles les joueurs répondent avec une ou plusieurs cartes. Afin de maximiser les occurrences pertinentes, seules les cartes ayant un /p t k/ en début de mot étaient incluses. Les participants ont été instruits sur la règle du jeu et ont été encouragés à commenter leurs observations linguistiques concernant les traductions.

Un entretien individuel entièrement en anglais a été mené pour deux participants francophones, pour lesquels aucune occurrence de /p t k/ en contexte monolingue anglais n'avait été collectée lors de l'enregistrement original. Trois paires de bilingues ont été recrutés pour cette étude, deux anglophones et quatre francophones aux profils linguistiques comparables.

La sélection des occurrences de /p t k/ et les mesures du VOT ont été effectuées manuellement par l'expérimentateur, à l'aide de Praat. Les mots tirés du corpus ont été codés pour : le locuteur, la langue matrice (la langue dominante lors d'une interaction), la langue cible (la langue d'un code-switch), les désignations L1/L2 correspondantes, la transcription orthographique, le phonème mesuré, et le VO (en ms) de ce dernier. Seuls les occlusives initiales sourdes suivis d'une voyelle ont été sélectionnés. Les groupes de consonnes ont été exclus. Toute allophone de / p t k/ permettant une mesure claire du VOT ont été incluses dans la collecte des données, telle que l'occlusive aspirée. Le VOT a été mesuré sur le signal acoustique du relâchement de la consonne jusqu'au début de la périodicité de la voyelle suivante.



Above is an example of VOT measurement in Praat. The token is from Speaker JD, whose signal is on Channel 1.

FIGURE 1: Mesure du VOT

3 Résultats

3.1 Résultats

Les résultats suivants sont basés sur l'ensemble des occurrences de /p t k/ collectés à partir du corpus pour chaque condition linguistique (énoncés monolingues en anglais et en français, et énoncés code-switch en anglais et en français) et pour chaque groupe de participants (anglophones et francophones de langue maternelle).

Des tests de significativité ont été utilisés pour déterminer les éventuels effets de la langue matrice et la langue du code-switch, et de la langue maternelle des locuteurs. Les seules conditions entre lesquelles il y avait des différences significatives dans les valeurs moyennes du VOT se trouvaient entre les occurences d'un contexte anglais monolingue et d'un contexte d'un code-switch vers l'anglais pour les participants francophones. Les participants francophones ont montré dans l'ensemble un VOT plus long pour les occurrences de /p/ dans la condition de code-switch que dans la condition monolingue. Les participants francophones ont montré un VOT globalement plus court pour les occurrences de /t/ dans la condition de code-switch que dans la condition monolingue.

Les conditions comparées dans cette étude sont les suivantes :

- VOT moyen des anglophones pour les énoncés monolingues en anglais vs VOT moyen des francophones pour les énoncés monolingues en anglais
- VOT moyen des anglophones pour les énoncés monolingues en anglais vs VOT moyen des anglophones pour les énoncés code-switch en anglais
- VOT moyen des francophones pour les énoncés monolingues en anglais vs VOT moyen des francophones pour les énoncés code-switch
- VOT moyen pour les énoncés monolingues français pour les anglophones vs VOT moyen des francophones pour les énoncés monolingues en français
- VOT moyen pour les énoncés monolingues français pour les anglophones vs VOT moyen des anglophones pour les énoncés français code-switch

Le VOT moyen des francophones pour les énoncés monolingues en français n'a pas été comparé avec le VOT moyen des francophones pour les énoncés code-switch en français, en raison de la faiblesse des données disponibles.

Des analyses de significativité supplémentaires sont en cours pour les conditions suivantes :

- VOT moyen des anglophones pour les énoncés monolingues en anglais vs VOT moyen pour les énoncés monolingues français pour les anglophones
- VOT moyen des francophones pour les énoncés monolingues en anglais vs VOT moyen des francophones pour les énoncés monolingues en français
- VOT moyen pour les énoncés monolingues en anglais pour tous les locuteurs vs VOT moyen pour les énoncés monolingues français pour tous les locuteurs

3.2 Comparaison du baseline VOT en anglais des francophones et des anglophones

Le VOT moyen en anglais pour les deux groupes de locuteurs (L1 anglais et L1 français) a été déterminé à la base des occurrences de /p t k/ en début de mots en anglais, lors des conversations où l'anglais était la langue matrice

Pour les participants anglophones, 114 occurrences totale de /p t k/ en anglais ont été extraits du corpus pour calculer le VOT de référence. La moyenne générale du VOT était de 50 ms. 30 occurrences de /p/ avec un VOT moyen de 43 ms, 38 occurrences de /t/ avec un VOT moyen de 56 ms, et 46 occurrences de /k/ avec un VOT moyen de 50 ms. Pour les participants francophones, 191 occurrences totale de /p t k/ en anglais ont été extraits du corpus pour calculer le VOT de référence. La moyenne générale du VOT était de 42 ms. 66 occurrences de /p/ avec un VOT moyen de 32ms, 63 occurrences de /t/ avec un VOT moyen de 50ms, et 62 occurrences de /k/ avec un VOT moyen de 46ms.

Les valeurs baseline du VOT ont été comparées à l'aide de T-tests dans R studio. Les moyennes du VOT ont été comparées entre les groupes de locuteurs (anglophones et francophones) pour chaque lieu d'articulation. Aucun test de significativité n'a été effectué pour tous les lieux d'articulation regroupés, car les différents lieux d'articulation présentent généralement des gammes de VOT différentes.

Pour les occurrences de /p/ et /t/, les locuteurs anglophones ont produit un VOT plus long dans leur discours monolingue anglais que leurs homologues francophones. Ce résultat était conforme à la prédiction que les anglophones montreraient des VOT plus longues en anglais que les francophones. Pour les occurrences de /k/, les francophones n'ont pas montré de différence significative par rapport aux anglophones. Cela ne correspond pas à la prédiction selon laquelle les francophones auraient un VOT plus court en moyenne que les anglophones en anglais.

Baseline English VOT – Anglophones vs Francophones

	Anglophones		**Francophones**	
Phoneme	# of Tokens	Mean VOT(ms)	# of Tokens	Mean VOT(ms)
P	30	43	66	32
T	38	56	63	50
K	46	50	62	46
All	114	50	191	42

Phoneme	p-value	Significant	Direction
P	0.00657	Yes	anglo > franco
T	0.04507	Yes	anglo > franco
K	0.2587	No	-

Difference in English Baseline VOT – Anglophones vs Francophones

TABLE 2 : VOT baseline en anglais : anglophones vs francophones

3.3 Comparaison du VOT en anglais des anglophones, entre le contextes monolingue et code-switch

Pour les participants anglophones et pour les participants francophones, les VOT moyens pour chaque lieu d'articulation ont été comparés entre le contexte monolingue et le contexte de code-switch.

Les deux participants anglophones n'ont pas montré de différence significative de contexte dans leur VOT anglais. Au total, 114 occurrences de /p t k/ en anglais d'un contexte monolingue ont été extraits du corpus, avec une moyenne de 50 ms. 30 occurrences de /p/ avec un VOT moyen de 43 ms, 38 occurrences de /t/ avec un VOT moyen de 56 ms, et 46 occurrences de /k/ avec un VOT moyen de 50 ms. Un total de 22 occurrences de /p t k/ des code-switchs anglais (L2-L1) ont été collectés, avec un VOT moyen de 54 ms. 9 occurrences de /p/, avec un VOT moyen de 49 ms, 5 occurrences de /t/ avec un VOT moyen de 66 ms, et 8 occurrences de /k/ avec un VOT moyen de 51 ms. Des tests ont été réalisés dans R Studio pour comparer le VOT moyen dans les différents contextes linguistiques, pour chaque lieu d'articulation séparément.

3.4 Comparaison du VOT en anglais des francophones, entre le contexte monolingue et code-switch

Les participants francophones ont montré des différences significatives de contexte linguistique (entre un contexte monolingue anglais et un code-switch vers l'anglais) pour leurs VOTs moyens en anglais, mais ces différences étaient bidirectionnelles. Le /p/ était significativement plus long lors des énoncés code-switch vers l'anglais. Le /t/ était significativement plus long lors des énoncés monolingues anglais.

Au total, 191 occurrences de /p t k/ ont été collectés dans le contexte monolingue. La moyenne générale du VOT était de 42 ms. 66 occurrences de /p/ avec un VOT moyen de 32ms, 63 occurrences de /t/ avec un VOT moyen de 50ms, et 62 occurrences de /k/ avec un VOT moyen de 46ms. 202 occurrences de /p t k/ en code-switch ont été collectés, soit un VOT moyen de 42 ms. 74 occurrences de /p/ avec un VOT moyen de 40ms, 70 occurrences de /t/ avec un VOT moyen de 43ms, et 58 occurrences de /k/ avec un VOT moyen de 43ms. Ces résultats sont résumés dans les tableaux suivants :

	Baseline (L2)		**Code-Switching (L1 to L2)**	
Phoneme	# of Tokens	Mean VOT(ms)	# of Tokens	Mean VOT(ms)
P	66	32	74	40
T	63	50	70	43
K	62	46	58	43
All	191	42	202	42

Francophone English VOT – Baseline and Code-Switching

Phoneme	p-value	Significant	Direction
P	0.01788	Yes	BL < CS
T	0.03572	Yes	BL > CS
K	0.386	No	-

Difference in Francophone English VOT – Baseline vs Codeswitch

TABLE 3 : VOT en anglais des francophones : contexte monolingue vs contexte code-switch

3.5 Comparaison du baseline VOT en français entre les anglophones et les francophones

Aucune différence significative de VOT lors des interactions monolingues françaises n'a été constatée entre les anglophones et les francophones.

Un total de 76 occurrences de /p t k/ dans un contexte monolingue a été collecté auprès des participants anglophones, avec un VOT moyen de 31 ms. 34 occurrences de /p/ avec un VOT moyen de 25ms, 16 occurrences de /t/ avec un VOT moyen de 35ms, et 26 occurrences de /k/ avec un VOT moyen de 37ms. Des participants francophones, un total de 97 occurrences lors d'un contexte monolingue français ont été extraits, avec un VOT moyen de 34 ms. 30 occurrences de /p/, avec un VOT moyen de 30ms, 27 occurrences de /t/ avec un VOT moyen de 37ms, et 40 occurrences de /k/ avec un VOT moyen de 35ms.

Des tests ont été réalisés dans R Studio entre les participants anglophones et francophones, pour chaque lieu d'articulation. Aucune différence significative n'a été constatée.

3.6 Comparaison du VOT en français des anglophones dans le contexte monolingue français et le contexte d'un code-switch vers le français

La comparaison du VOT moyen entre les conditions des énoncés monolingues et des énoncés code-switch n'a été effectuée que pour les anglophones, en raison de la faiblesse des données pour les francophones.

Un total de 76 occurrences de /p t k/ provenant d'un contexte monolingue français a été tiré du corpus pour les participants anglophones, avec un VOT moyen de 31 ms. 34 /p/ occurrences de avec un VOT moyen de 25ms, 16 occurrences de /t/ avec un VOT moyen de 35ms, et 26 occurrences de /k/ avec un VOT moyen de 37ms. Un total de 36 occurrences de code-switch (L1-L2) a été collecté, avec une moyenne globale de 25 ms. 21 occurrences de /p/ avec un VOT moyen de 18ms, 4 occurrences de /t/ avec un VOT moyen de 26ms, et 11 occurrences de /k/ avec un VOT moyen de 40ms.

Des tests ont été effectués dans R Studio pour chaque lieu d'articulation afin de voir s'il y avait une différence significative entre le VOT moyen français dans un contexte monolingue et dans un contexte code-switch pour les participants anglophones. Aucune différence significative n'a été constatée.

4 Discussion

En réponse à la première grande question de recherche de cette étude -- savoir si les VOT moyens pour les /p t k/ en début de mot diffèrent lorsqu'ils sont prononcés dans un contexte monolingue et lorsqu'ils sont prononcés dans un contexte de code-switch -- il semble que cela puisse être le cas.

Dans la présente étude, les participants francophones ont montré une moyenne de VOT de /p/ et /t/ en anglais significativement différente entre les contextes. La différence n'était cependant pas constante pour les deux consonnes. Les /p/ ont eu un VOT plus long dans le contexte code-switch par rapport au contexte monolingue. Les /t/ ont eu un VOT plus court dans le contexte code-switch que dans le contexte monolingue.

L'absence d'effet général de contexte linguistique sur la moyenne du VOT de l'anglais ou du français pour les participants anglophones est possiblement exagérée par le faible nombre d'occurrences dans le contexte code-switch, mais notre étude ne permet pas d'affirmer cela.

Dans l'ensemble, le protocole de cette étude a été une réussite. Il a atteint l'objectif visé : générer facilement un discours bilingue spontané que l'expérimentateur pouvait qu orienter vers des données pertinentes avec peu d'ingérence.

La comparaison du VOT moyen des occlusives sourdes en début de mot a été une mesure simple pour obtenir une première impression des performances phonologiques des anglophones et des francophones dans les énoncés code-switchés. Il permet également de contribuer à la quantité croissante de données pour ces mêmes types de critères d'étude.

Il existe de nombreux facteurs qui affectent sans aucun doute le VOT et qui n'ont pas été abordés dans la présente étude. Notamment : le débit de parole, la distance (temporelle ou en nombre de mots/phonèmes) entre l'occlusive mésurée et le moment de changement de la langue matrice à la langue cible d'un code-switch, la voyelle suivant l'occlusive, l'accent de la syllabe, l'allophonie. Les facteurs prosodiques, tels que les groupes de respiration, les contours d'intonation ou les pauses, méritent également d'être étudiés.

Références

ADEL, H., VU, N. T., SCHULTZ, T. (2013). Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Vol. 2, pp. 206-211).

ANTONIOU, M., BEST, C. T., TYLER, M. D., KROOS, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of phonetics*, 38(4), 640-653.

ANTONIOU, M., BEST, C. T., TYLER, M. D., KROOS, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. J*ournal of Phonetics*, 39(4), 558-570.

AUER, P. (1995). The pragmatics of code-switching : a sequential approach. One speaker, two languages : Cross-disciplinary perspectives on code-switching, pages 115-135

BULLOCK, B. E. (2012). Phonetic reflexes of code-switching, chapter 10, pages 163-181. Cambridge University Press, Cambridge.

BULLOCK, B. E., TORIBIO, A. J., GONZ´ALEZ, V., DALOLA, A. (2006). Language dominance and performance outcomes in bilingual pronunciation. *In Proceedings of the 8th generative approaches to second language acquisition conference* (pp. 9-16). Somerville, MA: Cascadilla Proceedings Project.

FRICKE, M., KOOTSTRA, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language,* 91, 181-201.

FOWLER, C. A., SRAMKO, V., OSTRY, D. J., ROWLAND, S. A., HALLE, P. (2008). Cross language phonetic influences on the speech of French/English bilinguals. *Journal of Phonetics*, 36(4), 649-663.

GARDNER-CHLOROS, P. (2009A). Code-Switching. Cambridge University Press, Cambridge.

GARDNER-CHLOROS, P. (2009B). Sociolinguistic factors in code-switching. Cambridge University Press.

GOLDRICK, M., RUNNQVIST, E., COSTA, A. (2014). Language switching makes pronunciation less nativelike. *Psychological science,* 25(4), 1031-1036.

GONZ´ALEZ-VILBAZO, K., BARTLETT, L., DOWNEY, S., EBERT, S., HEIL, J., HOOT, B., ... RAMOS, S. (2013). Methodological considerations in code-switching research. *Studies in Hispanic and Lusophone Linguistics*, 6(1), 119-138.

GROSJEAN, F. (1995). A psycholinguistic approach to codeswitching : The recognition of guest words by bilinguals. *One speaker, two languages : Crossdisciplinary perspectives on code-switching, pages 259-275*

NEAREY, T. M., ROCHET, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, 24(1), 1-18.

OLSON, D. J. (2013). Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics*, 41(6), 407-420.

OLSON, D., ORTEGA-LLEBARIA, M. (2010). The perceptual relevance of code switching and intonation in creating narrow focus. *In Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology* (pp. 57- 68).

PICCININI, P. AND ARVANITI, A. (2015). Voice onset time in spanish-english spontaneous code-switching. *Journal of Phonetics*, 52 :121-137

PICCININI, P. E., GARELLEK, M. (2014). Prosodic cues to monolingual versus code-switching sentences in English and Spanish. *In Proceedings of the 7th Speech Prosody Conference* (pp. 885-889).

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Où en sommes-nous dans la reconnaissance des entités nommées structurées à partir de la parole ?

Antoine Caubrière[1] Sophie Rosset[2] Yannick Estève[3] Antoine Laurent[1]
Emmanuel Morin[4]
(1) LIUM, Avenue Olivier Messiaen, 72085 Le Mans ; (2) LIMSI, Rue du Belvedère, 91405 Orsay
(3) LIA, 339 Chemin des Meinajaries, 84140 Avignon ; (4) LS2N, 2 Chemin de la Houssinière, 44322 Nantes
prénom.nom@[univ-lemans ; limsi ; univ-avignon ; univ-nantes].fr

RÉSUMÉ

La reconnaissance des entités nommées (REN) à partir de la parole est traditionnellement effectuée par l'intermédiaire d'une chaîne de composants, exploitant un système de reconnaissance de la parole (RAP), puis un système de REN appliqué sur les transcriptions automatiques. Les dernières données disponibles pour la REN structurées à partir de la parole en français proviennent de la campagne d'évaluation ETAPE en 2012. Depuis la publication des résultats, des améliorations majeures ont été réalisées pour les systèmes de REN et de RAP. Notamment avec le développement des systèmes neuronaux. De plus, certains travaux montrent l'intérêt des approches de bout en bout pour la tâche de REN dans la parole. Nous proposons une étude des améliorations en RAP et REN dans le cadre d'une chaîne de composants, ainsi qu'une nouvelle approche en trois étapes. Nous explorons aussi les capacités d'une approche bout en bout pour la REN structurées. Enfin, nous comparons ces deux types d'approches à l'état de l'art de la campagne ETAPE. Nos résultats montrent l'intérêt de l'approche bout en bout, qui reste toutefois en deçà d'une chaîne de composants entièrement mise à jour.

ABSTRACT

Where are we in Named Entity Recognition from speech ?

Named entity recognition (NER) from speech is usually made through a pipeline process that consists in (i) processing audio using an automatic speech recognition system (ASR) and (ii) applying a NER to the ASR outputs. The latest data available for named entity extraction from speech in French were produced during the ETAPE evaluation campaign in 2012. Since the publication of ETAPE's results, major improvements were done on NER and ASR systems, especially with the development of neural approaches for both of these components. In addition, recent studies have shown the capability of End-to-End (E2E) approach for NER / SLU tasks. In this paper, we propose a study of the improvements made in speech recognition and named entity recognition for pipeline approaches. For this type of systems, we propose an original 3-pass approach. We also explore the capability of an E2E system to do structured NER. Finally, we compare the performances of ETAPE's systems (state-of-the-art systems in 2012) with the performances obtained using current technologies. The results show the interest of the E2E approach, which however remains below an updated pipeline approach.

MOTS-CLÉS : Reconnaissance d'entités nommées structurées, Reconnaissance automatique de la parole, Chaînes de composants, bout en bout.

KEYWORDS: Named Entity Recognition, Automatic Speech Recognition, Pipeline, End-to-End.

Traduction de l'article accepté à LREC 2019 : "Where are we in Named Entity Recognition from speech ?"

Volume 1 : Journées d'Études sur la Parole, pages 64–72.
HAL : hal-02798516.

1 Introduction

La reconnaissance d'entités nommées (REN) consiste en la localisation de concepts, dans des textes non structurés, et en leurs classifications dans des catégories prédéfinies. Le projet Quaero (Grouin *et al.*, 2011) a permis la mise en place d'une définition étendue des entités nommées (EN) dans le cadre de données françaises. Cette définition possède une structure arborescente multiniveau au sein de laquelle différentes EN sont combinés pour définir les plus complexes. Dans le but de mieux décrire les entités nommées, le projet Quaero met en place la notion de composant d'EN. Ainsi, avec cette définition, la REN consiste en la localisation, la classification et la décomposition des entités. La campagne d'évaluation ETAPE (Galibert *et al.*, 2014) a utilisé cette définition étendue.

À notre connaissance, aucun nouveau résultat n'a été publié depuis les résultats de la campagne ETAPE concernant la REN structurées à partir de la parole sur des données françaises. En raison de leurs structures, la reconnaissance de ces types d'EN ne peut pas être abordée comme une simple tâche d'étiquetage de séquences. Lors de la campagne ETAPE, l'état de l'art était construit à l'aide de traitement en plusieurs étapes avant de reconstruire la structure arborescente des EN. Les CRF (Conditional Random Field (Lafferty *et al.*, 2001)) sont au cœur des approches d'étiquetage de séquence. Certaines approches utilisent, en plus des CRF, des grammaires probabilistes sans contexte (Johnson, 1998) (PCFG) leur permettant de mettre en œuvre un modèle en cascade. Les CRF sont appris sur les composants d'EN et les PCFG sont utilisés pour prédire l'ensemble de structure en arbre des EN. Toutefois, le système de REN ayant remporté la campagne ETAPE n'utilise que des CRF avec un modèle par concept (Raymond, 2013). La plupart des approches pour la REN à partir de la parole utilisent une chaîne de deux composants. Tout d'abord un système de reconnaissance de la parole (RAP) produisant des transcriptions automatiques, puis un système de REN est appliqué dessus. Dans cette configuration, le système de REN est appliqué sur des transcriptions automatiques et donc imparfaites. Cela signifie que la qualité des transcriptions a un impact important sur les performances finales de la chaîne (Ben Jannet *et al.*, 2015). En 2012, les systèmes basés sur les modèles de Markov cachés et les modèles à mélange de gaussienne (HMM-GMM) constituaient l'état de l'art en RAP. Depuis, les approches neuronales ont montré leur potentiel (Tomashenko *et al.*, 2016) en se basant sur une combinaison d'HMM et de réseau de neurones profonds (DNN). Les approches neuronales se sont également illustrées dans le cadre de la REN par la combinaison de CRF et de couches neuronales de type bLSTM (Lample *et al.*, 2016; Ma & Hovy, 2016). Dernièrement, une approche de bout en bout (E2E) a été proposée dans (Ghannay *et al.*, 2018) pour la REN directement à partir de la parole. Cette approche laisse un système apprendre l'alignement entre la parole et sa transcription manuelle enrichie avec des concepts d'EN non structurés. D'autres travaux utilisent des approches de bout en bout pour faire correspondre directement la parole aux concepts au lieu de la faire correspondre à des mots, puis ces mots aux concepts (Lugosch *et al.*, 2019). Ces travaux montrent l'intérêt grandissant des approches E2E pour ce type de tâche.

Dans cet article, nous proposons une étude des améliorations récentes pour la tâche de REN dans le cadre de la campagne d'évaluation ETAPE. Nous comparons une approche traditionnelle par chaîne de composants à une approche de bout en bout apprise selon deux stratégies. Notre première contribution consiste en une implémentation en trois étapes permettant d'aborder la REN structurées. Avec cette implémentation, nous séparons la structure arborescente en trois parties distinctes pour les appréhender comme différentes tâches d'étiquetage de séquence plus simple avant de reconstruire la structure arborescente. La seconde est une approche E2E pour la REN structurées. Elle consiste en l'apprentissage de l'alignement entre la parole et les transcriptions textuelles enrichies directement avec les EN structurées.

Nous commençons par une description de la tâche de REN structurées Quaero. Puis, nous décrivons notre implémentation en trois étapes, suivi de la description de nos systèmes état de l'art pour les tâches de RAP / REN dans le cadre d'une chaîne de composant sections 3.2, 3.3 et 3.4, de notre système de bout en bout section 4, de nos jeu de données sections 5 et 6, de nos expérimentation et l'analyse de nos résultats section 7, puis nous concluons.

2 Définition de la tâche

Dans ce travail, nous étudions la REN structurées suivant le formalisme d'annotation Quaero (Rosset *et al.*, 2011). Il permet une annotation suivant huit catégories principales : *amount*, *event*, *func*, *loc*, *org*, *pers*, *prod* and *time*. Ces catégories sont enrichies de sous-types dans le but de créer une hiérarchie. Elle permet de décrire davantage les concepts. Ce guide permet ainsi une annotation selon 39 concepts différents par exemple : *loc.add.phys* qui représente une adresse physique.

En complément des types d'EN, le guide Quaero met en place la notion de composants. Ils permettent de décrire davantage les EN et sont au nombre de 28. Le guide impose que chaque mot présent au sein d'une EN soit annoté en composant, sauf dans le cas de certains articles et mots de liaison. Presque tous les composants dépendent directement du type d'EN, par exemple : "day", "week" sont dépendant du type "time". Cependant certain sont transversaux, par exemple : "kind", "qualifier".

L'annotation finale en EN structurées est arborescente, ce qui signifie qu'une EN peut être composée de composants, mais aussi d'autre EN, sans limites d'imbrication. En suivant la méthode d'annotation proposée par le guide Quaero, la phrase : "la mairie de paris", serait annotée ainsi : "la <**org.adm** <**kind** mairie > de <**loc.adm.town** <**name** paris > > >".

3 Systèmes à chaîne de composants

3.1 Implémentation en trois étapes

Pour réaliser la REN structurées, les systèmes que nous mettons en place utilisent le format d'annotation BIO. Il consiste en un fichier à deux colonnes, la première pour un mot, la seconde pour le concept associé à ce mot. Le concept est préfixé par un "B", un "I" ou un "O", en fonction de la position du mot au sein du groupe de mots correspondant à ce concept. Un inconvénient de ce format est qu'il ne permet pas de représenter efficacement l'arborescence des EN structurées. En effet, cette arborescence implique qu'un mot puisse appartenir à plusieurs concepts. Le format BIO impose un unique label pour chaque mot. Pour représenter l'arborescence, il est nécessaire de concaténer les labels BIO obtenus avec les différents concepts associés à un mot. Nous illustrons cette concaténation dans la figure 1.

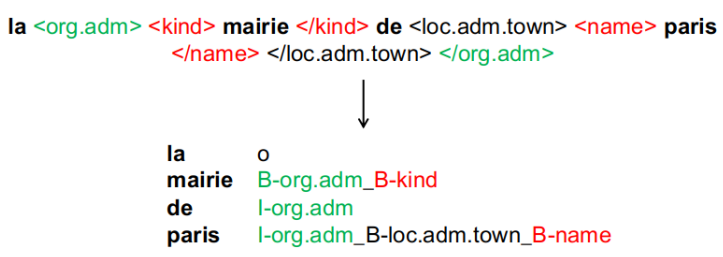


FIGURE 1 – Exemple de transformation d'une séquence d'EN arborescente au format BIO

Ainsi, le nombre final de labels possibles augmente drastiquement. Nous dénombrons désormais 1 690 labels prédictibles. À partir de la notion de composant du guide Quaero, nous pouvons déduire que la racine de l'arborescence est nécessairement un type EN. Nous pouvons également observer que ses feuilles sont en très grande majorité des composants. Enfin, les annotations entre la racine et les feuilles sont un mélange de composants et de type EN. L'augmentation drastique du nombre de labels possibles et ces observations nous motivent à mettre en place une annotation en trois niveaux :

— Le premier niveau représente la racine de l'arborescence. C'est-à-dire les concepts de plus haut niveau. Il est représenté par 96 labels distincts.
— Le troisième niveau représente les feuilles de l'arborescence. Il s'agit ainsi des concepts de plus bas niveau. Un total de 57 labels distincts sont nécessaires pour ce niveau.
— Le second niveau représente l'ensemble des concepts présents entre la racine et la feuille de l'arborescence. Pour ce niveau, nous utilisons 187 labels distincts.

Suite à ce découpage, nous proposons d'effectuer la REN structurées par l'intermédiaire de trois systèmes d'étiquetage de séquences fonctionnant de concert. Nous apprenons un système par niveau. Puis, avec la prédiction de chacun des systèmes nous pouvons reconstruire l'arborescence et ainsi obtenir la séquence finale annotée en EN structurées. Nous souhaitons aussi exploiter les liens existants entre les types EN et les composants. Ainsi, nous proposons d'injecter comme données additionnelles, les prédictions issues d'un modèle précédent dans le/les modèles suivant. C'est-à-dire, injecter les prédictions du premier niveau dans le second et le troisième et injecter les prédictions du second niveau dans le troisième. L'approche que nous proposons est représentée par la figure 2.

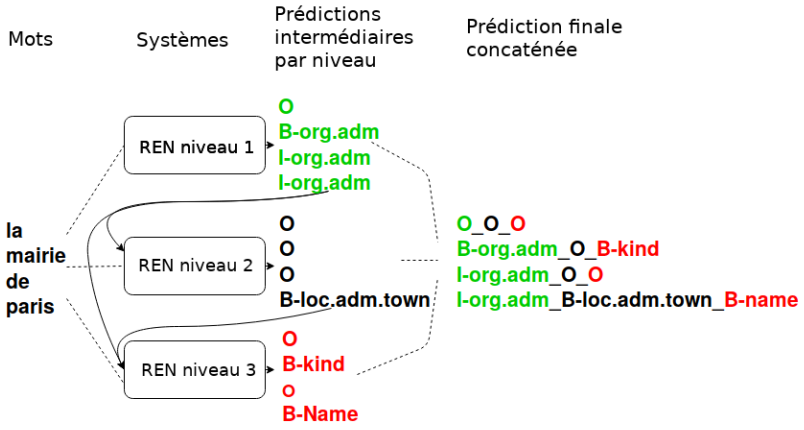


FIGURE 2 – Représentation de notre implémentation en 3 étapes

3.2 CRF

Le premier système de REN utilisé dans ces travaux est un CRF. Nous entraînons les modèles à l'aide du logiciel WAPITI (Lavergne *et al.*, 2010). Ils sont basés sur diverses caractéristiques :

— Les mots et les bigrammes des mots localisés autour des mots cibles sur une fenêtre [-2,+2].
— Les préfixes et les suffixes localisés autour des mots cibles sur une fenêtre [-2,+2].
— Plusieurs caractéristiques de types Oui / Non comme la présence de chiffre dans le mot ou la présence d'une majuscule comme première lettre.

En complément, nous utilisons des caractéristiques morphosyntaxiques extraites de la sortie de l'outil tree-tagger, ainsi que les hypothèses des modèles précédents (implémentation en 3 étapes). Tous les modèles sont entraînés à l'aide de l'algorithme rprop pour un maximum de 40 époques.

3.3 NeuroNlp2

Il s'agit d'un système d'étiquetage de séquences [1] proposé par (Ma & Hovy, 2016). Nous l'exploitons pour réaliser une tâche de REN à partir du texte. Ce système est un empilement de couches CNN, bLSTM et d'une couche CRF. La couche CNN permet l'extraction de plongements de caractères qui s'additionnent aux plongements de mots en entrée des couches bLSTM. Les vecteurs résultant des couches bLSTM sont placés en entrée du CRF finale. Dans nos expérimentations, nous conservons les paramètres par défaut, excepté le nombre de couches bLSTM, paramétré à 2, et le nombre d'unités par couches paramétré à 200.

3.4 Système de reconnaissance de la parole

Dans cette étude, nous utilisons un système de reconnaissance de la parole traditionnel. Ce système est construit à partir de Kaldi et est composé de modèles de Markov cachés et d'un réseau de neurones de types TDNN (Time-Delay Neural Network).

4 Système de bout en bout

L'implémentation [2] que nous utilisons dans cette étude est basée sur le système de RAP DeepSpeech 2 (Amodei *et al.*, 2016). Son architecture consiste en un empilement de deux couches CNN, cinq couches bLSTM et une couche de sortie Softmax. Ce système est entraîné à l'aide de la fonction de coût Connectionist Temporal Classification (Graves *et al.*, 2006). Cette fonction de coût permet au système d'apprendre l'alignement entre un segment audio et une séquence de caractères à produire. Les caractéristiques d'entrées de ce système correspondent aux log-spectrogrammes des segments audio, calculés sur des fenêtres de 20 ms.

Nous mettons en œuvre notre système E2E tel que nous l'avons proposé dans (Ghannay *et al.*, 2018). Les séquences à produire sont des séquences de caractères constituées de mots et des frontières des EN encadrant leurs valeurs. Comme DeepSpeech 2 produit une séquence de caractères, nous représentons ces frontières par l'intermédiaire de caractères uniques. Notre système va ainsi apprendre l'alignement entre des segments audio et des séquences de caractères enrichies des frontières d'EN. L'exemple annoté de la section 2, deviendrait : "la **$ &** mairie > de **% #** paris > > >".

5 Données pour la reconnaissance d'entités nommées

Nos expérimentations sont réalisées sur le corpus français ETAPE (Gravier *et al.*, 2012). Il est composé de données issues d'émissions de radio et TV françaises enregistrées entre 2010 et 2011. Ces émissions proviennent de quatre sources différentes : France Inter, LCP, BFMTV et TV8. Il contient 36 heures de parole séparées en trois parties distinctes : Développement (7 heures), Entraînement (22 heures) et Test (7 heures). Ce corpus possède une annotation en EN selon le guide Quaero. En complément, nous augmentons nos données d'entraînement à l'aide de l'ensemble Quaero (Grouin *et al.*, 2011). Ainsi, nous ajoutons 100 heures de parole entièrement transcrites et annotées en EN manuellement. Ces données sont aussi issues d'émissions de radio et TV françaises.

1. https ://github.com/XuezheMax/NeuroNLP2
2. https ://github.com/SeanNaren/deepspeech.pytorch

6 Données pour la reconnaissance automatique de la parole

Pour effectuer la tâche de RAP, nous utilisons les ensembles ESTER 1 et 2 , REPERE et VERA. Grâce à la combinaison de ces données, nous atteignons un total de 220 heures de parole nous permettant d'apprendre le modèle acoustique du système de RAP de notre approche par chaîne de composants. Le modèle de langage est appris à l'aide des transcriptions manuelles de ces mêmes données enrichies de textes issus de journaux français. Plus de détails sont présents dans la section 4.2.3 de l'article (Deléglise *et al.*, 2009). Notre système E2E est appris sur ces données, en plus de l'ensemble d'apprentissage d'ETAPE.

7 Expérimentations

Nos expérimentations sont évaluées à l'aide de l'ensemble de test d'ETAPE et de la métrique du Slot Error Rate (SER) (Makhoul *et al.*, 1999). Le système de REN remportant la campagne d'évaluation ETAPE (Raymond, 2013) consistait en une combinaison de 68 modèles CRF binaires. Un modèle par type d'EN et par composants. Ce système couplé au meilleur système de RAP permettait d'obtenir un SER de 59,3 %. Ce qui constitue notre référentiel. Afin d'utiliser des transcriptions automatiques annotées en EN structurées, les annotations manuelles d'EN sont projetées dans les transcriptions produites par les systèmes de RAP. Aussi, comme notre système E2E produit à la fois des mots et des EN, nous supprimons les EN produites pour obtenir les transcriptions automatiques. Pour être comparable aux résultats publiés (Galibert *et al.*, 2014), nous utilisons les scripts d'évaluation et de projection de la campagne.

7.1 Expérimentations par chaîne de composants

Nous comparons les résultats obtenus avec l'utilisation du meilleur système de RAP de la campagne (nommé RAP_{2012}) et notre système de RAP état de l'art (nommé RAP_{2020}). Le système RAP_{2012} obtenait un taux d'erreur sur les mots (WER) de 21,8 %, tandis que notre système RAP_{2020} obtient 16,5 %, représentant un gain relatif de 24,3 %. Nous réalisons des expérimentations avec plusieurs combinaisons des systèmes de RAP / REN dont les résultats sont reportés dans la table 1. Il comporte également des expérimentations comparant notre implémentation en 3 étapes (3-pass) avec une approche classique en une étape (1-pass).

TABLE 1 – Résultats expérimentaux des chaînes de composants

Système	**SER**
Sys 0. Baseline ETAPE 2012	59,3
Sys A. 1-pass – CRF – RAP_{2012}	69,4
Sys B. 3-pass – CRF – RAP_{2012}	59,5
Sys C. 3-pass – CRF – RAP_{2019}	55,0
Sys D. 3-pass – bLSTM-CRF – RAP_{2012}	56,1
Sys E. 3-pass – bLSTM-CRF – RAP_{2019}	**51,1**

Le système le plus simple (A) obtient un taux d'erreur de 69,4 %. Lorsque nous employons notre approche en trois étapes dans la même configuration (B), nous atteignons un SER de 59,5 %, soit un gain relatif de 14,3 % par l'utilisation de cette approche. Ces résultats sont proches de ceux de notre référentiel en utilisant uniquement 3 modèles CRF au lieu de 68. Sans surprise, la qualité des

transcriptions automatique améliore les performances globales. Pour un système de REN à base de CRF, les résultats vont de 59,5 % de SER à 55,0 %, soit un gain relatif de 7,6 % (B et C). Avec un système bLSTM-CRF, les résultats vont de 56,1 % à 51,1 % (D et E), permettant un gain relatif de 8,9 %. La mise à jour du système de REN permet également une amélioration. Avec le système de RAP de 2012, les résultats vont de 59,5 % de SER à 55,0 %, soit un gain relatif de 7,6 % (B et D). Tandis qu'avec le système de 2020, les résultats vont de 55,0 % à 51,1 %, soit un gain relatif de 7,1 %. Avec une chaîne de composants, nous obtenons nos meilleurs résultats en mettant à jour chacun des composants et en employant notre implémentation en 3 étapes (système E).

7.2 Expérimentations de bout en bout

Afin d'apprendre notre approche E2E, nous appliquons la même stratégie que dans nos travaux précédents pour compenser le manque de données audio annotées manuellement en EN (Ghannay *et al.*, 2018). Nous effectuons un entraînement multitâche qui consiste tout d'abord à apprendre un système de RAP, puis un système de REN par transfert d'apprentissage ($RAP -> REN_{struct}$). Pour le transfert d'apprentissage, nous conservons le modèle issu de la tâche de RAP, puis nous poursuivons son entraînement en ciblant la tâche de REN. Comme les labels de sorties changent entre les tâches de RAP et de REN, nous réinitialisons la couche de sortie softmax. L'apprentissage du modèle de RAP est effectué à l'aide des données de la section 6 et le modèle de REN à l'aide de celles décrites dans la section 5. Nos travaux précédents ont montré l'intérêt d'un transfert d'apprentissage piloté par une stratégie de curriculum (Caubrière *et al.*, 2019). Cette stratégie consiste à cibler des tâches organisées de celle considérée la plus générique vers celle considérée la plus spécifique. Comme les EN structurées sont composés de type et de composant d'EN, nous proposons d'exploiter cette stratégie. Pour ce faire, nous proposons d'entraîner la tâche de REN en deux apprentissages successifs. Le premier avec des annotations de type EN uniquement, puis le second avec l'annotation complète, incluant donc les composants. Nous supposons les composants comme plus spécifiques que les types EN puisqu'ils en dépendent directement. Ainsi, nous réalisons la chaîne d'apprentissage suivante : $RAP -> REN_{struct} -> REN_{full}$.

Avec notre système E2E, nous pouvons effectuer un décodage classique de type "Greedy", mais aussi un décodage de type "Beam Search" grâce à un modèle de langage. Nous apprenons un modèle de langage 4-gramme à l'aide des données d'apprentissage ETAPE et QUAERO. Les résultats des deux types de décodages pour les deux chaînes d'apprentissages sont donnés dans la table 2.

TABLE 2 – Résultats expérimentaux de l'approche de bout en bout

Système	**ML**	**SER**
$RAP -> REN_{struct}$	X	62,9
$RAP -> REN_{types} -> REN_{full}$	X	61,9
$RAP -> REN_{struct}$	4-gramme	57,3
$RAP -> REN_{types} -> REN_{full}$	4-gramme	**56,9**

Nos résultats montrent l'intérêt de notre stratégie de curriculum pour la REN structurées, par la réduction du SER de 62,9 % à 61,9 %. Ils montrent également l'intérêt du décodage Beam Search qui nous permet d'obtenir de meilleures performances en réduisant le taux de SER de 62,9 % à 57,3 %. La stratégie de curriculum conserve son utilité et nous permet d'obtenir nos meilleurs résultats en atteignant 56,9 % de SER.

7.3 Comparaison globale

Nous reportons les résultats de notre référentiel, de notre meilleur système E2E et de notre meilleur système par chaîne de composant dans la table 3.

TABLE 3 – Résultats reportés du référentiel et de nos meilleurs systèmes

Système	**SER**
(Sys 0) Baseline ETAPE 2012	59.3
$RAP- > REN_{types}- > REN_{full}$ (4-gramme)	56.9
3-passes – bLSTM-CRF – RAP_{2019}	**51.1**

Notre approche E2E nous permet un gain relatif de 4 % par rapport aux résultats de la campagne ETAPE. Toutefois, nos résultats montrent qu'une approche classique avec notre implémentation en 3 étapes et pour laquelle chaque composants est à jour est bien meilleure. Plaçant le nouvel état de l'art à 51,1% de SER. Cette approche classique nous permet un gain relatif de 13,8 % par rapport à notre référentiel. Enfin, en comparant les résultats de nos deux meilleurs systèmes ensemble, nous pouvons observer un gain relatif de 10,2 % à l'avantage de l'approche par chaîne de composants. Pour mieux comprendre les différences entre notre approche E2E et notre chaîne de composants en 3 étapes, nous comparons leurs réponses pour chaque type d'EN. Nous pouvons noter que les cinq concepts les plus représentés (name, kind, pers.ind, name.first, name.last) bénéficient tous d'une amélioration de reconnaissance autour de 16 %. Nous pouvons cependant noter que certains concepts (year, name.nickname) sont désavantagés par notre approche état de l'art, avec respectivement -21 % et -12,3 % de taux de reconnaissance par rapport à notre approche E2E. Ils restent tout de même peu impactant pour les performances globales, car peu représentés (seulement 95 et 65 concepts de références). Il serait nécessaire, dans de futurs travaux, d'effectuer une étude plus approfondie des différences entre nos deux types d'approches.

8 Conclusion

Dans ces travaux, nous présentons les performances désormais atteignables pour la campagne d'évaluation française ETAPE s'étant déroulée en 2012. Nos expérimentations comparent des approches traditionnelles par chaîne de composants et des approches de bout en bout plus récente. Nous proposons dans ce papier une nouvelle implémentation en trois étapes pour la reconnaissance d'entités nommées structurées, dans le cadre des approches par chaîne de composants. En séparant l'arborescence de ces EN en trois parties distinctes, nous sommes capables de réaliser trois tâches plus simples d'étiquetage de séquences. Cette implémentation nous permet d'obtenir des performances similaires au meilleur système de REN en 2012, avec trois CRF classiques au lieu de 68 CRF binaires. Basé sur nos précédents travaux sur les entités nommées non-structurées avec une approche de bout en bout, nous proposons ce type d'approche pour la reconnaissance des entités nommées structurées. Nous exploitons aussi nos précédents travaux sur le transfert d'apprentissage piloté par une stratégie de curriculum pour obtenir nos meilleurs résultats avec une approche de bout en bout. Nous obtenons un gain relatif de 4 % avec ce type d'approche par rapport aux résultats de la campagne ETAPE. Toutefois, nous obtenons nos meilleurs résultats avec une approche par chaîne de composants entièrement mise à jour et exploitant notre implémentation en 3 étapes. Les résultats expérimentaux montrent un gain relatif de 13,8 % entre les résultats de 2012 et le nouvel état de l'art.

Références

AMODEI D., ANANTHANARAYANAN S., ANUBHAI R., BAI J., BATTENBERG E., CASE C., CASPER J., CATANZARO B., CHENG Q., CHEN G. *et al.* (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of ICML'16*, p. 173–182.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015). How to evaluate asr output for named entity recognition ? In *Interspeech*, Dresden, Germany.

CAUBRIÈRE A., TOMASHENKO N., LAURENT A., MORIN E., CAMELIN N. & ESTÈVE Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Interspeech*, Graz, Austria.

DELÉGLISE P., ESTEVE Y., MEIGNIER S. & MERLIN T. (2009). Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate ? In *Interspeech*, Brighton, United Kingdom.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The ETAPE speech processing evaluation. In *Language Resources Evaluation Conference (LREC)*, Reykjavik, Iceland.

GHANNAY S., CAUBRIÈRE A., ESTÈVE Y., CAMELIN N., SIMONNET E., LAURENT A. & MORIN E. (2018). End-to-end named entity and semantic concept extraction from speech. In *IEEE SLT*.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC*, Istanbul, Turkey.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Linguistic Annotation Workshop*, p. 92–100, Portland, OR : ACL.

JOHNSON M. (1998). Pcfg models of linguistic tree representations. *Computational Linguistics*.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Annual Meeting of the Association for Computational Linguistics*, p. 504–513.

LUGOSCH L., RAVANELLI M., IGNOTO P., TOMAR V. S. & BENGIO Y. (2019). Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*, Graz, Austria.

MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv :1603.01354*.

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *DARPA Broadcast News Workshop*, p. 249–252, Herndon, United States.

RAYMOND C. (2013). Robust tree-structured named entities recognition from speech. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, Vancouver, Canada.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). Entités nommées structurées : guide d'annotation quaero. limsi–cnrs, orsay, france.

TOMASHENKO N., VYTHELINGUM K., ROUSSEAU A. & ESTÈVE Y. (2016). Lium asr systems for the 2016 multi-genre broadcast arabic challenge. In *IEEE Spoken Language Technology Workshop*.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

PTSVOX : une base de données pour la comparaison de voix dans le cadre judiciaire

Anaïs Chanclu[1] Laurianne Georgeton[2] Corinne Fredouille[1] Jean-François Bonastre[1]
(1) LIA - Avignon Université
(2) SCPTS - Police Nationale
{anais.chanclu, corinne.fredouille,
jean-francois.bonastre}@univ-avignon.fr,
laurianne.georgeton@interieur.gouv.fr

RÉSUMÉ

Cet article présente la base de données PTSVOX, créée par le Service Central de la Police Technique et Scientifique (SCPTS) spécifiquement pour la comparaison de voix dans le cadre judiciaire. PTSVOX contient 369 locuteurs et locutrices qui ont été enregistrés au microphone et au téléphone. PTSVOX a été conçue pour mesurer l'influence de différents facteurs de variabilité fréquemment rencontrés dans les cas pratiques en identification judiciaire, comme le type de parole, le temps écoulé et le matériel d'enregistrement. Pour cela, 24 des locuteurs de PTSVOX (12 hommes et 12 femmes) ont été enregistrés une fois par mois pendant 3 mois, en parole spontanée et en parole lue. Dans cet article, nous présentons dans un premier temps la base PTSVOX, puis nous décrivons des protocoles standards ainsi que les systèmes de référence associés à PTSVOX, avec une évaluation de leur performance.

ABSTRACT

PTSVOX : a Speech Database for Forensic Voice Comparison

This article introduces PTSVOX, a forensic voice comparison database created by the Service Central de la Police Technique et Scientifique (SCPTS). PTSVOX consists in 369 speakers, recorded using a microphone, and a telephone. The database has been conceived with the purpose of studying the influence of various variability factors which are commonly encountered in practical cases, such as the speaking style, the elapsed time between two recordings and the recording equipment. This is why 24 speakers(12 women and 12 men) have been recorded once a month for three months skeaking spontaneously and reading. In this article, we first present the PTSVOX database. Then, we describe the standard protocoles and the baselines associated with PTSVOX as well as an evaluation of the performance.

MOTS-CLÉS : comparaison de voix dans le cadre judiciaire, variabilité intra-locuteur, reconnaissance du locuteur, base de données PTSVOX.

KEYWORDS: forensic voice comparison, intra-speaker variability, speaker recognition, PTSVOX database.

Volume 1 : Journées d'Études sur la Parole, pages 73–81.
HAL : hal-02798519.

1 Introduction

La reconnaissance du locuteur a réalisé des progrès importants au cours des dernières décennies, notamment grâce à la mise à disposition de grandes bases de données à travers des campagnes d'évaluations comme les campagnes NIST-*Speaker Recognition Evaluation* (SRE) organisées par le *National Institute of Standards and Technology* (NIST). La variabilité inter-locuteur, à la source de la discrimination entre les locuteurs, est bien représentée dans les bases de données type NIST-SRE. Elle est cependant mélangée avec d'autres facteurs de variabilité, comme les accents régionaux, la langue parlée, le microphone ou le bruit ambiant. De nos jours, des taux d'égale erreur (*equal error rate* (EER)) en vérification du locuteur d'environ 1 % sont couramment enregistrés dans les grandes campagnes d'évaluation. Néanmoins, ces campagnes montrent des limites (Kahn *et al.*, 2010) car le nombre d'échantillons par locuteur reste faible alors que le contrôle des facteurs de variabilité est peu réalisé. La variabilité intra-locuteur est peu représentée alors qu'elle est à la fois un problème difficile à résoudre et une limite intrinsèque de la reconnaissance du locuteur. Elle concerne les caractéristiques vocales spécifiques de la personne qui parle mais également des informations telles que la langue, le style de parole, l'état émotionnel, le contenu phonétique (Ajili *et al.*, 2016a) et l'âge (Ajili *et al.*, 2016c; Kahn *et al.*, 2010). Si la mesure de performance autorisée par de telles campagnes et bases de données peut convenir pour des applications commerciales de la reconnaissance du locuteur — pour lesquelles certains facteurs de variabilité peuvent être contrôlés ou compensés — dans certaines applications, comme la comparaison de voix dans le cadre judiciaire, cela n'est pas le cas. Il est donc nécessaire de construire les bases de données et les protocoles adaptés à ce type d'applications.

La comparaison de voix dans le cadre judiciaire est une application spécifique de la reconnaissance du locuteur où deux types d'échantillons de voix sont analysés :

- la pièce de question ou trace, relative à l'enquête, qui représente l'objet de la comparaison de voix ;
- la pièce de comparaison, dont l'origine peut être multiple (mise sous écoute, audition...), peut être un enregistrement prélevé sur un suspect lors d'un entretien.

Dans le cadre judiciaire, la variabilité est une question encore plus prégnante qu'en reconnaissance du locuteur car les conditions d'enregistrements ne sont pas contrôlées, en tout cas pour la pièce de question. Alors que les ressources en vue d'étudier la variabilité inter-locuteurs ne manquent pas, très peu de ressources ont été conçues pour étudier la variabilité intra-locuteur (Ramos *et al.*, 2008; Vloed *et al.*, 2014). La base de données FABIOLE (Ajili *et al.*, 2016b) a été mise en place pour pallier ce manque, mais elle est uniquement composée d'hommes et d'enregistrements issus d'émissions de radio et de télévision, ce qui s'éloigne du contexte de la comparaison de voix judiciaire.

Cet article présente la base de données PTSVOX, qui a été créée dans le cadre du projet ANR-17-CE39-0016 VoxCrim pour correspondre spécifiquement au contexte de la comparaison de voix judiciaire. Les protocoles standards et les systèmes de référence proposés avec PTSVOX sont aussi présentés, ainsi que les résultats expérimentaux correspondants.

2 Description de la base

PTSVOX résulte de campagnes de prélèvement de voix organisées par le Service Central de la Police Technique et Scientifique (SCPTS) dans deux écoles de police situées à Chassieu et à Nîmes. 369 personnes (144 femmes et 225 hommes) ont été enregistrées. Le corpus est composé d'enregistre-

ments téléphoniques et microphoniques réalisés sous forme d'entretien. Les personnes chargées des entretiens ont reçu pour consigne de faire parler les locuteurs autant que possible avec pour objectif de favoriser la spontanéité. Le contenu phonétique et la durée des enregistrements en mode entretien sont donc variables. La majorité des locuteurs n'a été enregistrée qu'une fois alors qu'une sous-partie du corpus, composée de 12 hommes et 12 femmes, a été enregistrée à plusieurs reprises, en Octobre 2016 puis en Mars, Avril et Mai 2017. Pour cette sous partie du corpus, des enregistrements de textes lus ont été également réalisés à chaque session, à l'exception de la première.

2.1 Fichiers audio

Chaque session d'enregistrement contient au moins deux enregistrements de parole spontanée, l'un effectué au microphone et le second au téléphone. La base compte un total de 952 fichiers audio, ce qui correspond à plus de 80 heures de données.

Microphone et téléphones Un enregistreur de type H4n est utilisé, paramétré sur une fréquence d'échantillonnage à 44100 Hz, en stéréo (car il possède deux microphones), avec une résolution de 16 bits. Trois téléphones ont été utilisés pour le prélèvement de voix, un Huawei Ascend Y550 et deux Wiko Cink Slim. Nous avons utilisé l'application Call Recorder, développée par Appliqato, sous la version 4.1.1 d'Android pour enregistrer directement les fichiers audio sur l'appareil. Les enregistrements sont paramétrés sur une fréquence d'échantillonnage de 44100 Hz, en mono, avec une résolution de 16 bits. Cependant, 27 fichiers ont été enregistrés avec une fréquence d'échantillonnage de 8000 Hz.

Ré-échantillonnage des fichiers audio Pour unifier les fichiers, les enregistrements sont également proposés après un ré-échantillonage, en mono avec une fréquence d'échantillonnage à 8000 Hz, 16000 Hz ou 44100 Hz.

2.2 Transcriptions

Les enregistrements ont été transcrits manuellement en utilisant le logiciel Praat (Boersma & Weenink, 2001) pour préparer l'alignement phonétique qui consiste à segmenter le signal de parole en unités minimales, les phonèmes. Cette segmentation en phonèmes est fournie par un outil automatique d'alignement contraint par le texte, développé par le Laboratoire Informatique d'Avignon (LIA). Cet outil prend en entrée le signal de parole, accompagné d'une transcription orthographique du contenu linguistique et un lexique de mots phonétisés (pouvant comporter différentes variantes phonologiques pour un même mot) et fournit en sortie une liste de frontières (début et fin) pour chaque phonème présent dans la transcription. L'alignement phonétique a ensuite été corrigé manuellement par des réservistes citoyennes de la Police nationale. Au total, la base contient 706 560 tokens ("mots" décomposés en chaînes de caractères) et 2 104 237 phonèmes.

2.3 Locuteurs et locutrices

Tous les locuteurs, femmes et hommes, sont des étudiants de l'école de police qui ont signé un formulaire de participation et un formulaire de consentement.

Âge Une très grande majorité des locuteurs, 253, est âgée de 18 à 24. Seuls 5 locuteurs ont plus de 30 ans. Les 111 locuteurs restants ont entre 24 et 30 ans.

Langue maternelle Le français est la langue maternelle de 346 locuteurs. D'autres langues telles que le shimaoré, les créoles guyanais, guadeloupéen et réunionnais sont également mentionnées. Seize locuteurs ont également déclaré avoir le turc, le portugais, le berbère, le malgache, l'arabe, le bushi tongo, le kurde, le guinéen ou l'italien pour langue maternelle.

État de santé Avant chaque session d'enregistrement, les locuteurs devaient indiquer si leur voix était susceptible d'être affectée par une quelconque condition ou état de santé. L'interrogatoire a montré que :

— 130 locuteurs (80 hommes et 50 femmes) ont déclaré fumer ;
— 89 locuteurs ont dit être malade le jour de l'enregistrement (nez bouché, toux, mal de gorge) ;
— 20 locuteurs ont indiqué avoir eu recours à de l'orthophonie ;
— 7 locuteurs ont subi une opération dans la zone ORL.

2.4 Jeux de données

Nous découpons la base PTSVOX en fonction des locuteurs, en trois jeux de données décrits dans le tableau 1. Il n'y a aucun recoupement de locuteurs entre les trois jeux ainsi définis. Chaque enregistrement est découpé en « tours de parole » qui sont ensuite concaténés pour obtenir des $chunks$ d'une durée minimale de 30 secondes.

	Femmes	Hommes	Sessions par locuteur	$chunks$ 30 sec	
				Microphone	Téléphone
$PTSVOX_1$	100	180	1	494	472
$PTSVOX_2$	32	33	1	151	146
$PTSVOX_3$	12	12	2 à 4	194	184

TABLE 1 – Jeux de données de la base PTSVOX

3 Protocoles et systèmes

Cette section présente les protocoles standards définis pour la base PTSVOX ainsi que les deux systèmes de référence associés.

3.1 Protocoles

Dans les protocoles présentés ci-après, nous utilisons les jeux de données décrits dans la section 2.4. Nous utilisons les enregistrements originaux rééchantillonnés à 16000 Hz.

Les protocoles sont définis sous la forme de deux listes de paires de $chunks$ (extraits d'enregistrements audio) de 30 secondes, (a, b), pour la comparaison de voix. Pour les paires $target$, les $chunks$ a et b

proviennent du même locuteur quand pour les paires $nontarget$, les $chunks$ a et b ont été prononcés par des locuteurs différents mais du même sexe.

Pour créer les paires de test $target$, le jeu de données $PTSVOX_3$ est utilisé de deux façons distinctes :

1. $PTSVOX_{3a}$: tous les tests $target$ intrasession sont exclus ;
2. $PTSVOX_{3b}$: composé uniquement des tests intrasession.

Nous distinguons trois protocoles, P_2, P_{3a} et P_{3b} dont les spécificités sont présentés dans le tableau 2 :

1. P_2 : utilise le jeu de données $PTSVOX_2$;
2. P_{3a} : utilise le jeu de données $PTSVOX_{3a}$;
3. P_{3b} : utilise le jeu de données $PTSVOX_{3b}$.

	Microphone			Téléphone		
	$target$ intrasession	$target$ transsession	$nontarget$	$target$ intrasession	$target$ transsession	$nontarget$
P_2	1206	0	1206	1268	0	1268
P_{3a}	0	6484	6484	0	5338	5338
P_{3b}	2738	0	2738	2284	0	2284

TABLE 2 – Nombre de tests par protocole

Les tests $nontarget$ ont été échantillonnés aléatoirement afin de les équilibrer en nombre avec les tests $target$.

3.2 Systèmes de référence

Nous mettons en place deux systèmes de référence, que nous évaluerons avec les trois protocoles mis en place précédemment :

1. $S_{UBM-GMM}$: approche UBM-GMM ;
2. $S_{ivector}$: approche *i-vector*.

En utilisant les approches UBM-GMM et *i-vector* (Dehak *et al.*, 2010), nous construisons deux systèmes : $S_{UBM-GMM}$ et $S_{ivector}$. Ces deux systèmes sont conçus grâce au module LIA/SpkDet, partie intégrante du toolkit open-source ALIZE (Larcher *et al.*, 2013). Les paramètres acoustiques sont composés de 19 MFCC, ses dérivées et dérivées secondes. Une normalisation des paramètres est ensuite appliquée au niveau du fichier.

L'*Universal Background Model* (UBM) possède 512 composants, et est entraîné par un algorithme *Expectation Maximisation* (EM).

Dans le cas du système $S_{UBM-GMM}$, ce modèle générique est entraîné sur le jeu de données $PTSVOX_1$. Plusieurs UBM sont créés en fonction du genre et du matériel d'enregistrement.

Dans le cas du système $S_{ivector}$, ce modèle générique est entraîné sur les données des corpus ESTER (Galliano *et al.*, 2009), ETAPE (Larcher *et al.*, 2013) et REPERE (Giraudel *et al.*, 2012).

Les matrices T de variabilité totale sont également apprises sur ces données. Deux UBM et deux matrices T sont créés en fonction du genre.

Le score délivré lors d'une comparaison de voix est le logarithme du rapport de vraisemblance (*log-likelihood ratio* (LLR)) exprimé par :

$$score = log \frac{P(e_1, e_2|H_p)}{P(e_1, e_2|H_d)} \quad (1)$$

où H_p est l'hypothèse e_1 et e_2 proviennent de la même personne, et H_d est l'hypothèse que e_1 et e_2 ont été prononcés par des personnes différentes.

3.3 Performances

Les tableaux 3 et 4 indiquent les performances pour les deux systèmes de référence en fonction du protocole de test, avec une paramétrisation en bande large 0-8000 Hz. La performance est exprimée en taux d'égale erreur (EER, le taux d'erreur totale est ici le double de l'EER). Pour le système $S_{UBM-GMM}$, les résultats sont donnés pour les UBM appris sur des données téléphoniques et les UBM appris sur des données microphoniques.

		Femmes		Hommes	
		UBM Mic.	UBM Tél.	UBM Mic.	UBM Tél.
P_2	Microphone	1.32%	13.21%	0.66%	8.05%
	Téléphone	7.62%	1.07%	3.32%	1.84%
P_{3a}	Microphone	2.50%	39.02%	2.38%	34.16%
	Téléphone	14.47%	40.33%	11.80%	38.53%
P_{3b}	Microphone	0.60%	9.02%	0.31%	7.43%
	Téléphone	8.08%	6.26%	14.06%	5.18%

TABLE 3 – Performances (EER) pour le système $S_{UBM-GMM}$ avec une bande passante large 0-8000 Hz

		Femmes	Hommes
P_2	Microphone	0.00%	0.00%
	Téléphone	0.00%	0.14%
P_{3a}	Microphone	10.24%	6.44%
	Téléphone	43.52%	40.49%
P_{3b}	Microphone	4.55%	5.15%
	Téléphone	11.41%	9.27%

TABLE 4 – Performances (EER) pour le système $S_{ivector}$ avec une bande passante large 0-8000 Hz (UBM et matrice de variabilité totale T, appris sur ESTER-ETAPE-REPERE)

Les tableaux 5 et 6 indiquent les performances pour une situation comparable aux tableaux 3 et 4 mais avec une paramétrisation en bande passante étroite, 300-3400 Hz.

		Femmes		Hommes	
		UBM Mic.	UBM Tél.	UBM Mic.	UBM Tél.
P_2	Microphone	1.16%	1.99%	1.50%	2.82%
	Téléphone	4.46%	1.07%	4.38%	2.12%
P_{3a}	Microphone	7.82%	12.94%	2.18%	7.35%
	Téléphone	23.55%	20.41%	20.92%	21.47%
P_{3b}	Microphone	2.66%	6.79%	0.44%	3.37%
	Téléphone	5.56%	1.92%	6.96%	5.26%

TABLE 5 – Performances (EER) pour le système $S_{UBM-GMM}$ en bande étroite 300-3400 Hz

		Femmes	Hommes
P_2	Microphone	0.00%	0.17%
	Téléphone	0.00%	0.00%
P_{3a}	Microphone	27.16%	9.08%
	Téléphone	29.59%	32.20%
P_{3b}	Microphone	10.82%	5.33%
	Téléphone	8.28%	6.18%

TABLE 6 – Performances (EER) pour le système $S_{ivector}$ en bande étroite 300-3400 Hz (UBM et matrice de variabilité totale T, appris sur ESTER-ETAPE-REPERE)

Le tableau 7 montre la performance du système $S_{UBM-GMM}$ en utilisant l'UBM du système $S_{ivector}$, appris sur ESTER-ETAPE-REPERE, en bande étroite.

		Femmes	Hommes
P_2	Microphone	2.48%	2.16%
	Téléphone	2.16%	5.51%
P_{3a}	Microphone	17.94%	5.73%
	Téléphone	26.74%	21.93%

TABLE 7 – Performances (EER) pour le système $S_{UBM-GMM}$ en utilisant un UBM appris sur ESTER-ETAPE-REPERE, avec une bande passante étroite 300-3400 Hz

4 Discussion

Pour le protocole P_2 pour lequel il n'y a pas de paires $target$ inter-session, la performance obtenue par les deux systèmes de référence est bonne pour toutes les situations. Le système $S_{ivector}$ montre une supériorité, malgré le fait que ses données d'apprentissage ne viennent pas de PTSVOX. Pour

le protocole P_{3a}, qui ne contient que des paires inter-sessions, deux phénomènes sont observés. D'une part, la performance diminue légèrement comparativement à P_2 en données microphone pour $S_{UBM-GMM}$ et plus fortement pour $S_{ivector}$. D'autre part, pour les données téléphone, la performance s'écroule complètement pour les deux systèmes et se rapproche du hasard. Une hypothèse plausible pour expliquer ces résultats est la présence d'un facteur de variabilité important entre les différentes sessions d'enregistrement "téléphone" : les systèmes reconnaissent autant la session que le locuteur. Les résultats obtenus en utilisant le protocole P_{3b} confirment cette hypothèse : le niveau de performance est alors proche de celui obtenu pour P_2.

Il est difficile d'expliquer cette importante variabilité entre sessions téléphoniques autrement que par un artefact technique. L'analyse spectrale de plusieurs fichiers téléphoniques du protocole P_{3a} montrent une réflection du spectre entre 4000 Hz et 5000 Hz , ce qui pourrait expliquer les résultats observés précédemment. Les expériences en bande étroite confirment que si cette réflection est une part du problème, d'autres facteurs existent également : en bande 300-3400 Hz, alors que la réflexion est exclue, l'écart de performance entre P_{3a} et P_{3b} se réduit mais reste marqué, quel que soit le système employé (une expérience complémentaire avec une bande passante 0-4000 Hz a montré des résultats similaires). Les particularités ou des différences de configuration du logiciel d'enregistrement Call Recorder sont une deuxième source plausible de cet effet session. Par ailleurs, sans surprise, en données microphone, une dégradation des performances est observée quand la bande passante est réduite, dégradation logiquement beaucoup plus forte pour les voix de femmes que pour les voix d'hommes.

Le système $S_{UBM-GMM}$ utilisant un UBM appris sur d'autres données que PTSVOX (tableau 7) montre une perte de performance comparativement à un UBM appris sur PTSVOX (tableau 5) pour les données microphone mais une amélioration de la performance pour les données téléphone, ce qui renforce l'idée que les données téléphone de PTSVOX contiennent un facteur de variabilité important et rarement rencontré. Ceci explique également le très bon, voire trop bon, niveau de performance observé avec le protocole P_2 : l'information sur la session d'enregistrement est tellement marquée qu'elle est peut-être plus saillante même que l'information locuteur.

5 Conclusion et perspectives

Dans cet article, nous présentons la base PTSVOX, dédié à la comparaison de voix judiciaire. Le corpus contient des enregistrements téléphoniques et microphones de parole spontanée et de lecture provenant de 369 personnes, représentant environ 80 heures au total. La base a été transcrite orthographiquement et alignée phonétiquement. Trois sous corpus et trois protocoles expérimentaux ont été définis, ainsi que deux systèmes de comparaison de voix de référence. Les expériences menées avec ces protocoles et systèmes ont démontré la bonne qualité des systèmes de référence. Elles ont permis de repérer un effet session prononcé et non audible, indicateurs d'artefacts, pour les données téléphone. Ces expériences illustrent également la sensibilité de la comparaison de voix à des facteurs de variabilité inconnus et met en lumière la nécessité d'analyser et de comprendre les raisons d'une différence de performance, dans un sens ou l'autre.

Pour faciliter l'accès de PTSVOX, un outil développé avec MongoDB est en cours de développement. Cet outil permet de créer des protocoles de test spécifiques à travers une interface Web aisée d'accès. L'outil permettra d'automatiser l'exécution des systèmes de référence sur ces protocoles. La base PTSVOX, les systèmes de référence et les outils associés seront diffusés prochainement grâce à une licence libre.

Remerciements

Ce travail de recherche a été financé par le projet ANR-17-CE39-0016 VoxCrim qui inclut le Laboratoire Informatique d'Avignon, le Laboratoire Parole et Langage, le Laboratoire Phonétique et Phonologie, le Service Central de la Police Technique et Scientifique et l'Institut de Recherche Criminelle de la Gendarmerie Nationale. Les auteurs tiennent également à remercier Nathan Griot pour la conception de l'interface ainsi que les réservistes citoyennes de la Police nationale qui ont travaillé sur la base de données.

Références

AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & JULIETTE K. (2016a). Phonetic content impact on forensic voice comparison. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, p. 210–217.

AJILI M., BONASTRE J.-F., KAHN J., ROSSATO S. & BERNARD G. (2016b). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 726–733.

AJILI M., BONASTRE J.-F., ROSSETTO S. & KAHN J. (2016c). Inter-speaker variability in forensic voice comparison : a preliminary evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2114–2118.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5**(9/10), 341–345.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(4), 788–798.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *INTERSPEECH '09*. Brighton, Royaume-Uni.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1102–1107.

KAHN J., AUDIBERT N., ROSSATO S. & BONASTRE J.-F. (2010). Intra-speaker variability effects on speaker verification performance. In *Odyssey*, p.2̃1.

LARCHER A., BONASTRE J.-F., FAUVE B., LEE K. A., LEVY C., LI H., MASON J. & PARFAIT J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH '13*. Lyon, France.

RAMOS D., GONZALEZ-RODRIGUEZ J., GONZALEZ-DOMINGUEZ J. & LUCENA-MOLINA J. J. (2008). Addressing database mismatch in forensic speaker recognition with ahumada iii : a public real-casework database in spanish. In *Ninth Annual Conference of the International Speech Communication Association*.

VLOED D., BOUTEN J. & VAN LEEUWEN D. (2014). Nfi-frits : A forensic speaker recognition database and some first experiments. In *Proceedings of Odyssey Speaker and Language Recognition Workshop*, p. 6–13 : [Sl] : ISCA Speaker and Language Characterization special interest group.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Dis-moi comment tu varies ton débit, je te dirai qui tu es

Estelle Chardenon, Cécile Fougeron, Nicolas Audibert, Cédric Gendrot
Laboratoire de Phonétique et Phonologie - UMR 7018
Université Sorbonne Nouvelle - CNRS
19 Rue des Bernardins 75005 PARIS, France
(estelle.chardenon, cecile.fougeron, nicolas.audibert,
cedric.gendrot)@sorbonne-nouvelle.fr

RÉSUMÉ

Si l'étude de la variabilité entre locuteurs permet d'identifier des caractéristiques phonétiques potentiellement discriminantes, voire spécifiques, il est essentiel de comprendre, si et comment, ces caractéristiques varient chez un même locuteur. Ici, nous examinons la variabilité de caractéristiques liées à la gestion temporelle de la parole sur un nombre limité de locuteurs, enregistrés sur plusieurs répétitions dans une même session, et sur 6 à 7 sessions espacées d'une année. Sur cette vingtaine d'enregistrements par locuteur, nous observons comment le débit articulatoire, les modulations de ce débit, et la durée des pauses varient en fonction de la répétition et de la session et en interaction avec le locuteur. Les résultats montrent que c'est dans la variation de gestion temporelle de la parole que les locuteurs se distinguent les uns des autres, en termes de régularité ou non entre enregistrements et au sein d'un même enregistrement.

ABSTRACT

Tell me how you vary your speech flow, I'll tell you who you are

Studying inter-speaker variability makes it possible to identify discriminating or even identifying phonetic characteristics. It is however essential to understand if and how these characteristics vary within the same speaker. Here, we examine the variability of characteristics related to temporal organization of speech over a limited number of speakers, recorded over several repetitions in the same session, and over 6 to 7 sessions spaced one year apart. With this 20 recordings per speaker, we observe how the articulatory rate, the modulations of this rate, and the duration of pauses vary according to the repetitions and the sessions and interact with the speaker. The results show that it is in the variation of the temporal organization of speech that speakers differ from each other, in termes of regularity or not between recordings, and within the same recording.

MOTS-CLÉS: débit articulatoire, pause, variation intra et inter locuteur
KEYWORDS: articulatory rate, pause, inter-speaker variability, intra-speaker variability

Volume 1 : Journées d'Études sur la Parole, pages 82–90.
HAL : hal-02798520.

1 Introduction

Le débit d'un locuteur véhicule des informations spécifiques au locuteur. Par exemple, on sait que les hommes ont tendance à parler plus vite que les femmes même si l'écart de débit entre sexes n'est en général pas très élevé (Byrd, 1994 ; Jacewicz, 2009 ; Schwab, 2012). Des différences de variété régionale peuvent aussi être liées à des différences de débit. Ceci a été montré entre des variétés de l'anglais américain (Jacewicz, 2009), du néerlandais (Quené, 2007), mais aussi sur le français. Cependant, la comparaison entre certaines variétés du français (Suisse, Belge, Français), ne montre pas cette même tendance puisque les études se contredisent voire présentent des variabilités nulles (Schwab, 2015). Le débit articulatoire peut aussi différencier un locuteur natif d'un locuteur L2, même si celui-ci est très fluent (Schwab, 2012, Bordal et al. 2012 ; Barquero, 2012).

Si l'étude de la variabilité entre locuteurs permet d'identifier des caractéristiques de parole potentiellement discriminantes, voire identifiantes, il est essentiel de savoir comment ces caractéristiques varient pour un même locuteur pour qu'elles prennent sens (Kahn et al., 2010, Ajili et al, 2018). Par exemple, Campbell et al. (2009) montrent que les performances d'un système de reconnaissance automatique du locuteur sont fortement impactées lorsqu'il est testé sur deux enregistrements d'un même locuteur à plus d'un mois d'écart. Une des explications possibles à cette variation dans les performances serait liée à l'évolution naturelle de la parole en fonction de l'âge du locuteur. En effet, on sait que le vieillissement a des effets sur la parole, et la diminution du débit et la gestion des pauses chez les locuteurs âgés en est un des effets les plus connus (Ramig, 1983, Jacewicz, 2009, Dellwo, 2015). Pour autant, on ne sait pas sur quel laps de temps on peut observer ces effets. De plus, d'autres facteurs peuvent faire varier la façon dont un locuteur va organiser sa parole entre deux enregistrements. Cette variation peut être contrôlée et/ou volontaire, par exemple pour s'adapter à la situation de communication, au style de parole ou à la personne à qui le locuteur s'adresse (ex. personnes malentendantes ou non natives). On peut ainsi observer des différences de débit entre parole lue et parole spontanée, bien que les différences n'aillent pas toujours dans le même sens (Tsao et Weismer, 2006, Schwab, 2012, Dellwo, 2015, Crystal et House, 1982 ; Avanzi, Schwab, Bubosson et Goldman, 2012, Grosjean et Deschamps, 1975 ; Lucci, 1983). Une variation dans la gestion temporelle de la parole et des pauses peut aussi refléter des changements non contrôlés liés à son état général, par exemple sa consommation de tabac, une prise de médicament, le stress ou la fatigue. En outre, les personnes utilisant quotidiennement leur voix (acteurs, professeurs, chanteurs) sont les premières touchées par le syndrome de fatigue vocale alors qu'il existerait pourtant une meilleure stabilité vocale et articulatoire chez les personnes pratiquant à haut niveau une de ces activités (Lortie, 2016). Enfin, une habituation à la tâche lors d'une expérience avec plusieurs répétitions, ou une fatigue lors d'une longue prise de parole peut aussi survenir au sein de la même session d'enregistrement et impacter la vitesse de production d'un locuteur.

Si ces multiples facteurs pouvant impacter la parole ont été examinés ou mentionnés comme des prédicteurs potentiels de la variation intra et inter-locuteurs, ceux-ci ont rarement été comparés directement et évalués. Dans cette étude, notre objectif est d'examiner la variabilité de paramètres liés à la gestion temporelle de la parole entre un nombre limité de locuteurs, enregistrés plusieurs fois dans une même session, et sur 6 à 7 sessions espacées d'une année.

2 Méthode

2.1 Enregistrements et description de la population

Les enregistrements de 5 locuteurs de la base de données PATATRA (‘Parole AdulTe A TRavers les Ages’), en cours d’acquisition au LPP, ont été utilisés. Cette base est constituée d’enregistrements d’une dizaine de locuteurs, collectés une fois par an depuis 2013, sur plusieurs tâches de parole. Les enregistrements audio (et EGG) sont effectués en chambre sourde avec le même matériel, et un questionnaire de qualité de vie, axé sur les potentiels facteurs de variation de la voix/parole, est recueilli à chaque session.

Cinq locutrices francophones natives ayant entre 6 et 7 sessions d’enregistrement ont été sélectionnées. Les informations concernant ces locutrices sont résumées dans la table 1. Lors du premier enregistrement, les locutrices étaient âgées de 39 à 68 ans, les locutrices F05 et F04 étant les plus âgées et les trois autres étant dans la quarantaine. Elles présentent toutes un niveau d’étude similaire et vivent en région parisienne.

	F01	F02	F03	F04	F05
Age <2013-2018>	<39-45>	<42-48>	<42-48>	<68-73>	<57-63>
Résidence	IDF	IDF	IDF	IDF	IDF
Nb sessions (manque)	7	6 (2015)	7	6 (2013)	7
Pratique vocale	Oui	Oui	Oui	Oui	Oui
Fatigue vocale	Rare	Modérée	Modérée	Modérée	Modérée

TABLE 1 : Description des 5 locutrices (F01-F05) et des sessions d’enregistrement

2.2 Traitement des données

Notre étude se base sur la tâche de lecture du texte « la bise et le soleil » que le locuteur doit lire à 3 reprises avec 5 minutes de pause entre deux lectures. Si une erreur de lecture est commise, le locuteur est invité à recommencer au début du paragraphe lu. Pour chaque locutrice, le matériel analysé inclut 3 répétitions sur 6 ou 7 sessions, donc 18 à 21 enregistrements. De façon à pouvoir aussi évaluer la variabilité au sein d’une même répétition, le texte a été divisé en 18 *chunks* comprenant chacun une dizaine de syllabe, les pauses internes éventuelles et la pause finale éventuelle (signalées par un dièse).

Chunk	Transcription API	Nombre de phonèmes	Nombres de syllabes
1	/la biz e lə solɛj sə dispytɛ #/	22	10
2	/ʃakɛ̃ asyʁɑ̃ kil etɛ lə ply fɔʁ #/	23	11
3	/kɑ̃ ilz ɔ̃ vy ɛ̃ vwajaʒœʁ ki/	19	9
4	/savɑ̃se # ɑ̃vəlope dɑ̃ sɔ̃ mɑ̃to #/	21	10 ou 11
5	/il sɔ̃ tɔ̃be dakɔʁ kə səlɥi/	20	9
6	/ki aʁivəʁɛ lə prømje a/	18	9
7	/fɛʁ ote sɔ̃ mɑ̃to o vwajaʒœʁ #/	21	10
8	/səʁɛ ʁəgaʁde kɔm lə ply fɔʁ #/	22	9
9	/alɔʁ la biz sɛ miz a sufle/	20	9
10	/də tut se fɔʁs me plyz ɛl suflɛ #/	24	9
11	/ply lə vwajaʒœʁ seʁɛ sɔ̃ mɑ̃to/	23	10
12	/otuʁ də lɥi # e a la fɛ̃#/	15	8
13	/la biz a ʁənɔ̃se a lə lɥi fɛʁ ote #/	24	12
14	/alɔʁ lə solɛj a komɑ̃se a brije #/	24	12
15	/e o bu dɛ̃ momɑ̃ # lə vwajaʒœʁ/	20	10
16	/ʁeʃofe # a ote sɔ̃ mɑ̃to #/	16	9
17	/ɛ̃si la biz a dy ʁəkonɛtʁ/	19	9
18	/kə lə solɛj etɛ lə ply fɔʁ de dø #/	24	11

TABLE 2 : Découpage et transcription des chunks

Ce choix de découpage est justifié par le fait que nous voulions obtenir des extraits de parole contenant à peu près le même nombre d'unités de parole pour pouvoir les comparer entre eux et avoir un même découpage pour tous les locuteurs. Cette segmentation a été faite en respectant au mieux le découpage syntaxique, sans couper de mots. Les 18 chunks contiennent 20.8 phonèmes en moyenne (σ=2.7, 15<>24) correspondant à 9.7 syllabes en moyenne (σ =1.14, 9<>12).

La durée des 18 chunks ainsi que celles des pauses, a été mesurée de façon à calculer, pour chaque chunk, un *débit articulatoire* (nombre de phonème sur temps de parole sans pauses), *le nombre de pauses* éventuelles et *la durée de ces pauses*. Une mesure de *modulation du débit articulatoire* au long du texte a été estimée à partir des différences de débit entre chunks adjacents. Ainsi pour chaque paire de chunks adjacents (17 paires) nous avons calculé une différence relative de débit articulatoire entre chunks à l'aide de la formule suivante : $(d_N - d_{N+1})/((d_N + d_{N+1})/2)$ où d est le débit du $N^{ième}$ chunk.

Pour prédire les variations de débit articulatoire, de durée des pauses et de différence relative entre les chunks, nous avons utilisé des modèles mixtes testant les effets du locuteur, de la répétition, de la session, mais également l'interaction entre les facteurs : locuteurs et répétition, ainsi que locuteur et session en conservant l'identité du chunk comme facteur aléatoire (lmer(VD~ LOCUTEUR + RÉPÉTITION + SESSION + LOCUTEUR: RÉPÉTITION + LOCUTEUR: SESSION + (1| CHUNK)).

Afin de vérifier si les variations de débit observées entre les sessions et répétitions n'étaient pas dues à des différences intrinsèques de débit entre locuteurs, nous avons également fait des analyses sur des débits relatifs.

3 Résultats

	débit articulatoire	*durée des pauses*	*diff_inter_chunk*
LOCUTEUR	$\chi^2(4)= 425.29$ $p<0.0001$, $R^2=0.008$	$\chi^2(4)= 81.38$ $p<0.0001$, $R^2=0.007$	($\chi^2(4)= 33.47$ $p<0.0001$, $R^2=0.006$
RÉPÉTITION	$\chi^2(2)= 8.87$, $p=0.01$, $R^2=0.002$	NS	NS
SESSION	$\chi^2(6)= 57.87$,$p<0.0001$, $R^2=0.01$	$\chi^2(6)= 16.17$ $p=0.01$, $R^2=0.005$	NS
LOCUTEUR: RÉPÉTITION	$\chi^2(8)=32,52$, $p<0.0001$, $R^2=0.05$	$\chi^2(8)=22.73$, p=0.004, $R^2=0.03$	NS
LOCUTEUR: SESSION	$\chi^2(22)=167.82$, $p<0.0001$, $R^2=0.04$	$\chi^2(22)=74.46$, $p<0.0001$, $R^2=0.02$	NS

TABLE 3 : Résumé des résultats statistiques

Les résultats sont résumés dans la Table 3 et les figures 1, 2 et 3. Globalement, un effet du LOCUTEUR est observé sur tous les paramètres étudiés mais en interaction avec les facteurs RÉPÉTITION et SESSION pour le débit et la durée des pauses. Pour ces trois paramètres, ces interactions peuvent être expliquées par des profils de variation différents en fonction des locuteurs. Notre étude ne présente que des statistiques descriptives puisque nous n'avons qu'une seule valeur pour chaque enregistrement et locuteur.

En ce qui concerne le *débit articulatoire*, illustré Figure 1a, toutes les locutrices sauf F02 varient leur débit de manière significative entre sessions. Les comparaisons par paires de sessions montrent toutefois que les locutrices F04, F03 et F01 varient plus d'une session à l'autre que la locutrice F05 qui varie essentiellement entre sa première session et les suivantes. D'autre part, les différences de débit entre sessions semblent essentiellement aléatoires et ne suivent pas de tendance particulière qui pourrait refléter un ralentissement d'une année à l'autre ou sur les dernières sessions. Seul le fait que la première session soit globalement plus rapide ressort pour 3 locutrices (F01, F03, F04). La variation entre répétitions (intra session) dépend aussi du locuteur. Seuls les locutrices F03 et F04 varient leur débit de façon significative entre répétitions, avec une première répétition plus lente que les suivantes.

En ce qui concerne la *durée des pauses*, illustrée Figure 1b, les différences entre sessions et entre répétitions dépendent aussi du locuteur. F05 et F03 gardent une durée de pause stable entre les sessions. Les locutrices F04, F01 et F02 varient d'une session à l'autre, F04 étant la moins régulière. Pour ces trois locutrices, la variation entre sessions semble aléatoire sans suivre une tendance particulière au cours du temps. Entre répétitions, il y a moins de variation dans la durée des pauses : seuls deux locuteurs, F01 et F05, montrent des différences significatives, avec des pauses plus courtes sur leurs premières lectures.

La *mesure de modulation du débit*, illustrée Figure 1c, qui capture la variation de débit d'un chunk à l'autre au sein d'une répétition, est particulièrement intéressante car c'est la seule qui distingue les locuteurs indépendamment de la répétition ou la session. Cette mesure permet de distinguer deux groupes : des locutrices qui modulent beaucoup leur débit (les locutrices F01 et F05) et des locutrices qui modulent moins (F02 et F04). La locutrice F03 présente un profil intermédiaire : similaire à F04 mais modulant plus que F02. N'ayant qu'une seule valeur ici, nous n'avons pas de barres d'erreur.

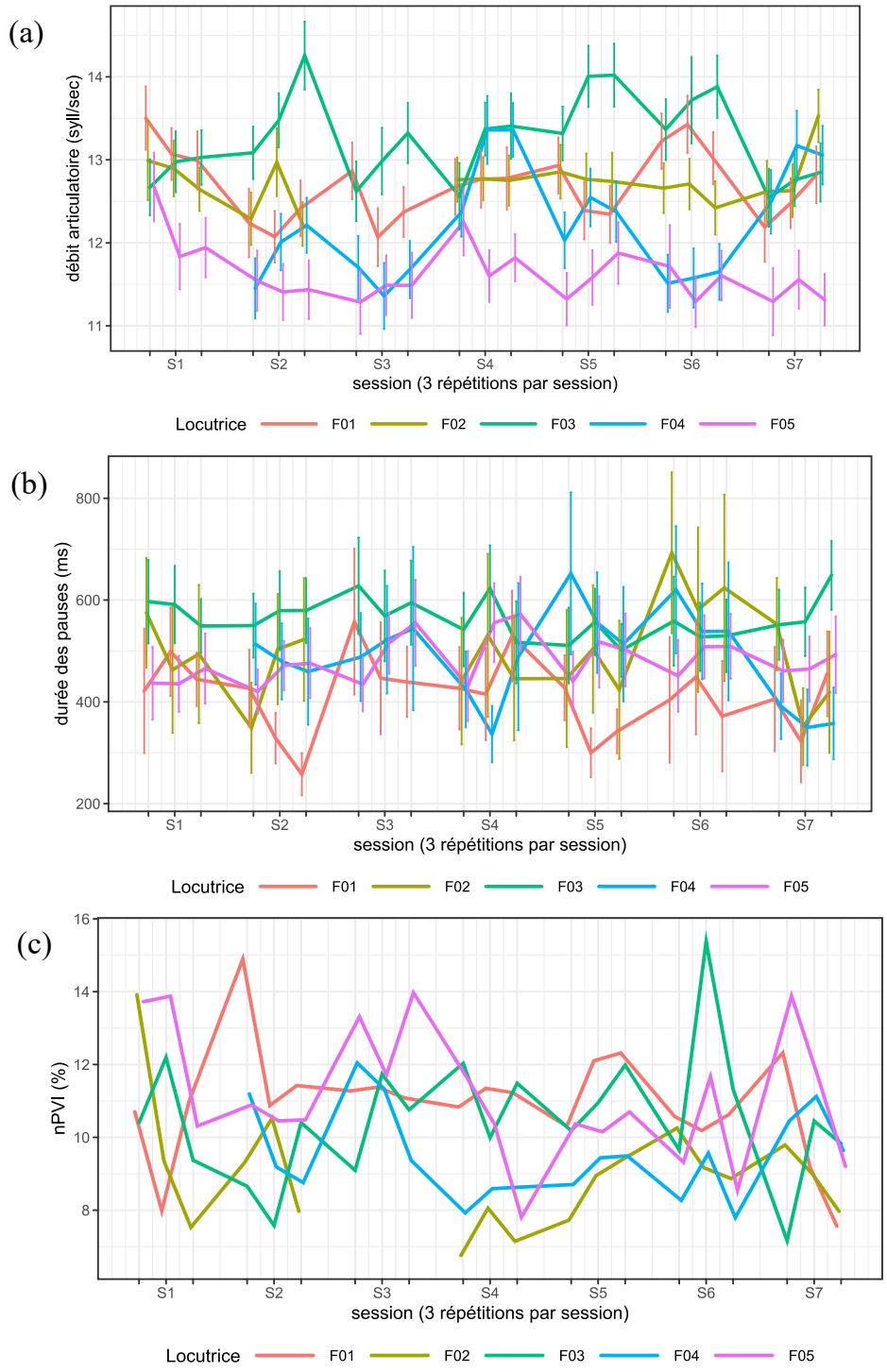


Figure 1 : Évolution (a) du débit articulatoire moyen, (b) de la durée moyenne des pauses, (c) de la modulation du débit (différence relative de débit articulatoire entre chunks consécutifs) au cours des 6/7 sessions successives et des 3 répétitions par session. Les barres d'erreur représentent la variation entre chunks au sein de chaque enregistrement.

4 Discussion et conclusion

Cette étude basée sur une vingtaine d'enregistrements par locuteur montre que c'est dans la variabilité de leur gestion temporelle (débit, pause, modulation de débit) que les locuteurs se distinguent les uns des autres.

Cette variabilité intra-locuteur s'observe premièrement dans la modulation de débit d'un chunk à l'autre au court d'une même lecture. Certaines locutrices ont un débit qu'on pourrait qualifier de monotone ou isochrone (F02, F04) alors que d'autres lisent le texte en variant beaucoup leur débit. Puisque le débit articulatoire et les pauses dépendent beaucoup du découpage prosodique, nous prévoyons dans les prolongements ultérieurs de cette étude de contrôler dans quelle mesure ces différents profils se retrouvent aussi dans d'autres aspects de l'organisation prosodique du texte produit par les locuteurs dans différents enregistrements.

Deuxièmement, cette variabilité intra-locuteur s'observe entre les différents enregistrements. Certaines locutrices sont très régulières d'une session à l'autre ou d'une répétition à l'autre. C'est le cas de la locutrice F02 qui ne change pas de débit articulatoire, mais aussi de la locutrice F05 qui change peu son débit et a aussi des pauses de durée stable. Au contraire, les locutrices F04, F03 et F01 adoptent des débits plus changeants, accompagnés de durées de pauses variées pour F01 et F04.

Le cas de la locutrice F04 est intéressant. Son débit est très régulier en termes de modulation au court d'une lecture, par contre elle varie beaucoup entre sessions et répétitions. Il est à noter qu'elle est aussi celle qui produit le plus de pauses. Un examen plus poussé de la structuration prosodique du texte lu sera nécessaire pour voir si celle-ci est également plus changeante d'une lecture à l'autre. Cette locutrice est la plus âgée de la cohorte, toutefois son débit n'est pas plus lent sur toutes les répétitions/sessions, ni ses pauses toujours plus longues que celles des autres. C'est donc bien en termes de variabilité que cette locutrice se distingue des autres. Ce résultat est en accord avec les études montrant que s'il y a une tendance à ralentir le débit avec l'âge (5% chez Verhoeven, 2004) ce ralentissement est très variable entre les individus (Decoster, 2000) mais c'est aussi la variabilité intra-sujet qui augmente avec l'âge (Morris et Brown 1994). Toutefois, il reste à démontrer si cette tendance est dû à l'âge ou à l'individu puisque la locutrice F05, également plus âgée que les autres locutrices, présente quant à elle plus de régularité.

Dans ces données, il ne ressort pas d'évolution particulière entre les premières années d'enregistrement et les dernières qui pourrait refléter une évolution longitudinale. Les enregistrements sont réalisés sur une période de 6 ou 7 ans, un laps de temps qui n'apparait pas suffisant pour montrer un effet d'âge, comparé à d'autres études longitudinales qui observent un changement sur un laps de temps beaucoup plus long (30 ans pour Descoster, 2000 ; 35 pour Harrington, 2007). De plus, l'âge auquel on commence à observer de nombreux changements dans la parole se situe entre 60 et 70 ans, ce qui coïncide avec des changements physiologiques (Decoster, 2000).

Cette étude sera poursuivie sur une plus longue durée en exploitant les enregistrements à venir de la base PATATRA. Il sera également intéressant d'explorer la variabilité intra-locuteur sur d'autres caractéristiques de la voix et la parole. Si les recherches en phonétique ont eu pour tradition dans le passé de moyenner les caractéristiques de parole entre répétitions, entre sessions, voire entre locuteurs, des résultats comme ceux présentés ici illustrent l'intérêt de mieux comprendre la variation dans la parole ainsi que la façon dont cette variation peut indexer des particularités individuelles.

Remerciements

Ce travail a été partiellement financé par l'ANR VoxCrim (ANR-17-CE39-0016), le programme Investissements d'Avenir (ANR-10-LABX-0083), et le subside FNS CRSII5_173711/1 Sinergia du FNRS.

Références

AJILI, M., ROSSATO, S., ZHAND, D., & BONASTRE, J. F. (2018). Impact of rhythm on forensic voice comparison reability. HAL : ⟨hal-01962531⟩

AVANZI, M., SCHWAB, S., DUBOSSON, P., & GOLDMAN, J. P. (2012). La prosodie de quelques variétés de français parlées en Suisse romande. *La variation prosodique régionale en français*. Bruxelles: De Boeck/Duculot, 89-118.

BARQUERO-ARMESTO, M. Á. (2012). A comparative study on accentual structure between Spanish learners of French interlanguage and French native speakers. *In Speech Prosody* 2012.

BORDAL, G., AVANZI, M., OBIN, N., & BARDIAUX, A. (2012). Variations in the realization of the French Accentual Phrase in the light of language contact. HAL : ⟨hal-01161304⟩

BYRD, D. (1994). Relations of sex and dialect to reduction. Speech Communication, 15(1-2), 39-54.

CAMPBELL, W., BONASTRE, J. F., SCHWARTZs, R., DRISS, M. (2009). Forensic Speaker Recognition. *Signal Processing Magazine*. IEEE 26.2 : 95-103. **DOI:** <10.1109/MSP.2008.931100>

CRYSTAL, T. H., & HOUSE, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. The JASA, 72(3), 705-716.

DECOSTER, W., & DEBRUYNE, F. (2000). Longitudinal voice changes: facts and interpretation. Journal of Voice, 14(2), 184-193. DOI : <doi.org/10.1016/S0892-1997(00)80026-0>

DELLWO, V., LEEMAN, A., KOLLY, M. J. (2015). Rythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *JASA*, 137(3): 1513-1528. DOI: <10.1121/1.4906837>

GROSJEAN, F., & DESCHAMPS, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, 31(3-4), 144-184. DOI : <doi.org/10.1159/000259667>

HARRINGTON, J., PALETHORPE, S., & WATSON, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In Eighth Annual Conference of the International Speech Communication Association.

JACEWICZ, E., FOX, R. A., O'NEILL, C., & SALMONS, J. (2009). Articulation rate across dialect, age, and gender. *Language variation and change*, 21(2), 233-256. DOI : <10.1017/S0954394509990093>

KAHN, J., AUDIBERT, N., ROSSATO, S., & BONASTRE, J. F. (2010). Intra- speaker variability effects on Speaker Verification performance. *In Odyssey* p. 21.

LORTIE, C. L., RIVARD, J., THIBEAULT, M., & TREMBLAY, P. (2017). The moderating effect of frequent singing on voice aging. *Journal of Voice*, 31(1), 112-e1.

LUCCI, V. (1983). Etude phonétique du français contemporain à travers la variation situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes). *Publications de l'Université des Langues et Lettres de Grenoble*, 1-360.

MORRIS, R. J., & BROWN Jr, W. S. (1994). Age-related differences in speech variability among women. *Journal of Communication Disorders*, *27*(1), 49-64.

QUENE, H. (2007). On the Just Noticeable Difference for tempo in speech. *Journal of Phonetics*. 35(3), 353– 362. DOI: <10.1016/j.wocn.2006.09.001>

RAMIG, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of communication disorders*, 16(3), 217-226. DOI : <doi.org/10.1016/0021-9924(83)90035-7>

SCHWAB, S., & AVANZI, M. (2015). Regional variation and articulation rate in French. *Journal of Phonetics*, 48, 96-105. DOI: <10.1016/j.wocn.2014.10.009>

SCHWAB, S., DUBOSSON, P., & AVANZI, M. (2012). Etude de l'influence de la variété dialectale sur la vitesse d'articulation en français (Dialectal Effect on Articulation Rate in French). *In Proceedings of the Joint Conference* JEP-TALN-RECITAL 2012,1: JEP, 521-528.

TSAO, Y.C., WEISMER, G., IQBAL, K. (2006). Interspeaker Variation in Habitual Speaking Rate: Additional Evidence. *Journal of Speech, Language and Hearing Research*. 49: 1156-1164.

VERHOEVEN, J., DE PAUW, G., KLOOTS, H. Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. Language and Speech 2004;47:297–308.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Caractérisation du locuteur par CNN à l'aide des contours d'intensité et d'intonation : comparaison avec le spectrogramme

Gabriele Chignoli Cédric Gendrot Emmanuel Ferragne
Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins, 75005 Paris, France
gabriele.chignoli@sorbonne-nouvelle.fr,
cedric.gendrot@sorbonne-nouvelle.fr, emmanuel.ferragne@u-paris.fr

RÉSUMÉ

Dans ce travail nous avons recours aux variations de f0 et d'intensité de 44 locuteurs francophones à partir de séquences de 4 secondes de parole spontanée pour comprendre comment ces paramètres prosodiques peuvent être utilisés pour caractériser des locuteurs. Une classification automatique est effectuée avec un réseau de neurones convolutifs, fournissant comme réponse des scores de probabilité pour chacun des 44 locuteurs modélisés. Une représentation par spectrogrammes a été utilisée comme référence pour le même système de classification. Nous avons pu mettre en avant la pertinence de l'intensité, et lorsque les deux paramètres prosodiques sont combinés pour représenter les locuteurs nous observons un score qui atteint en moyenne 59 % de bonnes classifications.

ABSTRACT

CNN speaker characterisation through prosody : spectrogram comparison

In this study we focused on f0 and intensity variation in four-second spontaneous speech sequences by 44 French speakers in order to evaluate the strength of prosodic parameters for speaker characterisation. We used a deep Convolutional Neural Network (CNN) with f0 and/or intensity values as input in a classification task, where the system had to classify 44 speakers in a closed dataset. Spectrograms were also used as input with the same CNN architecture as a benchmark for maximum possible performance. Results show that f0 and intensity are complementary as together they yield 59 % classification precision.

MOTS-CLÉS : Caractérisation du locuteur, intensité, f_0, prosodie, réseaux de neurones convolutifs.

KEYWORDS: Speaker characterisation, intensity, fundamental frequency, prosody, convolutional neural network.

1 Introduction

Les recherches sur la caractérisation phonétique du locuteur ont fréquemment pour objectif d'identifier des facteurs cohérents expliquant la variabilité inter-individuelle (Kahn, 2011). Ces facteurs comprennent notamment des mesures acoustiques issues de variables d'ordre biologique telles que l'âge ou des troubles de la parole. Par exemple, dans Schötz (2007), parmi plusieurs mesures acoustiques étudiées en relation avec l'âge, l'étendue de la pression acoustique se montre stable. Le style de parole (Arvaniti & Rodriquez, 2013) ou le sexe sont également des critères de variation importants. Dans Keating & Kuo (2012), des locuteurs hommes et femmes natifs de l'anglais ou du mandarin sont comparés à partir de mesures de fréquence fondamentale dans des tâches de lecture de mots isolés,

Volume 1 : Journées d'Études sur la Parole, pages 91–99.
HAL : hal-02798521.

de texte et de vocalisation dans différents contextes. Plusieurs mesures de ce descripteur, valeurs moyennes, maximales, minimales, moyenne de l'étendue ou écart type sont comparées dans deux analyses de variance prenant en compte la langue et le sexe comme facteurs de variabilité inter-locuteurs. Les variations de f_0 d'ordre physiologique sont jugées similaires dans les deux langues par les auteurs.

Étant donné la robustesse de la f_0 face à plusieurs facteurs perturbateurs comme les canaux de diffusion, l'effort vocal ou l'état d'esprit du locuteur (Lindh & Eriksson, 2007), elle s'est affirmée comme un des descripteurs les plus analysés non seulement pour essayer de différencier le sexe des locuteurs mais aussi pour distinguer les locuteurs de même sexe. Hudson *et al.* (2007) montrent comment la distribution des valeurs de f_0 pour 100 locuteurs hommes reste cohérente indépendamment de la tâche exécutée, avec la majorité des locuteurs (65 %) qui présentent une variation intra-individuelle ne dépassant pas les 20 Hz. Niebuhr & Skarnitzl (2019) analysent plusieurs mesures de f_0 sur 51 locuteurs, hommes et femmes, lors d'une tâche de conversation semi-spontanée pour décrire comment celles-ci peuvent représenter la variabilité des locuteurs. Les mesures prises en compte, dont la moyenne, l'étendue (pour 80% des locuteurs) et le kurtosis de la distribution des valeurs de f_0 pris tous les 5 ms, s'avèrent utiles pour expliquer la variabilité inter-locuteurs. Les résultats de cette étude montrent également que la f0 est un paramètre très dépendant du sexe. Dans Adami *et al.* (2003), c'est la complémentarité des indices qui est analysée avec 40 locuteurs, 20 hommes et 20 femmes, dont le contour intonatif et l'énergie de 15 mots sont utilisés pour modéliser les locuteurs et successivement tester un système de reconnaissance automatique du locuteur pour une des tâches du protocole NIST 2001. Les deux paramètres sont analysés d'abord séparément puis ensemble et montrent une forte complémentarité puisque l'erreur totale passe de 13 % à 3 %.

Contrairement à la f_0, l'intensité est un facteur prosodique peu utilisé dans la description de la variabilité inter-locuteurs. Sorin (1981) montre la cohérence de ce paramètre et la sensibilité de l'oreille humaine à sa variation ; cependant le rôle linguistique de l'intensité reste à approfondir. Pardo (2006) cite l'intensité comme l'un des facteurs pouvant potentiellement déterminer la similarité entre locuteurs. Dans la lignée de ces considérations sur l'intensité, Tweedy & Culling (2014) montrent comment ce facteur reste peu influencé chez des locuteurs soumis à des perturbations lors de tâches de conversation. Plus récemment nous retrouvons dans He *et al.* (2015) l'utilisation de mesures d'intensité parallèlement à des corrélats rythmiques pour la classification de 16 locuteurs à travers un système de réseaux de neurones. Dans cette étude, l'intensité permet d'atteindre un taux de bonnes classifications de 30 % qui augmente à 36 % lorsque les corrélats rythmiques et d'intensité sont observés ensemble. Les mesures utilisées se rapprochent des corrélats rythmiques déjà utilisés par les auteurs, nous renvoyons à cette étude et à d'autres (Dellwo *et al.*, 2015) pour plus de précision à ce sujet. La complémentarité des mesures prosodiques est un des points fondamentaux de l'étude que nous allons présenter puisque la production de la parole, et dans notre cas, la représentation de la variabilité inter-individuelle ne peut être expliquée que par l'interaction de plusieurs facteurs.

Dans ce travail nous essayons d'apporter un élément d'analyse supplémentaire à la caractérisation du locuteur à partir d'indices prosodiques, pour cela nous allons prendre en compte deux dimensions prosodiques classiques : la f_0 et l'intensité. Choisir ces deux mesures nous permettra de comparer leurs résultats respectifs et leur éventuelle complémentarité dans la caractérisation du locuteur. Dans la section suivante nous décrirons de manière plus précise les mesures acoustiques qui ont été sélectionnées, leurs méthodes d'extraction et de classification. La complémentarité entre les résultats des différentes mesures sera présentée dans la section 3 ainsi qu'une analyse plus spécifique de certains locuteurs et cas particuliers.

2 Protocole expérimental

Le corpus sur lequel nous avons travaillé est composé d'extraits de 4 secondes provenant de chacun des 44 locuteurs du corpus NCCFr (Torreira *et al.*, 2010) (pour les 45ème et 46ème les transcriptions n'étant pas intégralement présentes, nous les avons écartés). Nous utiliserons les chiffres de 1 à 44 pour identifier les locuteurs à la place des codes présents dans la base. Le corpus NCCFr se compose de conversations sur des thèmes divers entre binômes ou trinômes d'amis, dont la durée totale est d'une heure en moyenne, et enregistrées dans une chambre insonorisée à l'aide d'un micro casque, ce qui limite les variations d'intensité dûes aux mouvements de la tête. Les transcriptions ont été effectuées de manière semi-automatique au laboratoire LIMSI (Gauvain *et al.*, 2002). Nous avons considéré uniquement les extraits comportant un minimum de 20 phonèmes, en dessous de ce seuil, après écoute des extraits, nous avons souvent remarqué la présence de la voix des autres locuteurs de l'enregistrement. À partir de ces extraits nous avons obtenus 4 types de données que nous avons utilisées dans un système de classification à l'aide de réseaux de neurones : le spectrogramme complet de chaque séquence ; le contour de f_0 ; les valeurs d'intensité ; la représentation conjointe de f_0 et d'intensité. Le spectrogramme servira ici de référence puisqu'il comprend l'information la plus complète sur la séquence analysée.

Ces représentations ont été extraites avec les valeurs par défaut de Praat (Boersma, 2001) et converties en images codées sur 8 bits en niveaux de gris afin de préserver la mémoire du GPU. Pour l'extraction des spectrogrammes nous avons utilisé une fenêtre avec des trames de 5 ms avec un chevauchement de 90 % et une fréquence d'échantillonage de 16 kHz. Sur chaque trame de signal a été appliquée une fenêtre de Hamming pour obtenir 512 échantillons sur lesquels une FFT a été effectuée. Nous n'avons pas appliqué de pré-emphase et la dynamique a été fixée à 70 dB. Dans le but de ne pas saturer la mémoire du GPU, les spectrogrammes ont été redimensionnés à 800×257 pixels ($L \times H$), où chaque pixel correspond à 5 ms sur l'axe du temps (après compression de 8000 à 800 par interpolation bicubique), et à 31.13 Hz sur celui de la fréquence. Le même redimensionnement sur l'axe horizontal est utilisé pour les répresentations d'intensité et de f_0, sur l'axe vertical nous avons 1 seul pixel dont les niveaux de gris correspondent aux valeurs d'intensité ou de f_0. Pour la répresentation conjointe d'intensité et f_0 le format des images passe à 800×2 pixels.

Le modèle que nous avons utilisé pour la classification est une version légèrement modifiée d'un réseau de neurones convolutif avec architecture ResNet-18, à la base employé pour de la reconnaissance d'images. Les données de chacun des 44 locuteurs ont été aléatoirement divisées selon des ensembles d'entraînement (70 %), validation (10 %) et test (20 %). Le même découpage est utilisé les quatre expériences (f_0, intensité, f_0 et intensité, spectrogrammes). L'entraînement du réseau a été effectué avec l'optimiseur Adam et des mini-batches de taille 32 sur un GPU GTX 1080. Le nombre maximum d'itérations défini étant de 30, l'entraînement du modèle s'arrête avant cette limite si la valeur de la perte atteint un chiffre inférieur ou égal à sa valeur minimale dans les 10 dernières itérations. Tous les modèles ont été entraînés et testés à l'aide de la MATLAB Deep Learning Toolbox (Mathworks, 2019).

Pour pouvoir comparer et interpréter les résultats obtenus par la classification automatique nous avons aussi extrait des valeurs acoustiques à l'aide du logiciel Praat (Boersma, 2001) à partir des mêmes séquences de 4s. L'analyse de celles-ci nous a permis de comprendre dans quelle mesure les locuteurs se différencient ou se rapprochent du point de vue acoustique et ainsi d'interpréter les confusions faites par la classification automatique avec une approche phonétique. L'ensemble des mesures effectuées pour la f_0 et l'intensité comprend : les valeurs minimales, maximales, la moyenne,

l'écart type, le premier décile, le dernier, le décile du milieu et la différence entre les valeurs de f_0 et d'intensité (voir paragraphe 3.2). Nous avons aussi extrait d'autres valeurs comme la pente de f_0, le nombre de pics de f_0 et d'intensité mais nous les avons écartées de notre analyse finale car elles n'apportent pas d'éléments intéressants à celle-ci.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
Intonation	5	15	17	8	44	14	10	30	24	72	17	
Intensité	37	15	37	25	50	31	22	40	35	45	41	
Intens-into	45	41	77	50	88	62	62	88	80	82	40	
Spectrogramme	88	97	92	97	97	88	98	92	97	98	92	
	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	
Intonation	5	27	27	21	62	34	25	27	20	37	25	
Intensité	45	10	12	27	67	88	47	47	5	15	37	
Intens-into	72	48	51	51	87	100	71	65	51	42	70	
Spectrogramme	97	81	90	87	97	98	98	98	92	95	95	
	23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	
Intonation	15	12	11	41	20	55	37	34	15	25	8	
Intensité	10	10	4	55	44	34	44	17	42	8	18	
Intens-into	47	68	28	78	57	95	87	54	44	22	44	
Spectrogramme	94	97	94	92	94	91	95	95	92	85	98	
	34	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**Tot.**
Intonation	4	20	44	68	25	37	42	25	25	62	20	28
Intensité	17	8	22	62	24	62	12	31	21	61	48	32
Intens-into	27	37	62	62	50	44	60	31	28	80	48	59
Spectrogramme	88	75	91	97	95	94	77	92	100	97	92	93

TABLE 1 – Résultats obtenus par le réseau de neurones lors de la classification des différentes représentations, pourcentage de réussite pour l'ensemble des 44 locuteurs

3 Résultats

La tâche de classification consiste, pour chacun des 44 locuteurs, à donner au réseau de neurones les 70 extraits pris aléatoirement dans l'ensemble de départ et ne faisant pas partie des extraits d'entraînement. Cette même expérience a été répétée 4 fois pour chacune des représentations. La première considération concerne le score obtenu par les spectrogrammes, qui atteint un taux de bonnes classifications de 93 %, résultat attendu en considérant que cette représentation décline plusieurs éléments de la production de la parole alors que les 3 autres ne prennent en compte qu'une dimension voire deux à la fois. L'utilisation du contour intonatif, la répresentation à la fois des valeurs moyennes et des modulations de f_0, atteint le résultat moyen le moins élevé avec 28 %. Les valeurs d'intensité apportent quelques informations supplémentaires en totalisant un taux total moyen de 32 %, mais c'est lorsque les deux représentations sont combinées que le score est presque doublé et atteint 59 %.

3.1 Complémentarité des paramètres prosodiques

En considérant le spectrogramme comme la valeur de référence, nous remarquons deux locuteurs qui sont malgré tout mieux reconnus grâce à la prosodie seule : les locutrices 17 et 28. Elles présentent des profils similaires même si l'une se caractérise de manière plus marquée à travers l'intensité : la locutrice 17 obtient 88 % de bonnes classifications à partir des valeurs d'intensité et 34 % à partir de l'intonation, contre respectivement 24 % et 39 % pour la locutrice 28. Lorsque les deux dimensions sont combinées la locutrice 17 enregistre 100 % de bonnes classifications alors que la locutrice 28 obtient 95 %.

La locutrice 17 est caractérisée par des valeurs d'intensité toujours maximalement basses. La Figure 1(a) met en évidence cette tendance avec les valeurs minimales d'intensité pour les 44 locuteurs, nous prenons en exemple cette mesure puisque, avec les valeurs maximales et le dernier décile, elles permettent de distinguer de manière nette quelques groupes de locuteurs. Nous comprenons aussi pourquoi lors de la classification de la locutrice 17 la seule confusion existante se fait avec la locutrice 16, son binôme d'enregistrement, qui apparaît également dans la partie gauche de ce schéma avec des valeurs bien inférieures à la moyenne. Ceci nous renvoie aux considérations faites plus haut sur la convergence phonétique de l'intensité. Nous constatons la même tendance à avoir des valeurs d'intensité proches pour d'autres binômes : 25 et 26, 32 et 33, 34 et 35, 41 et 42. Cependant l'observation des valeurs de f_0 pour la locutrice 17 nous fait remarquer qu'elles ne s'écartent pas de la moyenne et forment un groupe compact avec les locutrices 24, 28, 35 et parfois 25. Nous observons cette fois une certaine distance avec la locutrice 16, qui quant à elle fait partie d'un groupe plus large. Cette différence évidente nous montre un exemple saillant de la complémentarité entre les mesures de f_0 et d'intensité, puisque grâce à l'une nous pouvons éliminer les confusions obtenues avec l'autre. Pour la locutrice 28, en utilisant la prosodie, nous n'observons pas une élimination totale

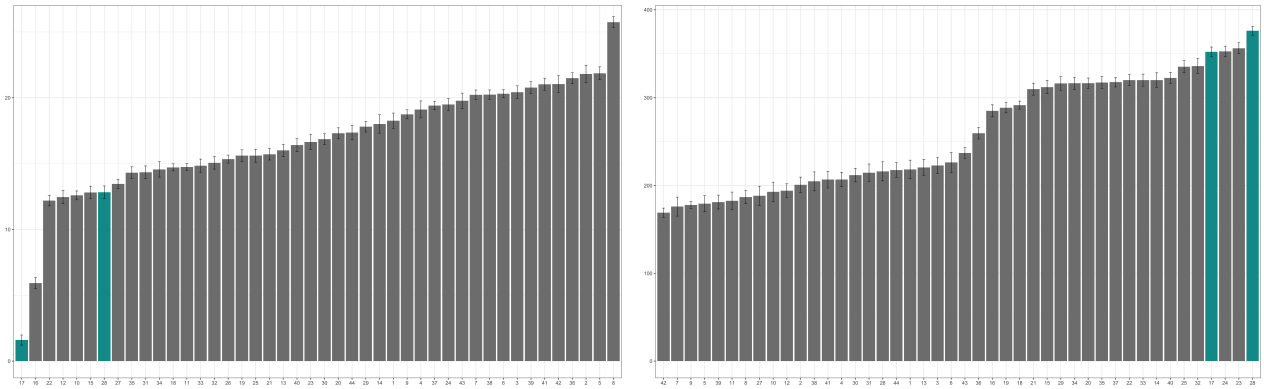


FIGURE 1 – Comparaisons des valeurs minimales d'intensité, en dB (Figure 1a, à gauche) et des valeurs maximales de f_0, en Hz (Figure 1b, à droite) pour les 44 locuteurs. Les locutrices 17 et 28 mises en évidence

des confusions mais plutôt une réduction du nombre de confusions effectuées par la classification. D'un côté, pour ses valeurs d'intensité, nous observons un petit groupe de locuteurs ayant des valeurs très basses mais rapprochés entre eux (Figure 1a) : il s'agit des locuteurs 10, 12 (les deux étant des hommes), 15, 22 et 28. Nous retrouvons tous ces locuteurs dans l'ensemble des confusions faites à partir des valeurs d'intensité pour la locutrice 28, mais aucun d'entre eux n'apparaît dans les résultats

obtenus à partir de la f_0. Pour cet autre paramètre nous retrouvons une confusion importante avec les locutrices 17 et 24, avec qui 28 est confondue respectivement 12 et 8 fois. Ces deux dernières locutrices, comme observé précédemment, font partie d'un groupe assez compact dont la locutrice 28 s'écarte grâce à ses valeurs maximales de f_0, que nous présentons en Figure 1(b). Le décile moyen et la valeur moyenne de f_0 sont aussi des valeurs permettant de distinguer certains groupes de locuteurs.

À partir des valeurs de f_0 une distinction nette entre hommes et femmes peut être faite. Effectivement, nous voyons au milieu de la Figure 1(b) un écart entre le dernier locuteur, 43, et la première locutrice, 36. Par conséquent, nous comprenons que les confusions entre locuteurs de sexes différents faites à partir des valeurs d'intensité peuvent disparaître avec cette autre dimension prosodique. La seule similarité qui n'est pas complètement effacée par la prosodie dans le cas de la locutrice 28, est avec la locutrice 24. Ces deux locutrices sont très écartées selon les valeurs d'intensité mais confondues dans tous les autres cas, même à partir des spectrogrammes. À partir des mesures que nous avons à notre disposition lors de cette étude nous pouvons affirmer que ces deux locutrices se distinguent de l'ensemble des locuteurs par leur intonation mais nous n'arrivons pas à déterminer quel élément prosodique peut nous permettre de les différencier nettement entre elles.

3.2 Cas de non-complémentarité

Les deux cas que nous venons d'évoquer montrent les deux tendances de la majorité de nos locuteurs : une caractérisation accentuée par l'intensité suivie d'un apport mineur mais complémentaire de la f_0 pour réduire les confusions ; ou des taux de classification légèrement différents entre les deux mesures. Dans les deux cas nous observons une augmentation importante du taux de bonnes réponses lorsque la représentation conjointe de f_0 et d'intensité est utilisée.

Cependant sur les 44 locuteurs, nous avons remarqué 6 cas dans lesquels la prosodie totalise un score égal ou inférieur à celui des deux mesures prises singulièrement : les locuteurs 11, 39, 41, 44 et les locutrices 32, 37. Parmi ces locuteurs, un premier groupe comprend les locuteurs 41 et 44 dont l'intensité totalise le même score des deux dimensions conjointes, avec des taux de bonnes classifications respectivement de 31 % et de 48 %. D'un autre côté, nous avons les locuteurs 11 et 39 qui ont aussi une meilleure classification à partir de l'intensité mais des scores inférieurs dans tous les autres cas. Enfin les locuteurs 32 et 37 atteignent un score plus important à partir de leur f_0.

Pour comprendre quels éléments influencent la non complémentarité nous avons décidé d'observer le locuteur 11 et ses confusions avec le locuteur 27, puisque ce dernier présente quant à lui des indices prosodiques complémentaires. Selon ses valeurs de f_0, le locuteur 11 est plus souvent identifié comme le 27 et inversement. En reprenant la Figure 1(a) comme exemple, nous observons que ces deux locuteurs ont des valeurs très proches et légèrement moins élevées que la moyenne. La f_0 et l'intensité du locuteur 11 ne montrant pas de complémentarité, il continue à être confondu avec le locuteur 27 à partir de la représentation conjointe de ces indices alors que cet autre locuteur atteint un taux de reconnaissance beaucoup plus important et ne montre pas de confusion avec le locuteur 11.

Nous avons essayé d'expliquer cette non complémentarité en analysant la relation entre les variations d'intensité et de f_0 à l'intérieur des séquences analysées. Nous avons observé cette relation sous forme de ratio et de différence des valeurs de f_0 et d'intensité. Nos observations montrent que le ratio n'apporte aucune information supplémentaire, pas de significativité statistique observée contrairement à la valeurs de la différence, et nous l'avons exclu de l'analyse statistique présentée dans le paragraphe suivant. La différence nous permet d'écarter les locuteurs 11 et 27 et d'autres cas similaires. Cependant,

ceci ne nous permet pas d'expliquer complètement pourquoi ces locuteurs sont moins caractérisés par la prosodie. Des indices contenus dans le spectrogramme – mais qui ne font pas partie de nos indices prosodiques – permettent ainsi d'améliorer les taux de classification de ces locuteurs. Effectivement, les locuteurs 11 et 27 atteignent respectivement 40 % et 57 % de bonnes classifications avec les deux indices conjoints alors qu'à partir du spectrogramme, ils atteignent respectivement 92 % et 94 %. Ces locuteurs, mais aussi les autres cités dans ces cas de non-complémentarité, atteignent à travers le spectrogramme plus de 90 % de bonnes classifications, nous pouvons supposer par conséquent qu'ils font partie d'un groupe de locuteurs dont la prosodie seule ne peut pas expliquer la variation et pour lesquels une représentation globale est nécessaire.

3.3 Analyse en Composantes Principales

Nous avons effectué une Analyse en Composantes Principales pour l'ensemble des mesures acoustiques afin de mieux comprendre leurs interactions. La Figure 2(a) montre les valeurs moyennées pour chacun des 44 locuteurs dans l'espace obtenus à partir des deux premières composantes. Nous n'avons retenu que les deux premières pour la représentation puisqu'elles expliquent ensemble 87.5 % de la variance, respectivement 55.3 % et 32.2 %. La Figure 2(b) montre la contribution des mesures acoustiques à la définition de chacune des composantes. Nous observons que les mesures de f_0 et la différence entre f_0 et intensité, dernière ligne du tableau, contribuent de manière importante à la première dimension. L'intensité contribue plutôt à la définition de la deuxième composante. Cette division nous confirme la non corrélation et par conséquent la complémentarité des deux paramètres. Seul l'écart type de l'intensité contribue de manière importante à définir la troisième composante, restant presque absent des deux premières. Tous les locuteurs sont représentés par les points dans la

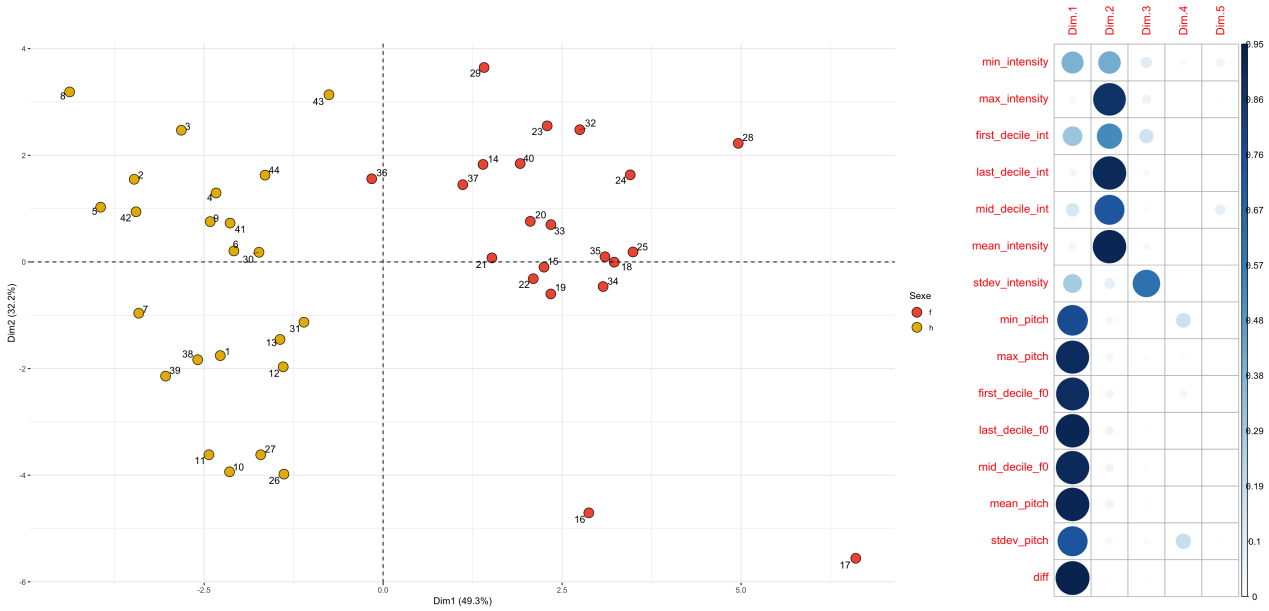


FIGURE 2 – Résultats de l'Analyse en Composantes Principales. À gauche (Figure 2a) la distribution des individus selon les dimensions 1 et 2 est représentée. La Figure de droite (2b) montre la contribution de nos mesures acoustiques à la définition de chaque composante

Figure 2(a) où l'espace défini par les deux premières composantes permet d'isoler 10 locuteurs : 3, 7, 8, 16, 17, 24, 28, 29, 36 et 43. Nous remarquons la correspondance entre certains de ces locuteurs et les cas de complémentarité des indices mis en évidence par la classification automatique.

La contribution des indices acoustiques n'étant pas la même pour les deux premières composantes, nous observons l'axe horizontal de l'ACP qui respecte la division par sexe grâce aux mesures de f_0. Sur le côté gauche, en jaune, sont disposés tous les locuteurs et sur la droite en rouge toutes les locutrices. Comme observé précédemment aussi (Figure 1b), le locuteur 43 et la locutrice 36 marquent la limite entre les deux sexes. À travers cette analyse nous pouvons également observer comment les paramètres prosodiques permettent de différencier 10 groupes de locuteurs plus ou moins compacts. Certains locuteurs sont plus proches que d'autres dans la représentation de l'ACP. Nous retrouvons parmi les groupes de locuteurs les mêmes confusions présentes lors de la classification à travers les deux paramètres conjoints. Certaines confusions, 38-39 et 11-10 par exemple, montrent une convergence phonétique pour les locuteurs d'un même binôme mais ne représentent qu'une minorité. Ceci renforce les observations faites dans le paragraphe précédent, concernant certains locuteurs qui sont moins caractérisés à partir de leur prosodie mais pouvant être ainsi associés entre eux et ensuite différenciés selon d'autres paramètres présents dans la production de la parole.

4 Discussion et conclusion

Dans ce travail nous avons présenté une analyse acoustique pour caractériser des locuteurs à partir de f_0 et d'intensité. L'utilisation d'une classification par réseaux de neurones et l'analyse des confusions opérées par le réseau nous ont aidé à comprendre dans quelle mesure le système utilise la prosodie comme facteur déterminant pour classer les locuteurs. Si la f_0 se montre très efficace pour trancher entre des locuteurs de sexes différents l'intensité atteint un score global de classification plus élevé. Nous avons ici analysé la contribution de ces dimensions prosodiques face au spectrogramme qui représente la référence puisqu'il englobe d'autres dimensions en plus de la prosodie et permet ainsi d'expliquer la majorité de la variation. Pour le sous-groupe de données que le réseau n'arrive pas à classer correctement en se basant sur le spectrogramme (7 % des données totale) l'utilisation conjointe du contour de f_0 et d'intensité arrive à donner 33 % de bonnes classifications.

La f_0 et l'intensité sont des paramètres classiques utilisés pour décrire deux aspects différents de la prosodie. Nous avons pu observer, à travers notre étude, comment l'utilisation d'une technique de classification non classique en phonétique peut faire apparaître des nouvelles caractéristiques liées à la description prosodique des locuteurs. L'Analyse en Composantes Principales, présentée en Figure 2, montre comment avec une approche statistique les mesures de f_0 contribuent à la description d'une partie conséquente de notre ensemble. L'intensité obtient un meilleur score que la f_0 lors de la classification par réseaux de neurones mais elle est moins présente dans l'ACP.

Pour diminuer cette disparité, il est envisageable d'aller regarder de manière plus approfondie d'autres mesures d'intensité plus robustes que celles que nous avons pu présenter. Nous avons observé pour ce paramètre que les valeurs minimales et le dernier décile sont les mesures plus précises pour la caractérisation des locuteurs. En ce qui concerne la f_0, la valeur moyenne et le décile du milieu sont les mesures qui se montrent plus adaptées pour la caractérisation.

Remerciements

Nous remercions l'ANR VOXCRIM (ANR-17-CE39-0016) ainsi que le LaBeX Empirical Foundations of Linguistics (EFL).

Références

ADAMI A. G., MIHAESCU R., REYNOLDS D. A. & GODFREY J. J. (2003). Modeling prosodic dynamics for speaker recognition. *Proceedings of ICASSP*, p. 788–791.

ARVANITI A. & RODRIQUEZ T. (2013). The role of rhythm class, speaking rate, and f0 in language discrimination. *Laboratory Phonology*, **4**.

BOERSMA P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5 :9/10**, 341–345.

DELLWO V., LEEMANN A. & KOLLY M.-J. (2015). Rhythmic variability between speakers : Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, **137**(3), 1513–1528.

GAUVAIN J., LAMEL L. & ADDA G. (2002). The limsi broadcast news transcription system. *Speech Communication*, **37**, 89–108.

HE L., GLAVITSCH U. & DELLWO V. (2015). Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability : an artificial neural networks approach. *Proceedings of ICPhS*.

HUDSON T., DE JONG G., MCDOUGALL K., HARRISON P. & NOLAN F. (2007). F0 statistics for 100 young male speakers of standard southern british english. *Proceedings of ICPhS*, p. 1809–1812.

KAHN J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale*. Thèse de doctorat, ED 536.

KEATING P. & KUO G. (2012). Comparison of speaking fundamental frequency in english and mandarin. *Journal of Acoustical Society of America*, **132**, 1050–1060.

LINDH J. & ERIKSSON A. (2007). Robustness of long time measures of fundamental frequency. *Proceedings of Interspeech*, p. 2025–2028.

MATHWORKS (2019). *MATLAB Deep Learning Toolbox R2019a*. Mathworks, Natick, MA, USA.

NIEBUHR O. & SKARNITZL R. (2019). Measuring a speaker's acoustic correlates of pitch - but which ? a constrastive analysis based on perceived speaker charisma. *Proceedings of ICPhS*, p. 1774–1778.

PARDO J. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, **119**.

SCHÖTZ S. (2007). Analysis and synthesis of speaker age. *Proceedings of ICPhS*.

SORIN C. (1981). Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, **9**, 359–374.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**, 201–212.

TWEEDY R. S. & CULLING J. F. (2014). Does the signal-to-noise ratio of an interlocutor influence a speaker's vocal intensity ? *Computer Speech and Language*, **28**, 572—-579.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

C'est "*mm-hm*, *oui*" ou "*mm-hm*, *non*" ? Propositions pour une grammaire des composantes acoustiques des *interactions nasalisées*

Aurélie Chlébowski, Nicolas Ballier
Université de Paris, CLILLAC-ARP, F-75013 Paris, France
aurelie.chlebowski@hotmail.fr, nicolas.ballier@u-paris.fr

RÉSUMÉ

Cet article se propose d'envisager l'existence d'une grammaire spécifique aux *interactions nasalisées* (Chlébowski et Ballier, 2015). Notre proposition se fonde sur une annotation des composantes acoustiques de cette sous-catégorie de *sons non-lexicaux* (Ward, 2006) dans le corpus CID (Bertrand et al., 2008). Nous voudrions présenter les contraintes combinatoires et régularités qui semblent s'appliquer à ces composantes acoustiques, ainsi que discuter leur structuration. Les résultats préliminaires de l'analyse des composantes acoustiques semblent suggérer des plages de valeurs par défaut pour les réalisations des IN (notamment pour la durée). La violation de ces usages peut donner lieu à une analyse de type gricienne d'implicature.

ABSTRACT

A modest proposal for the pragmatic of nasal grunts in the CID corpus.
This paper proposes to consider that a specific grammar might govern the production of *nasal grunts* (Chlébowski et Ballier, 2015). This proposal is based on an annotation of the acoustic components of this sub-category of *non-lexical conversational sounds* (Ward, 2006) in the CID corpus (Bertrand et al., 2008). We would like to show the combinatory constraints and regularities that apply to these acoustic components as well as debate their potential organisation. Our preliminary observations suggest that certain ranges of realisations are expected (especially for duration), the violation thereof leading to Gricean implicature.

MOTS-CLÉS : interactions nasalisées, compositionalité, CID.
KEYWORDS: nasal grunts, compositionality.

1 Introduction

D'abord exclus du champ de l'analyse linguistique (soit parce qu'ils étaient considérés comme événements marginaux, soit au prétexte qu'ils entraveraient la *bonne* production et compréhension du message principal), les sons « non-lexicaux » (entres autres dénominations, *cf.* Dingemanse, 2020) ont, depuis, suscité la curiosité, de par leur omniprésence dans la conversation. Ces dernières décennies témoignent d'un regain d'intérêt significatif pour leur analyse, principalement motivé par le désir de rendre compte de leurs multiples rôles en interaction. On peut citer, entre autres, les études qui interviennent dans la compréhension ou le partage de l'état émotionnel, attitudinal ou cognitif du locuteur (Duez, 2001 ; Corley et Hartsuiker, 2011), celles qui s'intéressent au niveau de maîtrise d'une langue et à sa perception (Cenoz, 1998 ; Niebuhr et Fischer, 2019) ou encore celles qui traitent de la structuration des tours de parole (Peters et Wong, 2015). Quels que soient leurs

Volume 1 : Journées d'Études sur la Parole, pages 100–108.
HAL : hal-02798522.

motivations, méthodes ou supports d'analyse (enregistrements, transcriptions, et même, corpus écrits), ces études se rejoignent pour ce qui est de leur approche fondamentalement onomasiologique : elles reposent sur des catégories majoritairement fonctionnelles, voire sémantiques (*backchannels*, *disfluences*, pour ne citer que celles-ci). Certaines ont proposé des hypothèses quant à la relation qu'entretiennent le sens et la forme des sons non-lexicaux (Duez, 2001 ; Clark et Tree, 2002 ; Dingemanse et al., 2013, entre autres) mais pas d'explication unificatrice sur la capacité qu'aurait une même forme sonore (telle qu'ils la posent *a priori*, hors toute considération de variation) d'endosser différentes fonctions ou de dénoter diverses significations (Ward, 2006). Néanmoins, ces études suggèrent que la production de ces sons n'est pas arbitraire : pour les employer correctement, il faut en apprendre et en comprendre l'usage. D'autres études se sont proposées de s'émanciper des catégories fonctionnelles susnommées. Compte tenu de la nature acoustique des *sons* non-lexicaux, Ward (2006) défend une approche sémasiologique, qui part des signes, et plus particulièrement d'une étude multidimensionnelle du signifiant sonore, pour remonter vers le concept. Il postule que ces sons sont régis par des règles différentes de celles qui s'appliquent au lexique : le sens transmis par un son non-lexical provient en grande partie de sa composition acoustique.

A défaut d'une sémantique, qui serait assez difficile à définir et sujette à caution pour la formulation de tests de perceptions qu'il faudrait établir (Chlébowski, 2016), nous proposons de discuter des régularités et contraintes qui nous paraissent à l'œuvre dans l'agencement des composantes acoustiques des sons non-lexicaux. Nous focalisons notre étude sur l'analyse des composantes des *interactions nasalisées* (ci-après, IN) dans le corpus CID (*Corpus of Interactional Data* ; Bertrand et al., 2008). La section suivante pose le cadre théorique, explicite les composantes acoustiques et problématise leur superposabilité. La section 3 présente le traitement des données tirées du CID. La section 4 esquisse la typologie des contraintes observées et la dernière section discute les résultats et conclut.

2 Cadre théorique

Cette section se propose de poser le cadre théorique dans lequel se situe notre étude, en explicitant d'abord la notion de *modèle compositionnel* (Ward, 2006), puis en proposant une représentation des composantes acoustiques des IN monosyllabiques telles qu'on peut les observer dans le CID.

2.1 Le modèle compositionnel de Ward (2006)

L'analyse acoustique des sons non-lexicaux n'est pas sans précédent et certaines études ont laissé entrevoir l'importance de la forme sonore de ces sons pour signifier une intention de sens. Néanmoins, la majorité d'entre elles se focalise sur les réalisations acoustiques d'une unique interprétation fonctionnelle de ces sons non-lexicaux. Les hypothèses formulées sont alors contraintes, selon nous, par les limites qu'impose cette catégorie unique en ce qu'elles ne permettent pas de rendre compte de ce qui est, ou n'est pas, partagé par différentes réalisations fonctionnelles (*backchannels*, *disfluences* …) des sons non-lexicaux. Ward (2006) propose, pour l'anglais américain, de répertorier tous les sons non-lexicaux d'un corpus de taille moyenne, indépendamment des catégories fonctionnelles qui leur sont généralement prêtées. Son étude, principalement auditive, des formes acoustiques de ces sons est fondée sur une conception *décompositionnelle* de leur substance sonore. Il propose un modèle componentiel (*compositional model*) selon lequel les sons non-lexicaux sont décomposables en un nombre fini d'entités acoustiques (que nous appellerons ici des *composantes acoustiques*). Il relève une dizaine de

composantes (/h/ ou voix soufflée, voix craquée, nasalité, /o/, /a/, entre autres) qui se présenteraient de manière superposée (voire concaténée dans de rares cas) et dont le nombre de combinaisons est difficilement estimable. Selon ce modèle, chaque composante acoustique d'un son non-lexical porte une valeur sémantique qui s'exprime indépendamment du contexte dans lequel le son est produit. La signification du non-lexical, contrairement au lexical, n'est pas à chercher dans la succession des signifiants. La production des composantes acoustiques dans les sons non-lexicaux n'étant pas astreint à la linéarité du signifiant, il serait possible de dériver le sens d'un son non-lexical par l'inspection de sa composition acoustique. Bien que la saillance perceptive de certaines composantes soit notable (comme les variations de mélodie ou de durée), chaque composante d'un son non-lexical apporte bien une contribution au sens final du son non-lexical. Il n'est alors pas question de hiérarchiser les composantes mais plutôt de les traiter sur le même plan. Nous illustrerons l'intérêt du modèle compositionnel de Ward par l'exemple suivant :

(1a) A : J'ai fait rentrer le chat.
B : Hein ?
A : J'ai fait rentrer le chat.

(1b) A : J'ai fait rentrer le chat.
B : Ah ?
A : Oui, il miaulait à la porte.

Posons que les composantes à l'œuvre dans les sons non-lexicaux dans (1a) et (1b) sont les mêmes, à l'exception de la qualité du segment prononcé par B (/ɛ̃/ *vs.* /a/). Sans entrer dans les détails, on remarque que la réponse de A est cohérente avec le son non-lexical produit par B. L'issue de la conversation varie considérablement selon le timbre vocalique choisi par B, ce qui semble accréditer l'idée qu'une seule composante acoustique de l'IN (telle qu'ici le timbre vocalique) puisse être signifiante. On perçoit ici les relations qu'entretiennent sens et forme sonore d'un son non-lexical mais aussi la reconnaissance de ces relations par le locuteur, qui est capable de les employer, et par l'interlocuteur, qui est capable de les décoder (Duez, 2001). Au-delà des applications sémantiques et pragmatiques impliquées par la sélection des composantes acoustiques dans la constitution d'un son non-lexical, nous estimons qu'il est d'abord nécessaire de déterminer la nature de ces composantes (et leur inventaire) dans une langue donnée. A la différence du modèle de Ward (2006), nous estimons que toutes les composantes acoustiques n'ont peut-être pas le même statut combinatoire.

2.2 Anatomie des composantes acoustiques des IN monosyllabiques autonomes

La représentation stratificationnelle proposée en Figure 1 (à gauche), qui superpose les composantes d'une IN monosyllabique *autonome* (voir section 3), nous paraît réaliser le programme de recherche de Ward. L'acte de production (ou de *vocalisation*) d'une IN contraint la présence de certaines composantes acoustiques, notamment : un son syllabique nasal, une durée, une valeur de fréquence fondamentale (F0), un degré d'intensité et un registre. Comme exposé dans Chlébowski (2020), nous considérons ces composantes comme *fondamentales* à la production d'une IN par rapport à d'autres qui sont, de fait, *optionnelles*. Chaque strate de cette représentation offre un éventail de valeurs possibles. Par exemple, la variation de la F0 peut être déclinée en contours *plat*, *montant* ou *descendant*. L'aspect catégoriel ou continu des valeurs que peut revêtir chaque composante n'est pas tranché (Ward, 2006). Il est tout à fait possible d'imaginer qu'il existe un continuum de *montant* à *descendant* (en passant par le niveau *plat*) au niveau de la 3^{e} strate, de *bouche fermée* à *bouche ouverte* pour la 1ère strate (Chlébowski, 2020), ou encore continuum de *haut* à *bas* (en passant par le niveau *médian*) pour la 5^{e} strate.

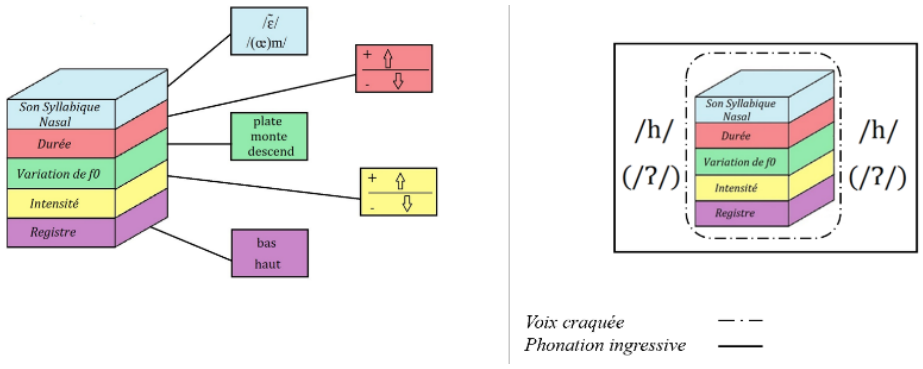


FIGURE 1 : Modèle stratificationnel des composantes *fondamentales* (à gauche) et modèle pour l'interaction avec les composantes *optionnelles* (à droite), pour les IN autonomes dans le CID.

Dans notre représentation, les composantes *optionnelles* (qui, selon nous, ne relèvent pas de paramètres pathologiques ou stylistiques) sont : la voix craquée, /h/ ou voix soufflée, la voix ingressive et le coup de glotte. Ces composantes sont sujettes à des contraintes physiologiques, voire phonotactiques : certaines seront davantage successives que simultanées. La propriété essentielle de leur combinatoire est la possibilité d'être potentiellement non-congruentes avec la durée totale des composantes principales de l'IN. Nous avons représenté ces composantes *additionnelles* à droite sur la Figure 1. La composante /h/, ou voix soufflée, peut apparaître en début ou en fin de IN. Lorsqu'elle est en position finale, elle est toujours discrétisable, contrairement à la position initiale où elle est susceptible de se propager sur les composantes principales de l'IN. Il semblerait que les coups de glotte aient une distribution comparable (sans la propagation). La phonation ingressive, quant à elle, commence avant les composantes principales de l'IN et se prolonge (fréquemment, après leur réalisation dans l'IN). Le craquement lui, est co-occurrent avec le son syllabique mais présente la particularité de pouvoir être discontinu. Dans une optique compositionnelle, cette capacité de désynchronisation des composantes optionnelles permettrait de varier le sens transmis par une IN[1].

3 Constitution et traitement des données

Dans la continuité des travaux de Ward (2006), Chlébowski et Ballier (2020) ont proposé un protocole pour l'annotation des composantes acoustiques de 947 *interactions nasalisées* (*nasal grunts*)[2] sélectionnées dans le corpus CID. Les IN y sont définies comme une sous-catégorie de sons non-lexicaux qui partagent dans leur combinaison le trait acoustique de la *nasalité*. Les items

[1] Bien que cela complique les analyses, on pourrait proposer de considérer que ces composantes soient elles-mêmes divisibles en sous-composantes, par exemple : durée, intensité et position vis-à-vis des composantes principales.

[2] Chlébowski et Ballier (2015) emploient le terme *grunt* dans leur étude sur les *nasal grunts* en anglais Geordie, suivant (Ward, 2000), qui entérine à la fois le caractère non technique et stigmatisant du terme. La traduction de ce terme par « *grognement* » n'est pas, à notre connaissance, attestée dans cet emploi en français. Notons que l'ensemble de ses synonymes renvoie à des verbes de parole connotés négativement (*grommellement*, *bougonnement*, *marmonnement*, *ronchonnement*). Notre terme *interactions nasalisées* souligne l'aspect profondément interactionnel de ces sons.

susceptibles d'entrer dans cette catégorie dans le CID sont alors les suivants : *hein, han, hum* (ou *euh+mh*) et *mh* (ou *hm*). Ils soutiennent que l'observation d'indices acoustiques minimiserait la subjectivité qu'induit la perception auditive sur ces sons. Ils ont proposé un guide d'annotations de leurs composantes acoustiques, basées majoritairement sur des critères visuels, conduites sous Praat (Boersma et Weenink, 2019). Le tableau 1 ci-dessous récapitule la nature des observations proposées pour les composantes des IN. Les composantes *segment*, *registre* et *syllabation* ont, pour diverses raisons que nous ne re-exposeront pas ici, échappé à une analyse exclusivement visuelle. De plus, si le repérage de /h/ et de la phonation ingressive peut se faire visuellement, la distinction entre les deux nécessite l'écoute de l'annotateur. Il semble que seule la durée soit directement objectivable. Ils proposent une liste d'étiquettes restreintes, pour non seulement signifier la présence d'une composante mais aussi déterminer sa localisation dans l'IN. Ils soulignent que leurs annotations se veulent préliminaires à des études plus approfondies des corrélats acoustiques et de leurs interactions et reconnaissent que leur protocole d'annotation réalisé par un annotateur unique devra être soumis à validation inter-juges[3].

D'autre part, Chlébowski et Ballier (2020) proposent d'annoter la position de l'IN dans l'énoncé du locuteur de l'IN (en début, milieu, fin d'énoncé ou comme étant *isolée* du reste de l'énoncé par des pauses) et en termes d'interaction entre locuteur et interlocuteur (voir la distinction en termes *self vs. others* dans Fruehwald (2016)). Nous avons retenu pour le présent article les IN prononcées *à l'isolée* dans l'énoncé du locuteur et qui peuvent donc être considérées comme *autonomes*[4]. En revanche, nous n'avons pas tenu compte de la position de l'IN en interaction.

Composantes	**Type d'observation**
Segment ; registre	Auditive
Syllabation Distinction entre /h/ et phonation ingressive	Visuelle et auditive
Variations de la F0 ; coup de glotte ; voix craquée ; /h/ ; phonation ingressive	Visuelle
Durée	Par mesure à partir des frontières des annotations
Intensité	

TABLE 1 : Degrés de la subjectivité des observations de Chlébowski et Ballier (2020)

4 Typologie des contraintes observées

Nous avons montré en 2.2 l'importance de contraintes physiologiques pour le statut superposable (craquement) ou concaténable (coup de glotte) des composantes acoustiques. Des

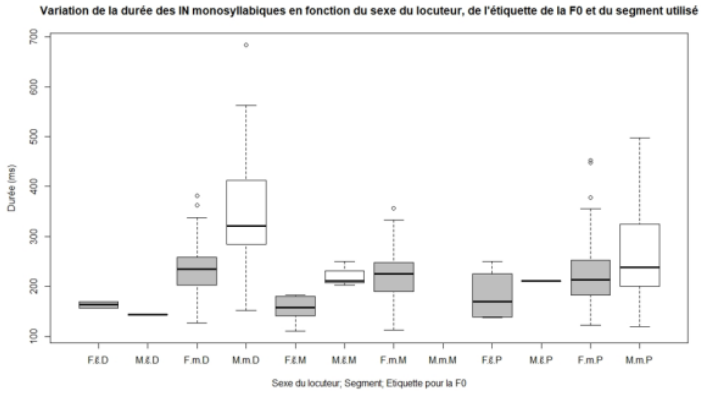
[3] Les procédures envisageables sont, en fonction des composantes, visuelles, auditives et/ou automatiques. Pour la composante *syllabation*, par exemple, nous pourrions mettre en place des tests d'annotation croisée auditifs ou visuels (basés sur la recherche d'indices acoustiques). Une approche plus instrumentale pourrait être envisagée. Par exemple à l'aide de la détection automatique de pics syllabiques (De Jong et Wempe, 2009).

[4] A l'inverse des IN prononcées en début, fin ou milieu d'énoncé (par exemple, *hein* en tant que marqueur de discours), les IN entourées par des pauses ne sont pas constituantes d'une unité prosodique plus importante qu'elles-mêmes.

contraintes d'ordre phonotactique peuvent également être observées. Ward (2006) et Chlébowski (2020) notent des combinaisons limitées pour certaines composantes. On retrouvera par exemple la succession /j/ + /e/ dans *yeah* (en anglais américain) ou la succession /œ/ + /m/ dans *hum* (en français dans le CID). A l'inverse, les séquences */e/+/j/ et */m/+/œ/, ne sont pas attestées et fort peu probables. Une partie de la distribution des formes des IN recensées pourrait s'expliquer par le jeu d'interrelations entre les composantes. Chlébowski et Ballier (2020) notent que la phonation ingressive se rencontre essentiellement avec les IN de type segmental /m/ et /ɑ̃/, mais pas avec celles de type segmental /ɛ̃/. Dans cette section, nous voulons montrer des tendances distributionnelles des composantes des IN mono- et dissyllabiques de type segmental /ɛ̃/ ou /m/ dans le CID.

4.1 Contraintes distributionnelles pour les monosyllabiques autonomes en voix modale

Nous concentrons nos premières observations sur la durée, la catégorisation du contour, et le sexe du locuteur pour les IN monosyllabiques autonomes en voix modale de type segmental /ɛ̃/ ou /m/. Les IN de type segmental /m/ semblent être préférées en position *isolée*, et ce, indépendamment du sexe du locuteur (tableaux de contingence en Figure 2). Le contour montant (M) est préféré pour les monosyllabes de type segmental /ɛ̃/ et les contours descendant (D) et plat (P) pour les monosyllabes de type segmental /m/ (p-valeurs pour le test exact de Fisher de 0,0009 et 0,002 pour les locuteurs et locutrices du CID, respectivement). En ce qui concerne les variations de durée (voir Figure 2, à gauche), nous rejoignons les résultats de Chlébowski et Ballier (2020) sur un autre jeu de données. L'hétéroscédasticité des données a été contrôlée avec un test de Levene (p-valeur < 0,001) sous R (R Core Team, 2017) et les résultats de l'ANOVA montrent une forte interaction de la durée avec le sexe du locuteur et le type segmental de l'IN. La catégorisation du contour ne semble pas jouer de rôle majeur dans les variations de la durée d'une IN. En moyenne, les locuteurs masculins produisent des IN plus longues que celles locutrices féminines et les IN de type segmental /m/ sont souvent plus longues que celles de type segmental /ɛ̃/. Nous remarquons aussi que les IN de type segmental /ɛ̃/ ne dépassent pas une certaine durée de production (jusque 250 ms.) alors que les IN de type segmental /m/ présentent une plus large plage de variation de durée (jusque 684 ms.). Certes, il est physiologiquement possible de réaliser un monosyllabe de type /ɛ̃/ d'une certaine durée, mais l'absence d'observations de durée supérieure à 250 ms. pourrait laisser penser qu'une réalisation de type /ɛ̃/ n'est pas compatible avec l'interprétation sémantique d'une durée plus longue. On voit par là qu'il pourrait exister des plages de valeurs par défaut qui sont susceptibles d'interagir avec l'interprétation.



	/ɛ̃/	/m/	TOTAL (M)
D	1	29	30
M	3	0	3
P	1	7	8
TOTAL (M)	5	36	41

	/ɛ̃/	/m/	TOTAL (F)
D	2	61	63
M	8	40	48
P	4	143	147
TOTAL (F)	14	244	258

FIGURE 2 : Variation de la durée des IN monosyllabiques autonomes en voix modale en fonction du sexe du locuteur (F/M), du segment et de l'étiquette de la F0 (à gauche) et tableaux de contingence pour ces mêmes paramètres, à l'exception de la durée (à droite).

4.2 Dissymétrie de la durée dans les IN dissyllabiques

Nous ne reviendrons pas sur les difficultés du découpage syllabique exposées dans Chlébowski et Ballier (2020)[5]. Nous voudrions argumenter en faveur d'une attente des locuteurs en termes de réalisations des IN dissyllabiques en montrant que, par défaut, la durée de la première syllabe semble inférieure à celle de la seconde (Figure 3). Chlébowski (2020) propose que cette distribution s'explique non pas par des contraintes de production (il est physiologiquement possible de produire la première partie de l'IN plus longue que la seconde) mais plutôt par ce qui est de l'ordre des attentes pragmatiques. Allonger la première syllabe d'une IN serait interprété dans un contexte interactionnel de la même manière que pour les violations du principe de coopération de Grice (1975) : comme une implicature. S'il est conventionnel que la première syllabe soit plus courte, l'allonger produirait un effet non-conventionnel souvent sarcastique, glosable, entre autres, par « ah ouais, tu m'en diras tant ». En d'autres termes, cette interprétation sarcastique de l'IN fonctionnerait à la manière d'une méta-règle (Ballier, 2005) : si la durée de l'IN ne respecte pas la répartition attendue, l'interaction est d'ordre sarcastique (et pas dans l'approbation, pour reprendre la problématique du titre de notre article). Nous proposons de vérifier cette hypothèse par un exemple en contexte. Sur la Figure 3, les IN pour lesquelles le premier segment et nettement plus long que le second (environ 100 ms.) sont les IN 905 et 896.

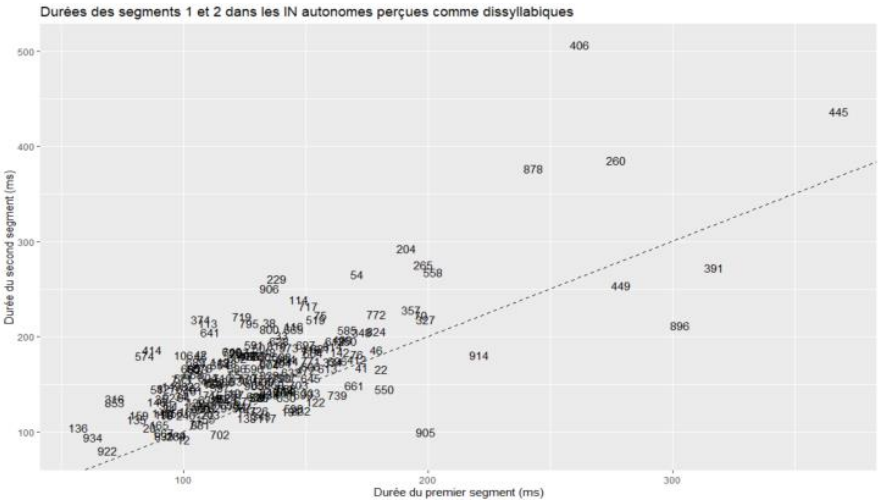


FIGURE 3 : Durées des premiers et seconds segments des IN autonomes perçues comme dissyllabiques (la nature de la réalisation de la découpe syllabique n'est pas prise en compte)

Nous focaliserons notre analyse sur l'IN 905 de la conversation entre les participants AP et LJ (bien que L'IN 896 soit passible d'une même analyse). Dans cette conversation, AP explique à LJ comment il s'est retrouvé à faire la cuisine, sans aucune expérience, dans un restaurant peu recommandable, à l'hygiène et aux méthodes douteuses. L'exemple (2) ci-dessous présente un extrait de cette conversation.

[5] Nous soutenons néanmoins leur hypothèse qui pose que la présence des composantes /h/ ou *coup de glotte* en position médiane d'une IN fonctionne comme un marqueur de coupe syllabique (/m.hm/ et /mʔm/). Les résultats d'une première expérience d'annotation croisée (impliquant trois participants, étudiants en phonétique) semblent aller en ce sens, avec plus de 75% d'accord sur la forme dissyllabique des IN lorsqu'elles sont présentées avec un /h/ ou un coup de glotte médian. D'autres analyses sont en cours.

(2)	LJ (576,72)	t'as fait la cuisine antillaise @@
	AP (577,90)	ouais cuisine antillaise ben la cuisine antillaise en fait je croyais que c'était vachement dur quoi et en_fait j'étais un super cuistot
	LJ (582,75)	@@
	AP (583,01)	il suffisait d'ouvrir euh les les d- de piocher dans les trois congélateurs
	LJ (586,28)	IN#905
	AP (586, 81)	de prendre euh les les trucs tout faits

S'il n'est pas perceptible de prime abord dans la transcription donnée en (2), le sarcasme impliqué par cette IN l'est à l'oreille, selon nous. L'IN sarcastique « tu m'en diras tant, piocher dans les congélateurs, c'est pas être un super cuistot » semble alors tout à fait à propos pour ponctuer le côté absurde du contexte.

5 Discussion et conclusion

Cet article, s'il ne constitue que des propositions et observations, a néanmoins le mérite de proposer une interprétation systémique des résultats obtenus. Nous avons voulu montrer le caractère superposable et/ou concaténable des composantes acoustiques à l'œuvre dans les IN du CID, en rendant compte de certaines contraintes et régularités.

Au-delà de l'importance des contraintes physiologiques qui semblent diviser les composantes en ce qu'elles seraient *fondamentales* ou *optionnelles,* nous avons suggéré l'importance que joue la sémantique dans leur combinatoire. Il est nécessaire de distinguer les impossibilités physiologiques de certaines productions (à la manière des articulations jugées impossibles des cases noires du tableau de l'API) et leur improbabilité, qu'elle soit agrammaticale (cas de *mh* + *euh*) ou apragmatique (produire la première syllabe d'une IN dissyllabique plus longue que la seconde). Les composantes acoustiques des IN seraient sujettes à une sorte de « grammaire » (Chlébowski, 2020) où la sélection des composantes serait principalement motivée par des choix sémantiques susceptibles d'exploitations pragmatiques, comme on vient de le voir. Notre analyse s'inscrit dans un courant plus général qui milite en faveur d'une grammaire de l'interaction (Ginzburg et Poesio, 2016).

Nous avons bien conscience qu'un certain nombre de questions restent en suspens. Entre autres, pour ces IN, selon les langues considérées, se pose toute la question de la grammaticalisation des composantes acoustiques (l'utilisation et l'investissement sémantique de telle ou telle composante acoustique). Les contraintes de production (ou les habitudes articulatoires ?) ne sont pas nécessairement les mêmes en français et en anglais, ou même entre variétés d'une même langue (Chlébowski, 2020). Le CID étant un corpus de français méridional, il ne semble pas exclu d'avoir des réalisations septentrionales qui soient différentes. Est-ce à dire pour autant qu'il n'y pas d'universal pour ces *interactions nasalisées* ? Nous défendons que seule une approche intégrative (Ward, 2006) des études sur les sons non-lexicaux pourrait nous permettre de mieux comprendre leur *modus operandi*. En ce sens, nous ne considérons pas notre approche comme *meilleure* mais comme *complémentaire* de toute autre étude sur le sujet.

Remerciements

Nous tenons à remercier Roxane Bertrand pour nous avoir donné accès au CID et les deux relecteurs anonymes pour leurs commentaires.

Références

Ballier, N. (2005). De l'exception à la régulation des exceptions pour la phonologie de l'anglais: d'un système de métarègles?. *Faits de Langues*, *25*(1), 245-253.

Bertrand, R., Blache, P., Espesser, R., Ferre, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique Des Langues*, *49*(3), 105-134.

Boersma, P. & Weenink, D. (2019). *Praat: doing phonetics by computer.* Version 6.1.06.

Cenoz, J. (1998). Pauses and Communication Strategies in Second Language Speech. *Reports-Research*, *143*, 1-11

Chlébowski, A. (2016). The Meaning of "Nasal Grunts" in the Necte Corpus. A Preliminary Perceptual Investigation. *Research in Language*, *14*(1), 43–59.

Chlébowski, A. (2020). A Semasiological Approach to Non-Lexical Conversational Sounds: Issues, Benefits and Impact. In Proc. *Laughter and Other Non-Verbal Vocalisations Workshop*

Chlébowski, A., & Ballier, N. (2015). "Nasal grunts" in the NECTE corpus, Meaningful interactional sounds. In Proc. *EPIP4-4th International Conference on English Pronunciation: Issues & Practices*, 54–58.

Chlébowski, A., & Ballier, N. (2020). A Manually Annotated Resource for the Investigation of Nasal Grunts. In Proc. *LREC 12th*

Clark, H. H., & Tree, J. E. F. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Corley, M., & Hartsuiker, R. J. (2011). Why *um* helps auditory word recognition: The temporal delay hypothesis. *PloS One*, *6*(5), e19792.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Dingemanse, M. (2020). Between sound and speech: Liminal signs in interaction. *Research on Language and Social Interaction*, *53*(1), 188–196.

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "*Huh*?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PloS One*, *8*(11), e78273.

Duez, D. (2001). Signification des hésitations dans la parole spontanée. *Revue Parole*, 17–18.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 41–49.

Ginzburg, J., & Poesio, M. (2016). Grammar is a system that characterizes talk in interaction. *Frontiers in psychology*, *7*, 1938.

Grice, H. P. (1975). Logic and conversation, syntax and semantics. *Speech Acts*, *3*, 41–58.

Niebuhr, O., & Fischer, K. (2019). Do not hesitate!–Unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In Proc. *Interspeech 2019*, 544–548.

Peters, P., & Wong, D. (2015). Turn management and backchannels. In *Corpus pragmatics: A handbook* (408–429). Cambridge University Press.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Ward, N. G. (2000). Issues in the transcription of English conversational grunts. In Proc. *1st SIGdial Workshop on Discourse and Dialogue*, 29–35.

Ward, N. G. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*(1), 129–182.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Variation prosodique des styles de parole et interface syntaxe-prosodie: Étude sur corpus à grande échelle

George Christodoulides[1]
(1) Service de Métrologie et des Sciences du Langage, Université de Mons,
18 Place du Parc, 7000 Mons, Belgique
george@mycontent.gr

RÉSUMÉ

La mutualisation et diffusion des grands corpus de parole permet de réexaminer des analyses précédentes effectuées sur des corpus plus petits, afin de vérifier si les conclusions de ces analyses se généralisent aux nouvelles données. Dans cette étude, nous présentons les résultats préliminaires d'une analyse de la variation des styles de parole en français, basée sur un corpus à grande échelle (300 heures, 2500 locuteurs). Le corpus a été réaligné au niveau des phones, syllabes et mots, et une annotation morphosyntaxique et syntaxique a été ajouté en améliorant les annotations existantes. Plusieurs caractéristiques acoustiques et prosodiques sont automatiquement extraites et une analyse statistique (analyse en composantes principales, ACP) est effectuée afin d'explorer les caractéristiques des styles de parole et leur variance. Nous explorons aussi la relation entre frontières prosodique et syntaxiques comme méthode pour discriminer les styles de parole. [1]

ABSTRACT

Speaking Style Prosodic Variation and the Prosody-Syntax Interface : A Large-Scale Corpus Study

As large spoken language corpora become available, we can revisit previous analyses based on smaller datasets and verify whether the conclusions generalise to the new data. We present an analysis of speaking style variation in French, based on a large-scale corpus (300 hours, 2500 speakers). The corpus was segmented at the phonetic, syllabic and word level. Automated annotation in parts-of-speech and syntactic dependencies was performed, enhancing existing annotations, and a multitude of acoustic and prosodic features were automatically extracted. Statistical analysis (principal component analysis, PCA) is performed to explore the characteristics of speaking styles and their variance. We finally explore congruency and mismatch between prosodic and syntactic boundaries, as a method to discriminate speaking styles.

MOTS-CLÉS : style de parole, variation prosodique, classification et regroupement (clustering) de styles de parole, interface prosodie-syntaxe, linguistique de corpus.

KEYWORDS: speaking style, prosodic variation, classification and clustering of speaking styles, prosody-syntax interface, corpus linguistics.

1. Cet article est une traduction en français de la communication Christodoulides, G. (2020) Speaking Style Prosodic Variation and the Prosody-Syntax Interface : A Large-Scale Corpus Study, *10th International Conference on Speech Prosody*, 24-28 May 2020, Tokyo, Japan.

Volume 1 : Journées d'Études sur la Parole, pages 109–117.
HAL : hal-02798523.

1 Introduction

La variation situationnelle, c'est-à-dire la variation linguistique liée aux différentes situations de communication, est importante pour l'étude du langage et de la parole. L'étude systématique de cette variation, à la croisée de la sociolinguistique et de la phonétique/phonologie, est du domaine de la sociophonétique. Dans ce contexte, la pertinence de la dichotomie entre « parole de laboratoire » et « parole spontanée » a été remise en question, étant insuffisante pour décrire la grande variation observée dans des corpus qui recensent plusieurs situations (Wagner *et al.*, 2015). Les caractéristiques de la parole sont déterminées à la fois par le contexte situationnel et les différences individuelles (Llisterri, 1992; Eskenazi, 1993; Léon, 1993). L'apport du contexte situationnel à un style de parole (« phonostyle ») est mieux décrit en utilisant plusieurs dimensions (cf. le modèle proposé dans (Koch & Österreicher, 2012)) plutôt que des distinctions binaires (p.ex. « discours formel » vs « discours informel »).

La variation de caractéristiques prosodiques selon les styles de parole ont été étudiés en analysant des corpus (Goldman *et al.*, 2014; Beliao *et al.*, 2013), ou parfois en vue d'une application spécifique, comme la classification automatique des échantillons sonores (Veiga *et al.*, 2012). La mutualisation et la diffusion de plus grands corpus de parole nous amène à revoir les analyses précédentes qui étaient basées sur des corpus plus petits, afin de vérifier si les résultats se généralisent, ainsi que pour effectuer des analyses plus fines. Cette contribution présente une étude sur un corpus de français parlé à large échelle (voir section 2.1) et sa méthodologie s'inspire de (Goldman *et al.*, 2014). Nous présentons également une analyse de la relation entre segmentation prosodique et syntaxique, et de sa variation selon les styles de parole, sur la base du même corpus.

2 Méthodologie

2.1 Corpus

La présente étude a été réalisée sur le sous-corpus de parole du *Corpus d'Étude pour le Français Contemporain* (CEFC) (Benzitoun *et al.*, 2016). Il s'agit d'une collection de corpus du français parlé provenant de diverses sources (Branca-Rosoff *et al.*, 2012; Cresti *et al.*, 2005; Baldauf-Quilliatre *et al.*, 2016; Avanzi *et al.*, 2016; Carruthers, 2008; Équipe Delic, 2004) qui ont été homogénéisés, transcrits, annotés automatiquement en parties du discours et en syntaxe des dépendances, et ont été approximativement alignés au niveau des tokens. La composition du corpus est présentée dans le tableau 1. Le corpus contient 900 échantillons, a une durée de 300 heures et contient plus de 2500 locuteurs. Sa taille est d'environ 3,7 millions tokens, ce qui après la phonétisation correspond à environ 4,6 millions de syllabes.

Une grande variété de situations de communication sont représentées dans le corpus CEFC, du fait qu'il est composé de plusieurs sources. Afin d'organiser les échantillons, les métadonnées du corpus CEFC proposent quatre dimensions principales : secteur (public ou privé), type/genre (fournissant une description large de l'activité ou de la situation communicative), milieu (p.ex. : amical, familial, affaires, politique, universitaire, etc.) et modalité (discours en public, face à face, radio, télévision, téléphone). Une distinction est également faite entre les monologues, les dialogues et les conversations multipartites. Nous avons combiné l'attribut secteur avec l'attribut type/genre (regroupant certaines situations de communication qui sont sous-représentées), afin d'arriver à une

catégorisation des styles de parole, qui est présentée dans le tableau 1. Dans cet article, nous présentons nos résultats principalement à travers ce regroupement en *genres* / *styles de parole* ; cependant, des analyses supplémentaires peuvent être effectuées sur les *sous-genres* plus détaillés, ou selon une autre dimension.

Style de parole	Nb	Dur	Loc	Tok	Syll
Privé					
Activité	10	2,3	14	21,4	27,4
Conversation	174	65,5	488	902,5	1075,2
Repas	12	8,0	39	102,0	120,7
Entretien	351	137,2	829	1672,3	2076,4
Narration	37	8,3	43	89,4	111,7
Public					
Activité	13	8,1	126	62,7	77,9
Conversation	81	6,2	209	67,6	86,5
Entretien	9	3,2	31	42,1	52,6
Didactique	16	4,3	63	35,3	49,2
Média	31	10,4	271	117,2	163,0
Réunion	50	28,1	319	348,6	443,6
Narration	86	15,8	96	148,2	188,3
Discours public	30	5,9	66	58,9	84,3
Total	900	303	2594	3668,4	4556,9

TABLE 1: Composition du corpus CEFC : nombre d'échantillons, durée (heures), nombre de locuteurs, syllabes (en milles) et tokens (en milles).

2.2 Traitement de données

Dans un premier temps et afin d'améliorer la performance de l'alignement automatique texte-parole, une procédure de restauration et d'amélioration a été appliquée à tous les échantillons audio du corpus, en utilisant le logicieel *iZotope RX 6 Audio Editor*. Les filtres suivants ont été appliqués en séquence : de-clip (restaurer les échantillons écrêtés à haute qualité), de-click (supprimer les clics aléatoires), de-hum (supprimer le bruit parasite et ses harmoniques), de-noise (réduction adaptative du bruit), equaliser (en mode dialogue) et leveller (normalisation des niveaux audio, respectant la dynamique du dialogue). Les métadonnées du corpus (fichiers XML encodés en TEI) et les annotations (fichiers CoNLL-U avec les informations sur les locuteurs et l'alignement approximatif de chaque token) ont tous été importés dans une base de données SQL à l'aide du logiciel de gestion de corpus *Praaline* (Christodoulides, 2014) pour le traitement ultérieur.

Une transcription phonétique avec des variantes de prononciation a été produite à partir de la transcription orthographique, et a été alignée au niveau du phone, de la syllabe, du token et de l'énoncé, en utilisant le système d'alignement forcé de *Praaline*, qui pour le français, utilise le système de reconnaissance vocale *Kaldi* (Povey *et al.*, 2011) et un lexique de prononciation basé sur *GLÀFF* (Hathout *et al.*, 2014).

Le corpus a été ré-annoté en utilisant *DisMo* (Christodoulides & Barreca, 2017) qui fournit une annotation morphosyntaxique détaillée, une détection des unités polylexicales ainsi qu'une annotation automatique des disfluences. Ces annotations ont été combinées avec l'annotation en syntaxe de dépendance originale. Le corpus aligné a été analysé à l'aide de *ProsoGram* (Mertens, 2004), qui détecte le noyau de chaque syllabe en fonction de la fréquence fondamentale (F0) et l'intensité ; la courbe F0 est ensuite stylisée selon un système perceptif, produisant une annotation en segments

tonales. Nous avons ensuite appliqué le plug-in *Promise* afin d'effectuer une détection automatique des syllabes proéminentes (Christodoulides & Avanzi, 2014) et une détection automatique des frontières prosodiques majeures et mineures (Christodoulides, 2018) ; les algorithmes statistiques de *Promise* ont été entraînés sur des corpus du français annotés manuellement, comme détaillé dans les références de l'outil. Une segmentation automatique en unités prosodiques (phrases intonatives et groupes accentués) a été effectuée sur la base de ces annotations, et est corrélée à l'annotation syntaxique (voir section 3.3).

Nous avons finalement appliqué les outils d'analyse statistique de *Praaline* (plug-ins d'analyse temporelle, profil prosodique et extraction d'unités) afin d'extraire plusieurs caractéristiques (mesures) acoustiques et prosodiques pour chaque échantillon du corpus et pour la participation de chaque locuteur à chaque échantillon. Ces mesures peuvent être regroupées comme suit : mesures temporelles (par exemple durée de pauses silencieuses et pleines, débit de parole) ; mesures de dynamique conversationnelle (longueur des tours de parole, pauses interlocuteurs et chevauchements) ; mesures prosodiques (ex. registre et mouvements dynamiques) ; mesures de proéminence ; et les mesures de segmentation (unités prosodiques, unités syntaxiques et leur corrélation). La base de données SQL de *Praaline* a été liée au logiciel *R* (R Core Team, 2017) pour l'analyse statistique.

3 Résultats

Dans ce qui suit, nous présentons des résultats statistiquement significatifs, selon la classification en *styles de parole*, pour certains groupes de mesures.

3.1 Mesures temporelles et prosodiques

Le taux d'articulation (pourcentage du temps d'enregistrement pendant lequel un locuteur articule) par situation (style de parole) est présenté dans la figure 1. Le style "narration privée" (narration spontanée d'expériences personnelles), ainsi que le style "narration prof" (contes de fées récités par des professionnels) ont un taux plus faible, indiquant que les pauses silencieuses sont plus nombreuses et/ou plus longues. Une observation similaire peut être faite pour le style "prof-leçon" (conférences académiques et leçons scolaires), ainsi que le style "prof-public speech" (où les pauses sont principalement utilisées pour l'effet rhétorique).

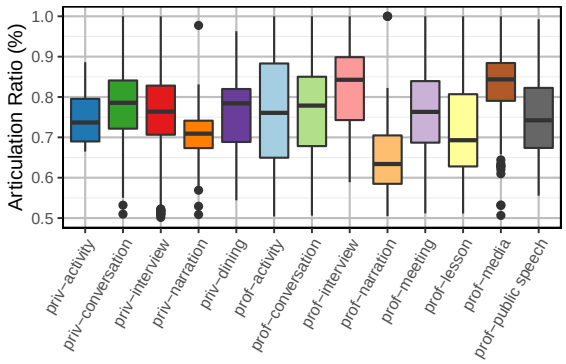


FIGURE 1 – Taux d'articulation (%) par situation.

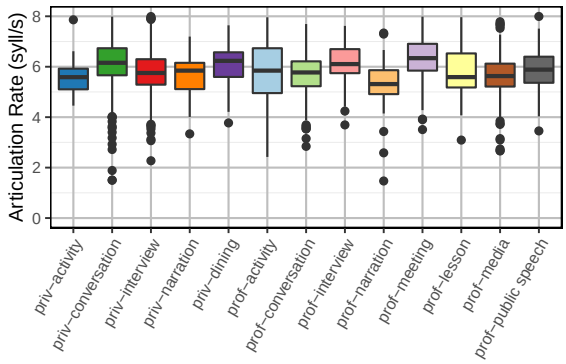


FIGURE 2 – Débit d'articulation (syll/s) par situation.

Nous observons également que la narration professionnelle a un débit d'articulation (syllabes par seconde) plus faible, et que la variabilité est plus élevée dans les conversations spontanées, tandis que les différences entre les autres styles de parole restent mineures. Le nombre de pauses pleines (normalisé au nombre de tokens) est présenté dans la figure 3 et est un descripteur prosodique qui discrimine les styles de parole avec un faible degré de planification, ou spontanés (p.ex. conversation, que ce soit dans un cadre privé ou professionnel, entretiens). Cependant, nous observons que les hésitations ainsi que les autres types de disfluences sont présentes dans tous les styles de parole (à l'exception de la narration professionnelle).

La figure 4 montre la distribution des trajectoires mélodiques (mouvements de F0 stylisée en demi-tons par seconde) par modalité de communication. Nous observons que les deux activités médiatiques (radio et télévision) ont des trajectoires mélodiques plus importantes, ce qui peut s'expliquer par l'adoption d'un style plus expressif par ces locuteurs professionnels.

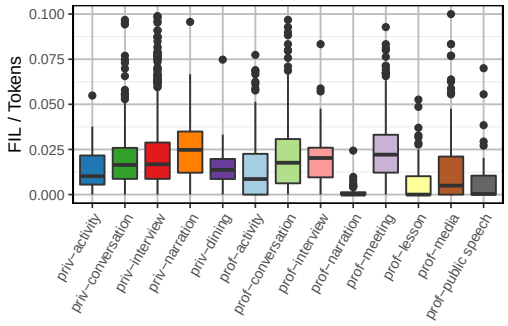


FIGURE 3 – Pauses pleines (normalisé au nombre de tokens) par situation.

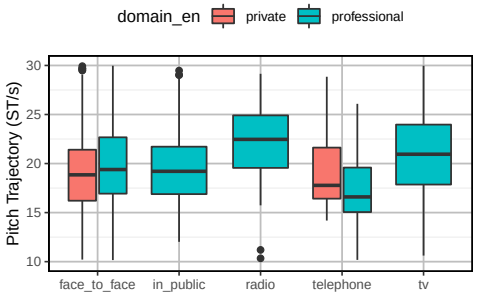


FIGURE 4 – Trajectoire mélodique (ST/s) par médium.

3.2 Dynamique conversationnelle

En ce qui concerne la dynamique conversationnelle, la figure 5 montre la durée moyenne (en secondes) des tours de parole.

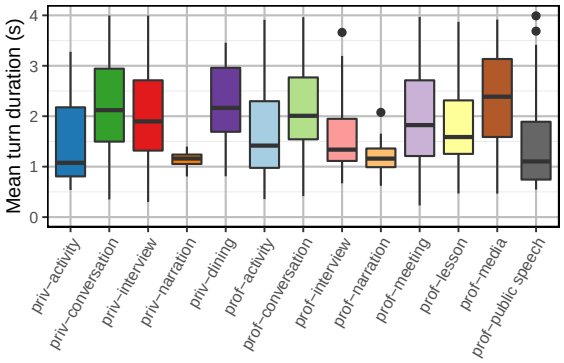


FIGURE 5 – Durée de tour de parole moyenne (s) par situation.

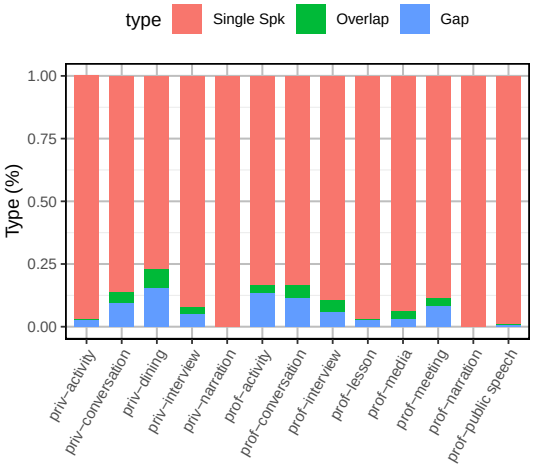


FIGURE 6 – Dynamique conversationnelle par situation.

Les différences entre les situations de communication plus ou moins interactives peuvent ainsi être observées, surtout si la durée du tour est combinée avec le pourcentage de chevauchements et de gaps (pauses interlocuteurs, entre deux tours de parole appartenant à des locuteurs différents). Cette distribution par situation est illustrée dans la figure 6.

3.3 Interface prosodie-syntaxe

Le corpus CEFC est annoté en syntaxe de dépendance en utilisant un nombre réduit de relations de dépendance, proposé dans le cadre du projet ORFEO. L'annotation syntaxique de la parole présente des difficultés et des choix doivent être faits concernant le traitement des faux départs (phrases inachevées), des parenthèses et des disfluences. Pour cette raison, dans la présente étude, nous avons utilisé l'annotation fournie et limité les analyses à la relation entre les unités syntaxiques majeures (c'est-à-dire les unités de dépendance complètes) et les unités prosodiques majeures. L'annotation en dépendance définit les unités syntaxiques (SU) et leur taille en syllabes, selon la situation de communication, est illustrée à la figure 7.

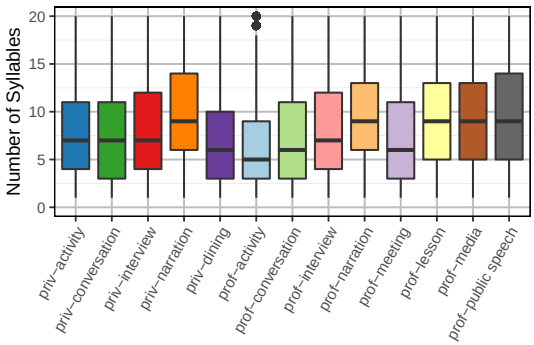


FIGURE 7 – Nombre de syllabes par unité syntaxique, par situation.

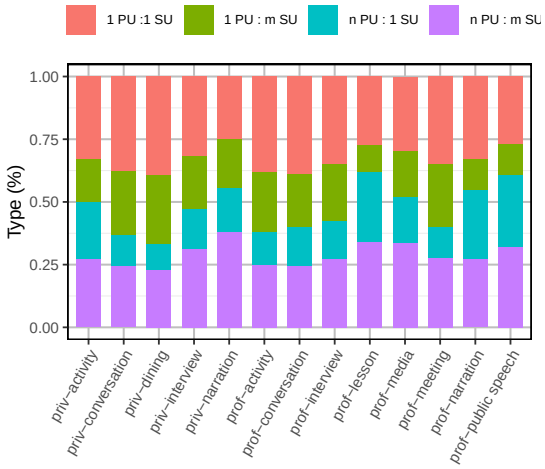


FIGURE 8 – Relation entre unités syntaxiques et prosodiques, par situation.

Sur la base de la détection automatique des frontières prosodiques et des syllabes proéminentes, nous avons annoté le corpus en unités prosodiques (PU) majeures (groupes intonatifs) et mineures (groupes accentués). Il existe des congruences et des décalages entre ces unités, créant quatre configurations possibles : un PU correspond à un SU (1 :1), un PU regroupe plusieurs SU (1 :m), plusieurs PU sont regroupés dans un SU (n :1), et une série de décalages entre les frontières prosodiques et syntaxiques qui conduisent à plusieurs PU correspondant à plusieurs SU (n :m). Cette méthode d'analyse est similaire à celles précédemment présentées par (Beliao *et al.*, 2013) et (Martin *et al.*, 2014) pour des corpus multi-genres du français parlé, de taille plus petite. Les résultats de l'analyse de la relation entre les unités syntaxiques et prosodiques (congruence/décalage), selon la situation, sont présentés dans la figure 8. Les situations de communication sont caractérisées par des différences dans la planification de la parole, et celles-ci affectent la longueur des structures syntaxiques, ainsi que le nombre et la durée des pauses silencieuses (qui sont le principal corrélat acoustique des frontières prosodiques majeures) ; on peut donc utiliser ces mesures de congruence / décalage entre les unités syntaxiques et prosodiques majeures afin de différencier (certains) styles de parole.

3.4 Analyse en composantes principales

Nous avons constaté qu'aucun paramètre prosodique unique n'est suffisant pour différencier les styles de parole. Étant donné que plusieurs mesures sont fortement corrélées entre elles, nous avons procédé en appliquant une analyse en composantes principales (ACP), qui réduit l'ensemble de mesures à un petit ensemble de *composantes principales* non corrélés linéairement (chaque composante principale est une combinaison linéaire des mesures initiales). Les résultats de l'ACP indiquent que les 2 premières composantes principales (PC) expliquent 25,1 % de la variance, les 4 premiers PC expliquent 44,6 % de la variance et 8 PC expliquent 71,0 % de la variance. On peut comparer ces résultats à ceux de (Goldman *et al.*, 2014) (avec 9 styles de parole et 105 échantillons), où les 2 premiers PC expliquaient seulement 43 % de la variance et les 8 premiers expliquaient 78,2 %.

Dans la figure 9, chaque point représente la participation d'un locuteur dans un échantillon de corpus, codé par couleur selon la situation / le style de parole, et est tracé sur un plan défini par les 2 premières composantes principales (PC1 sur l'axe x et PC2 sur l'axe y). Les ellipses de confiance indiquent la variabilité de chaque style de parole sont présentées dans la figure 10 : on peut ainsi observer que certains styles de parole sont très homogènes (par exemple les présentations médiatiques), tandis que les conversations et les entretiens sociolinguistiques ont une variabilité plus élevée.

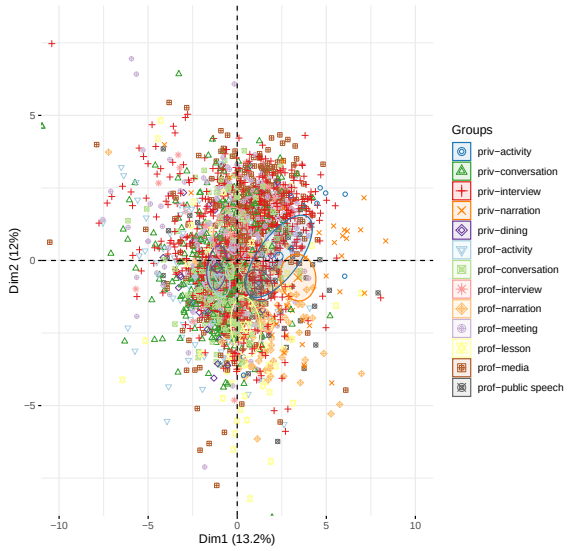


FIGURE 9 – Deux premières composantes principales, tous les échantillons.

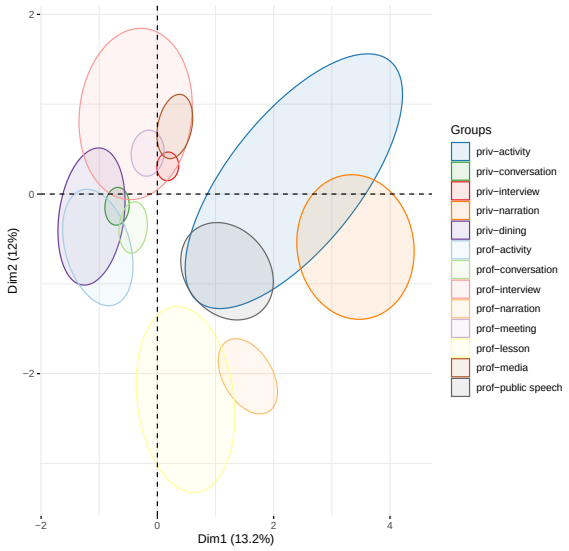


FIGURE 10 – Deux premières CP et ellipses de confiance pour chaque situation / style de parole.

4 Conclusion

Dans cet article, nous avons présenté une analyse de la variation prosodique des styles de parole dans un corpus à grande échelle (300 heures, plus de 2500 locuteurs) du français parlé. Alors que les tendances générales des études antérieures sur les corpus plus petits sont largement confirmées, l'analyse de ces «big data» indique qu'une approche beaucoup plus nuancée de la variation du style de parole est nécessaire : certaines situations de communication ont des contraintes très spécifiques (par exemple, le discours professionnel dans les médias, ou le discours politique en public, mais aussi narrations spontanées d'expériences personnelles) qui influencent plusieurs paramètres prosodiques

à la fois et sont ainsi plus faciles à distinguer et à classer. Cependant, la variation individuelle des conversations spontanées, voire des entretiens sociolinguistiques, est plus élevée. Ce résultat de notre étude sur grand corpus renforce les appels à encore plus de diversité dans la recherche sur la parole : un effort doit être fait non seulement pour étudier les phénomènes phonétiques et prosodiques à travers les différents styles de parole, mais aussi pour intégrer la variation individuelle dans les analyses. De plus, dans des études futures, nous envisageons une analyse plus détaillée de l'interface prosodie-syntaxe (entre autres, en incluant plusieurs niveaux d'unités syntaxiques et prosodiques) ; une exploration de méthodes alternatives pour décrire les styles de parole (un autre regroupement au niveau des métadonnées) ; et étendre les mesures extraites (en particulier au niveau segmental). Enfin, notre travail a enrichi et amélioré l'annotation d'un grand corpus du français parlé : le corpus est disponible sous la licence Creative Commons, et notre travail est rendu disponible sous les mêmes conditions.

Références

AVANZI M., BÉGUELIN M.-J. & DIÉMOZ F. (2016). De l'archive de parole au corpus de référence. Le corpus oral de français de Suisse romande (OFROM). *Corpus*, **15**, 309–342. Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

BALDAUF-QUILLIATRE H., COLON DE CARVAJAL I., ETIENNE C., JOUIN-CHARDON E., TESTON-BONNARD S. & TRAVERSO V. (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, **15**, 165–194. Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

BELIAO J., KAHANE S. & LACHERET A. (2013). Modéliser l'interface intonosyntaxique. In *Prosody-Discourse Interface Conference 2013, Proceedings*. DOI : 10.13140/2.1.1701.1205.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, **15**, 91–114. Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M. (2012). *Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*.

CARRUTHERS J. (2008). Annotating an oral corpus using the Text Encoding Initiative : Methodology, problems, solutions. *Journal of French Language Studies*, **18**(1), 103–119.

CHRISTODOULIDES G. (2014). Praaline : integrating tools for speech corpus research. In *LREC 2014 – 9th International Conference on Language Resources and Evaluation, May 26–31, Reykjavik, Iceland, Proceedings*, p. 31–34.

CHRISTODOULIDES G. (2018). Acoustic correlates of prosodic boundaries in french : A review of corpus data. *Revista de Estudos da Linguagem, Belo Horizonte*, **26**(4), 1531–1549. aop13597.2018.

CHRISTODOULIDES G. & AVANZI M. (2014). An evaluation of machine learning methods for prominence detection in french. In *Interspeech 2014 – 15th Annual Conference of the International Speech Communication Association, September 14–18, Singapore, Proceedings*, p. 116–119.

CHRISTODOULIDES G. & BARRECA G. (2017). Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo. *Corela : Cognition, Représentation, Langage*, **HS-21**. journals.openedition.org/corela/4867.

CRESTI E., BACELAR DO NASCIMENTO F., MORENO SANDOVAL A., VERONIS J., MARTIN P. & KALID C. (2005). *The C-ORAL-ROM Corpus. A Multilingual Resource of Spontaneous Speech for Romance Languages*. John Benjamins Publishing Company.

ÉQUIPE DELIC (2004). *Autour du Corpus de référence du français parlé*. Publications de l'université de Provence. Recherches sur le français parlé No 18, 265 pp.

ESKENAZI M. (1993). Trends in speaking styles research. In *Proceedings of Eurospeech*, p. 501–509.

GOLDMAN J.-P., PRŠIR T., CHRISTODOULIDES G. & AUCHLIN A. (2014). Speaking style prosodic variation : an 8-hour 9-style corpus study. In *7th International Conference on Speech Prosody, May 20–23, Dublin, Ireland, Proceedings*, p. 105–109.

HATHOUT N., SAJOUS F. & CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. In *LREC 2014 – 9th International Conference on Language Resources and Evaluation, May 26–31, Reykjavik, Iceland, Proceedings*.

KOCH P. & ÖSTERREICHER W. (2012). Language of immediacy – language of distance : Orality and literarcy from perspective of language theory and linguistic history. In C. LANGE, B. WEBER & G. WOLF, Éds., *Communicative spaces : Variation, contact, and change*, p. 441–473. Frankfurt : Peter Lang.

LLISTERRI J. (1992). Speaking styles in speech research. In *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language, Dublin, Ireland, 15–17 July 1992*, p. 15–17.

LÉON P. (1993). *Précis de phonostylistique, Parole et expressivité*. Paris : Nathan Université.

MARTIN L., DEGAND L. & SIMON A. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus*, **13**, 243—-265.

MERTENS P. (2004). The Prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proc. of Speech Prosody 2004, March 23–26, Nara, Japan*, p. 549–552.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

R CORE TEAM (2017). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

VEIGA A., CELORICO D., PROENÇA J., CANDEIAS S. & PERDIGÃO F. (2012). Prosodic and phonetic features for speaking styles classification and detection. In D. TORRE TOLEDANO, Éd., *Advances in Speech and Language Technologies for Iberian Languages. Communications in Computer and Information Science*, volume 328, p. 15–17. Berlin, Heidelberg : Springer.

WAGNER P., TROUVAIN J. & ZIMMERER F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, **48**, 1–12. DOI : 10.1016/j.wocn.2014.11.001.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Proximité rythmique entre apprenants et natifs du français Évaluation d'une métrique basée sur le CEFC

Sylvain Coulange[1] Solange Rossato[2]
(1) LIDILEM, Université Grenoble Alpes, France
(2) LIG, Université Grenoble Alpes, France
{sylvain.coulange, solange.rossato}@univ-grenoble-alpes.fr

RÉSUMÉ

Cette étude a pour objectif de proposer une quantification de l'accent étranger se basant sur des mesures rythmiques. Nous avons utilisé le Corpus pour l'Étude du Français Contemporain, qui propose plus de 300 heures de parole aux profils de locuteurs et aux situations variés. Nous nous sommes concentrés sur 16 paramètres temporels estimés à partir des durées de voisement et de syllabes. Un mélange gaussien a été appris sur les données de 1 340 natifs du français, puis testé sur des extraits de 146 natifs tirés au hasard (NS), sur ceux des 37 non-natifs présents dans le corpus (NNS), ainsi que sur des enregistrements de 29 apprenants japonais de niveau A2 d'un autre corpus. La probabilité que les NNS aient une log-vraisemblance inférieure aux NS ne dépasse pas la tendance ($p = 0,067$), mais celle pour les apprenants japonais est beaucoup plus significative ($p < 0,0001$). L'étude de la répartition des paramètres entre les différents groupes met en avant l'importance du débit de parole et des durées de voisement. [1]

ABSTRACT

Rhythmic Proximity Between Natives And Learners Of French – Evaluation of a metric based on the CEFC corpus

This work aims to quantify foreign accent in French based on rhythmic measurements. We used the Corpus pour l'Étude du Français Contemporain, which contains +300 hours of speech from a wide variety of speaker profiles and situations. We focused on 16 temporal parameters computed from voicing and syllables intervals. A Gaussian mixture model was trained on 1,340 native speakers of French, then tested with 146 natives (NS) and 37 non-native speakers (NNS) from the same corpus, as well as 29 A2-level Japanese learners of French from another corpus. The probability of NNS having inferior log-likelihood to NS was only a tendency ($p = .067$), but a much bigger probability was obtained for Japanese learners ($p < .0001$), where all speakers were A2 level. Parameter efficiency analysis reveals the importance of speech rate and voicing duration. [2]

MOTS-CLÉS : Modélisation du rythme, Accent étranger, Prononciation de la L2, Débit de parole, séquences voisées et non voisées.

KEYWORDS: Rhythm modelling, Foreign accent, L2 pronunciation, Speech rate, Voiced and Unvoiced sequences.

1. Une version anglaise de cet article est également parue dans les actes de la conférence internationale LREC 2020.
2. An English version of this article was also published in the proceedings of LREC 2020.

Volume 1 : Journées d'Études sur la Parole, pages 118–126.
HAL : hal-02798525.

1 Introduction

La perception de l'accent étranger est principalement due aux différences de prononciation entre ce que le locuteur dit et une norme attendue et partagée par les natifs de la langue cible (Alazard, 2013). Cette différence a été largement décrite au niveau segmental, à travers les théories de l'acquisition du langage, et le rôle de la prosodie a été amplement démontré, notamment par des études de perception (De Meo *et al.*, 2012; Pellegrino, 2012). Aujourd'hui, il est reconnu que le segmental et le suprasegmental jouent tous les deux un rôle important dans la perception de cette différence par les locuteurs natifs. Parmi les paramètres prosodiques, nous nous intéressons à des paramètres rythmiques, le rythme étant défini ici comme une récurrence de patterns de marquages forts ou faibles d'éléments dans un environnement temporel (Gibbon & Gut, 2001). Ces éléments peuvent être des alternances de syllabes longues et courtes, ou de segments vocaliques et consonantiques. Beaucoup d'études tentent d'ailleurs de classifier les langues en se basant sur les durées vocaliques et consonantiques, et nécessitent donc une transcription alignée. Des études ont toutefois montré des résultats similaires à partir des durées de voisement ou des durées de syllabes (Fourcin & Dellwo, 2013; Dellwo *et al.*, 2015), paramètres qui peuvent être quant à eux détectés automatiquement directement à partir du signal audio, sans recours à une transcription. C'est ce qui nous intéresse étant donné que la transcription automatique de la parole non-native reste encore difficile.

Les catégories rythmiques des langues sont cependant peu distinctes et la plupart des études font part de limites, dues à de trop petits effectifs de locuteurs, ou à des biais d'élicitation, facteurs qui influent aussi sur les caractéristiques rythmiques (Fourcin & Dellwo, 2013; Ramus *et al.*, 1999; Gibbon & Gut, 2001; Grabe & Low, 2002). Il est donc nécessaire d'utiliser un corpus volumineux, prenant en compte la variation selon les situations et les locuteurs, et notamment lors de production spontanée (Bhat *et al.*, 2010). Cela inclue également les enregistrements provenant de différents pays francophones, et de milieux sociaux variés.

Dans cette étude, nous proposons de modéliser le rythme du français à travers le récent Corpus d'Étude pour le Français Contemporain (CEFC, Benzitoun *et al.* 2016). Afin de modéliser l'ensemble de ces variations, nous avons entraîné un modèle de mélange gaussien sur 16 paramètres rythmiques. La log-vraisemblance obtenue avec ce modèle global pour un nouvel extrait de parole est comparée pour des locuteurs natifs et locuteurs non-natifs que nous avons réservés pour le test, ainsi qu'avec un corpus indépendant d'apprenants japonais du français. Nous avons également étudié la distribution de chaque paramètre entre les extraits de parole des locuteurs natifs et non-natifs.

2 Méthodologie

2.1 Corpus

Le corpus CEFC regroupe, uniformise et complète les annotations d'un ensemble ou de partie de 13 corpus, tels que Valibel [3] ou encore Clapi [4]. Les enregistrements peuvent provenir de différentes régions de France, de Belgique ou de Suisse. Les situations d'énonciation varient de conversations entre amis aux réunions professionnelles, en passant par des repas de famille, des débats médiatiques, des lectures de contes traditionnels ou encore des conversations enregistrées dans des magasins. Les

3. Valibel : https://uclouvain.be/fr/instituts-recherche/ilc/valibel/corpora.html.
4. Clapi : http://clapi.icar.cnrs.fr

extraits de parole sont constitués de dialogues (481 enregistrements sur 900), d'interactions à plus de 2 locuteurs (277) et des monologues (144). Ce corpus totalise environ 4 million de mots, avec 300 heures d'enregistrement. Tous les enregistrements sont transcrits et alignés, ce qui nous a permis d'identifier les extraits de parole de chaque locuteur. Sur un total de 2 587 locuteurs, les femmes représentent la majorité des locuteurs avec 1 373 locutrices, contre seulement 1 048 locuteurs, et 166 dont le genre n'est pas renseigné. [5]

Le corpus contient également la parole d'une cinquantaine de locuteurs non-natifs : ils ont été réservés pour la partition de test de notre modèle et exclus en totalité des enregistrements utilisés pour construire le modèle. En effet, les seules contraintes pour que l'extrait de parole soit inclus dans les données d'apprentissage du modèle est que la langue parlée soit le français, et que le locuteur soit natif. Le corpus de test est constitué d'environ 10% de l'ensemble des enregistrements de locuteurs natifs, choisis au hasard et mis de côté comme base de comparaison avec les locuteurs non-natifs.

Le niveau de français des locuteurs non-natifs n'étant pas précisé dans les métadonnées, nous avons inclus un second corpus de parole de locuteurs non-natifs, pour lequel nous connaissons la langue maternelle et le niveau de compétence en français de chaque locuteur. Ce corpus est constitué des enregistrements d'une évaluation en production orale pour 29 étudiants de l'Université de Langues Étrangères de Kyōto, de la même classe et tous de langue maternelle japonaise. Nous disposons également de leurs résultats à l'examen de fin de semestre, évaluant les 4 habiletés langagières : compréhension de l'oral et de l'écrit, production orale et écrite, et dont les enregistrements constituent une partie de l'évaluation.

2.2 Les paramètres acoustiques

Nous avons mesuré 16 paramètres largement utilisés soit pour la classification des langues (White & Mattys, 2007; Fourcin & Dellwo, 2013; Pettorino *et al.*, 2013, entre autres), la caractérisation des locuteurs (Rossato *et al.*, 2018), ainsi que la perception de l'accent étranger (Bhat *et al.*, 2010; Fontan *et al.*, 2018). Les paramètres rythmiques sont basés sur les segments voisés détectés par Praat [6], et les noyaux syllabiques détectés grâce à un script de De Jong & Wempe (2009) :

— Le débit de parole, ratio entre le nombre de noyaux syllabiques et la durée du segment SR ;
— Le pourcentage de voisement $\%V$;
— La moyenne μV, l'écart type σV et le coefficient de variation ρV des durées des intervalles voisés V_i ;
— La moyenne μU, l'écart type σU et le coefficient de variation ρU des durées des intervalles non-voisés U_i ;
— La moyenne μP, l'écart type σP et le coefficient de variation ρP des durées de la paire $P_i = V_i + U_i$;
— L'indice de comparaison brut $rPVI$ et normalisé $nPVI$ de couples successifs d'intervalles voisés (Grabe & Low, 2002; Fourcin & Dellwo, 2013) ;
— La moyenne $\mu\Delta NV$, l'écart type $\sigma\Delta NV$ et le coefficient de variation $\rho\Delta NV$ des durées ΔNV_i entre deux noyaux syllabiques successifs.

Les paramètres rythmiques sont calculés sur des segments constitués par la concaténation d'unités

5. Des statistiques plus détaillées sur les locuteurs du CEFC sont présentées dans Coulange (2019), mémoire de master de Sciences du langage parcours industries de la langue, Université Grenoble Alpes dirigé par Solange Rossato.
6. Praat : doing phonetics by computer. Version 6.0.37, téléchargée en mars 2019 depuis http://www.praat.org/.

entre pauses (UEP) [7] consécutives d'un même locuteur, jusqu'à atteindre une durée minimale de 30 secondes. Aucune UEP n'est coupée avant sa fin, il arrive donc que certains segments soient assez longs. Nous pensons que cette durée permet d'avoir suffisamment de parole pour obtenir des mesures fiables. Nous ne gardons que les locuteurs ayant au moins un segment et pour lesquels le statut de la langue française est connu (natif ou non). Le corpus d'apprentissage contient 16 884 segments de 1 340 locuteurs natifs. Les trois partitions de test sont constituées comme suit : 146 locuteurs natifs NS, 37 locuteurs non-natifs NNS et le corpus de 29 apprenants japonais JpNNS. Le tableau 1 récapitule ce partitionnement.

Set	Training	Test NS	Test NNS	Test JpNNS
French status	native	native	non-native	non-native
Corpus	CEFC	CEFC	CEFC	Jp learners
#speakers	1,340	146	37	29
#segments	16,884	1,919	268	96

TABLE 1 – Constitution du corpus d'apprentissage et des 3 corpus de test

2.3 Modèle de mélanges gaussiens

Selon Ferrer *et al.* (2015), les modèles de mélanges gaussiens (GMM) sont réputés pour modéliser la variation. Un GMM est une densité de probabilité calculée à partir d'une somme de gaussiennes pondérées. Cette fonction suit au mieux la distribution des données. L'apprentissage du modèle revient à trouver les meilleurs paramètres de ces gaussiennes (moyennes et matrices de covariance) et leur pondération pour représenter les données grâce à l'algorithme d'espérance-maximisation EM. La probabilité d'un vecteur $\vec{x}$, étant donné un GMM de paramètres $\{w_k, \vec{\mu}_k, \Sigma_k\}_{k=1}^{K}$ est alors :

$$p(\vec{x}) = \sum_{k=1}^{K} w_i \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) \tag{1}$$

où K est le nombre de gaussiennes, w_k est le poids de la gaussienne k, tel que $\sum_{k=1}^{K} w_k = 1$, et $\mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)$ la fonction normale de $\vec{x}$ de moyenne $\vec{\mu}$ et de covariance Σ de k. Nous utilisons la covariance diagonale pour alléger l'apprentissage, même si certains paramètres acoustiques sont corrélés entre eux. Nous avons également choisi de limiter notre GMM à 1 024 gaussiennes. L'apprentissage du modèle a été implémenté en Python grâce à la librairie *SciKit Learn* Gaussian Mixture [8].

Pour calculer la proximité de la parole d'un locuteur X au modèle, nous avons utilisé le produit de la vraisemblance du modèle pour chacun de ses segments de parole $\vec{x_n}$. Pour simplifier ces calculs, nous avons transformé le produit en somme en utilisant la log-vraisemblance $log\ p(X)$, et normalisé celle-ci par le nombre N de segments de chaque locuteur qui peut varier beaucoup en fonction des locuteurs :

$$log\ p(X) = \frac{1}{N} \sum_{n=1}^{N} log\ p(\vec{x}_n) \tag{2}$$

7. Toute pause supérieure à 1 seconde ou tout changement de locuteur mettant fin à une UEP.
8. https ://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

Nous avons calculé la log-vraisemblance moyenne de chaque locuteur natif de la partition de test (NS), et l'avons comparée à celle des locuteurs des corpus non-natifs (NNS ou JpNNS) avec un test de Wilcoxon-Mann-Whitney. Pour les locuteurs non-natifs, la log-vraisemblance obtenue par le modèle gaussien est interprétée comme un score de proximité rythmique au français et elle est comparée aux performances de l'apprenant évaluées par l'enseignant de français pour les locuteurs du Test JpNNS.

2.4 Comparaison des paramètres rythmiques entre natifs et non-natifs

Nous comparons la distribution de chacun des 16 paramètres entre les locuteurs natifs et les locuteurs non-natifs du test NNS ou ceux du test JpNNS. Or, le nombre de segments sur lesquels sont extraits ces paramètres est bien supérieur parmi les locuteurs natifs (1 919) par rapport à celui des segments disponibles pour les non-natifs du CEFC (268) et plus encore à celui des japonophones (96). La comparaison est faite sur un échantillon de 96 segments de paroles par corpus, sélectionnés aléatoirement. Pour les natifs, plusieurs tirages aléatoires sont effectués. Les comparaisons entre chaque paramètre mesuré s'effectuent donc sur 96 valeurs pour chaque groupe natif et non-natif. Un test t permet de tester si la différence observée entre les deux groupes est significative. Nous avons également calculé la valeur de l'êta-carré η^2 qui rend compte de la proportion de variance du paramètre expliquée par la variable natif/non-natif.

3 Résultats

3.1 Proximité rythmique des locuteurs

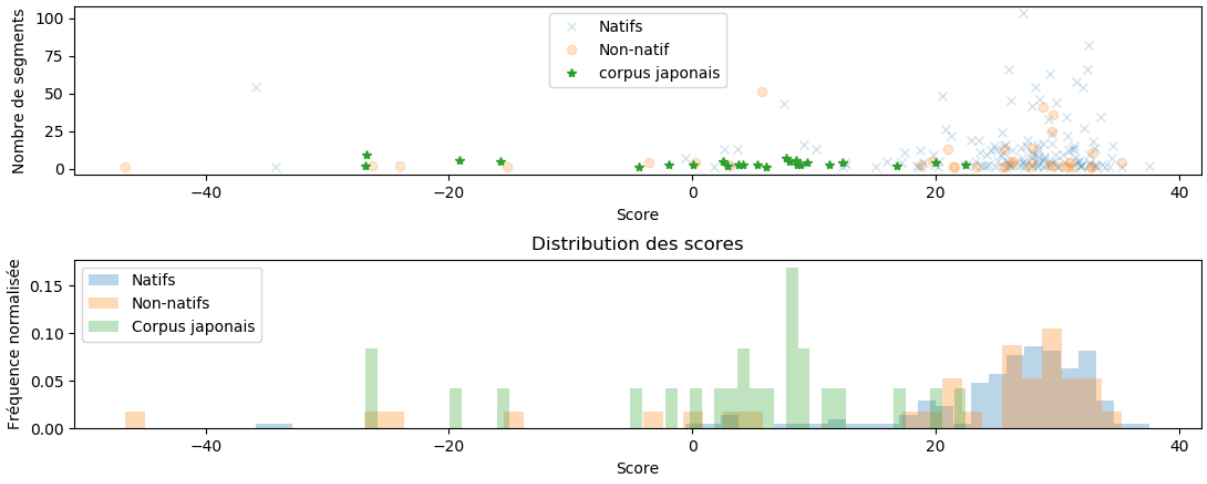


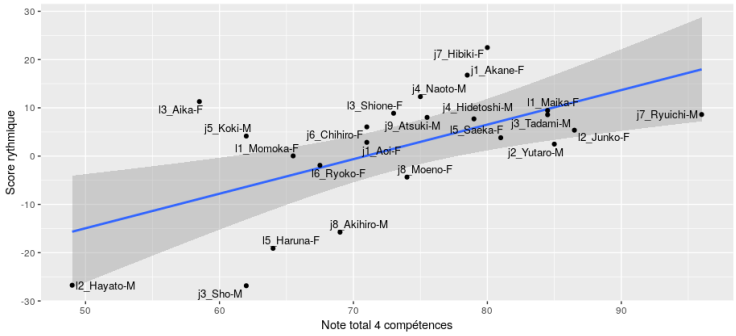
FIGURE 1 – Distribution des scores des locuteurs natifs NS (bleu), des non-natifs NNS (orange) et des apprenants japonais JpNNS (vert) (supérieurs à -50)

Nous avons d'abord comparé les scores de proximité rythmique des locuteurs natifs (NS) avec ceux des non-natifs du CEFC (NNS). Il s'avère que l'hypothèse selon laquelle les locuteurs non-natifs obtiennent un score de proximité rythmique inférieur à celui des natifs (et sont donc plus éloignés du modèle) ne dépasse pas la tendance ($p = 0.067$).

Nous avons ensuite comparé le score des natifs à celui des non-natifs du corpus d'apprenants japonais (JpNNS). Cette fois-ci, la différence est très significative ($p < 0.001$). L'écart entre les scores de proximité des locuteurs natifs et ceux des locuteurs non-natifs est bien plus net.

La figure 1 présente une distribution de ces scores, zoomée sur les scores supérieurs à -50 (ce qui correspond à 96,6% des NS, 94,6% des NNS et 79,3% des JpNNS). Les NS et les NNS sont majoritairement entre 20 et 40, tandis qu'on ne trouve aucun JpNNS avec un score supérieur à 22,48 (la majorité d'entre eux se situe en 0 et 10). La proximité moyenne de chaque population est respectivement de 25,0, 21,48 et 0,74, si on ignore les scores inférieurs à -50 qui pourraient être dus à de mauvaises détections de voisement ou de noyaux syllabiques à cause de voix trop faibles [9]. Dans cette figure, nous avons également fait apparaître le score de l'enseignant, francophone natif de la classe d'apprenants japonais, dont la voix a également été enregistrée. Son score est de 19,96.

3.2 Corrélation avec le niveau de compétence en langue



(a) Scores de proximité rythmique des apprenants en fonction de leur note globale à l'examen

	Global	Oral	Fluency
r	.598 (p = .003)	.257 (p = .237)	.410 (p = .052)
r^2	.358 (p = .003)	.066 (p = .237)	.168 (p = .052)
ρ	.478 (p = .021)	.315 (p = .144)	.228 (p = .295)

(b) Tests de corrélation entre les scores de proximité rythmique et les notes globales (gauche), de production orale (milieu) et de fluence (droite)

FIGURE 2 – Corrélation entre scores de proximité rythmique et niveau de français

Avec les enregistrements du corpus d'apprenants japonais, nous disposons également de 3 notes pour chaque étudiant : la note globale obtenue à l'examen de fin de semestre, type DELF, évaluant les 4 habiletés langagières (compréhension et production, orale et écrite), la note obtenue en production orale (PO) pour ce même examen, et le nombre de points obtenus spécifiquement pour la fluence du discours. La fluence est évaluée sur 5 points dans la partie de production orale, qui représente elle-même $\frac{1}{4}$ de la note globale sur 100 points. Les enregistrements du corpus sont ceux de la production orale des étudiants lors de l'examen, sur laquelle sont les notes de production orale et de fluence.

Nous avons mesuré la corrélation entre ces 3 notes et le score de proximité rythmique obtenu par chaque étudiant. Le tableau 2b donne le coefficient de corrélation de Spearman (r), le coefficient de détermination (r^2) et le coefficient de Pearson (ρ) pour chaque type de note : globale, production orale, et fluence ; avec les p-values associées. La note globale est assez bien corrélée avec le score de proximité rythmique ($r = .598$, $p < .005$; $r^2 = .358$, $p < .005$ et $\rho = .478$, $p < .05$). Les notes de production orale et de fluence sont quant à elles trop proches les unes des autres (17 à 24 pour la PO et 3 à 5 pour la fluence), et la corrélation n'est pas significative. La figure 2a montre les scores de proximité rythmique des 29 étudiants en fonction de leur note globale lors de l'évaluation.

9. Ces scores vont de -57,83 à -12 544,14 ; deux seulement sont compris entre -50 et -100.

3.3 Analyse des paramètres rythmiques

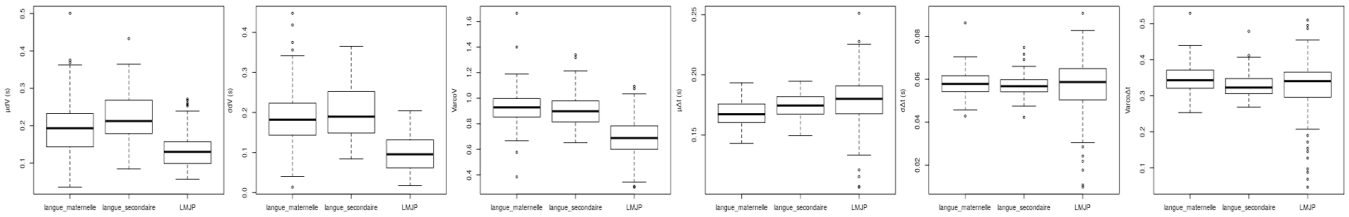


FIGURE 3 – Distribution de (μV), (σV), ρV des durées d'intervalles voisés ; et de $(\mu \Delta NV)$, $(\sigma \Delta NV)$ et $\rho \Delta NV$ des écarts entre noyaux syllabiques (respectivement NS, NNS et JpNNS).

La figure 3 illustre les distributions de certains paramètres en fonction des groupes NS, NNS et JpNNS. Pour certains paramètres, comme le débit de parole SR, on observe un comportement proche entre le groupe NS, et le groupe NNS avec respectivement 4,0 syll/s et 3,9 syll/s mais seulement 1,3 syll/s pour les JpNNS. De même, le pourcentage de voisement $\%V$ est de 57% pour les natifs, 58% pour les NNS et seulement 15% pour les JpNNS.

La durée moyenne des intervalles voisés μV est plus longue pour les locuteurs NS (190 ms) et pour les NNS (210 ms) que pour les JpNNS (130 ms). Les écarts de durée de ces intervalles sont toutefois plus importants chez les NS et les NNS que chez les apprenants japonais (σV à respectivement 180, 190 et 90 ms). ρV permet une mesure indépendante du débit de parole, il varie de 0,93 chez les natifs et 0,90 chez les NNS à 0,70 chez les JpNNS.

En ce qui concerne les métriques faisant intervenir les durées d'intervalles non-voisés, on constate des valeurs de μU d'environ 150 ms pour les NS et NNS, et de 580 ms pour JpNNS. σU reflète un même ordre de grandeur pour les NS et NNS (190 ms) et des variations très importantes pour les JpNNS (730 ms). ρU montre peu de différence entre les groupes : 1,21 pour les NS, 1,27 pour les NNS et 1,22 pour les JpNNS.

Paramètre	η^2	P-value		$\mu(\eta^2)$	$\sigma(\eta^2)$
SR	.745	$3.07e^{-58}$	<.0001	.705	.038
$\%V$	.673	$5.63e^{-48}$	<.0001	.647	.039
ρV	.415	$7.23e^{-24}$	<.0001	.391	.028
σV	.373	$4.79e^{-21}$	<.0001	.359	.029
$nPVI$	.360	$3.81e^{-20}$	<.0001	.339	.030
$rPVI$	.349	$1.99e^{-19}$	<.0001	.327	.035
μV	.281	$2.64e^{-15}$	<.0001	.246	.033
μU	.125	$4.78e^{-07}$	<.0001	.123	.003
μP	.118	$1.11e^{-06}$	<.0001	.116	.002
σU	.085	$3.94e^{-05}$	<.0001	.084	.002
$\mu \Delta NV$	.084	$4.37e^{-05}$	<.0001	.066	.017
σP	.082	$5.46e^{-05}$	<.0001	.081	.001
$\rho \Delta NV$	.046	.003	<.01	.029	.015
ρU	.012	.124	>.05	.012	.010
$\sigma \Delta NV$	.011	.157	>.05	.006	.004
ρP	.008	.206	>.05	.008	.008

FIGURE 4 – η^2 et p-value entre les NS et les JpNNS pour le 1er rééchantillonnage, et la moyenne et l'écart type des η^2 sur les 3 rééchantillonnages

Les indices de comparaisons $nPVI$ et $rPVI$ de durée entre couples successifs d'intervalles voisés présentent des valeurs très proches pour NS et NNS, et plus basses pour JpNNS (NS = 78,54% et NNS = 79,14% ; JpNNS = 64,57%), indiquant que les locuteurs JpNNS ont tendance a avoir moins d'écart de durée entre deux intervalles voisés successifs, ce qui est cohérent avec une valeur de σV plus faible. Lorsque l'on s'intéresse à l'écart temporel entre deux noyaux syllabique ΔNV, la valeur moyenne des durées séparant les noyaux syllabiques varie peu entre les groupes, avec 166 ms pour NS, 176 ms pour NNS et 180 ms pour JpNNS. Les écart-types et les coefficients de variations varient

peu entre les groupes mais montrent une dispersion plus importante pour les JpNNS (cf. les 2 derniers graphiques de la figure 3).

Nous avons donc voulu comparer NS et JpNNS afin de déterminer si, pour chaque paramètre, les différences observées étaient significatives ou non et calculé les êta-carrés. La figure 4 présente les êta-carrés (η^2) du premier rééchantillonnage, avec les p-values associées, ainsi que la moyenne des η^2 sur 3 rééchantillonnages successifs et leur écart type. On remarque que la variance du débit de parole SR est expliquée à 75% par le facteur natif/non-natif, suivi de près par le pourcentage de voisement ($\%V$, 67%). Arrivent ensuite les métriques impliquant les durées d'intervalles voisés : le coefficient de variation des durées d'intervalles voisés (ρV), l'écart type et la moyenne de durée de ces intervalles (σV, μV), comme les indices de comparaison brut et normalisé de leur paires successives ($nPVI$, $rPVI$). La variance de ces paramètres est expliquée pour 28 à 42% par le facteur natif/non-natif.

4 Discussion

Nous avons proposé une modélisation informatique du rythme du français, apprise sur un corpus volumineux de parole variée, et basée sur les mesures de 16 paramètres rythmiques détectés de manière entièrement automatique. Ce modèle a permis de mesurer un score de proximité rythmique à partir d'extraits de parole de minimum 30s. La comparaison des scores de proximité rythmique entre des locuteurs natifs et des locuteurs non-natifs du CEFC n'a pas montré de différence significative. Plusieurs raisons peuvent expliquer ce phénomène, comme l'hétérogénéité probable des niveaux de français chez les non-natifs du CEFC, la diversité de leurs langues maternelles ou encore les conditions d'enregistrement. Nous savons que les 37 locuteurs viennent d'au moins 18 pays différents, or le rythme de la langue maternelle (ou des langues maternelles) influence grandement l'acquisition des autres langues, tout comme d'autres facteurs individuels tels que la durée de séjour dans un pays où la langue cible est dominante, ou encore l'âge de première exposition à la langue (Piske *et al.*, 2001; Flege, 1988).

La comparaison d'un groupe plus homogène d'apprenants japonophones de niveau A2 a permis de montrer des différences significatives entre les scores de proximité rythmiques des apprenants et ceux des locuteurs natifs, avec une corrélation, certes peu élevée, avec le niveau global de français des étudiants japonais. Les notes de production orale et de fluence des étudiants ne nous ont pas permis de tirer de conclusion. Il nous faudrait réitérer l'expérience avec des notes plus détaillées, et plus hétérogènes entre les étudiants. Il serait également intéressant de mesurer la corrélation entre les scores rythmiques et les résultats d'un test de perception de l'accent étranger sur ces mêmes locuteurs.

Une étude plus fouillée des 16 paramètres confirme que le débit de parole est bien plus faible chez les apprenants, avec une diminution de la proportion de voisement, sans doute due à une augmentation des pauses intra UEP, inférieures à 1s. L'écart-type des intervalles voisés σV et sa version normalisée ρV ainsi que l'indice de comparaison des durées entre couples d'intervalles voisés successifs $rPVI$ et sa version normalisée $nPVI$ ne dépendent pas du débit pour leur version normalisée et montrent une tendance à uniformiser la durée des intervalles voisées chez les apprenants par rapport aux locuteurs natifs. Ces résultats corroborent les connaissances des enseignants qui savent que la parole des apprenants doit être accélérée, avec moins de pauses et plus de variation de durées de voisement pour se rapprocher de la parole native. Il serait intéressant maintenant de réitérer l'expérience avec différents niveaux de compétence en langue, et des langues maternelles appartenant à différentes familles rythmiques, pour voir comment varient les paramètres rythmiques en fonction de ces facteurs.

Références

ALAZARD C. (2013). *Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Étrangère*. Thèse de doctorat, Université Toulouse 2. Thèse de doctorat dirigée par Michel Billières et Corine Astesano.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet ORFÉO : un corpus d'études pour le français contemporain. *Corpus*, **15**, 91–114.

BHAT S., HASEGAWA-JOHNSON M. & SPROAT R. (2010). Automatic fluency assessment by signal-level measurement of spontaneous speech. *Second Language Studies : Acquisition, Learning, Education and Technology*.

DE JONG N. & WEMPE T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, **41**(2), 385–390.

DE MEO A., PETTORINO M. & VITALE M. (2012). Comunicare in una lingua seconda. Il ruolo dell'intonazione nella percezione dell'interlingua di apprendenti cinesi di italiano. In *La voce nelle applicazioni. Proceedings of the 7th Congress of Italian Association of Speech Sciences AISV*, p. 117–129.

DELLWO V., LEEMAN A. & KOLLY M.-J. (2015). Rhytmic variability between speakers : Articulatory, prosodic and linguistic factors. *The Journal of the Acoustical Society of America*, **137**(3).

FERRER L., BRATT H., RICHEY C., FRANCO H., ABRASH V. & PRECODA K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, **69**(C), 31–45.

FLEGE J. (1988). Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America*, **84**(1), 70–79.

FONTAN L., LE COZ M. & DETEY S. (2018). Automatically measuring l2 speech fluency without the need of ASR : A proof-of-concept study with Japanese learners of French. In *Interspeech 2018*, p. 2544–2548.

FOURCIN A. & DELLWO V. (2013). Rhythmic classification of languages based on voice timing. *Tranel Review*, p. 87–107.

GIBBON D. & GUT U. (2001). Measuring speech rhythm. In *EUROSPEECH 2001*, p. 95–98.

GRABE E. & LOW E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, **Vol. 7**, 515–546.

PELLEGRINO E. (2012). The perception of foreign accented speech. Segmental and suprasegmental features affecting degree of foreign accent in italian l2. *Mello H. et al. (Eds.) Proceeding of the 8 GSCP Conference*, p. 261–267.

PETTORINO M., MAFFIA M., PELLEGRINO E., VITALE M. & DE MEO A. (2013). *VtoV : a perceptual cue for rhythm identification*. University of Leuven (KU Leuven).

PISKE T., MACKAY I. & FLEGE J. (2001). Factors affecting degree of foreign accent in an l2 : a review. *Journal of Phonetics*, **29**(2), 191 – 215.

RAMUS F., NESPOR M. & MEHLER J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73**, 265–292.

ROSSATO S., ZHANG D., AJILI M. & BONASTRE J.-F. (2018). Suivre le rythme de tes paroles. In *Proc. XXXIIe Journées d'Études sur la Parole*, p. 37–45.

WHITE L. & MATTYS S. (2007). Calibrating rhythm : First language and second language studies. *J. Phonetics*, **35**, 501–522.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude comparative des paramètres d'entrée pour la synthèse expressive audiovisuelle de la parole par DNNs

Sara Dahmani[1] Vincent Colotte[1] Slim Ouni[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sara.dahmani@loria.fr, vinvent.colotte@loria.fr, slim.ouni@loria.fr

RÉSUMÉ

Dans le passé, les descripteurs contextuels pour la synthèse de la parole acoustique ont été étudiés pour l'entraînement des systèmes basés sur des HMMs. Dans ce travail, nous étudions l'impact de ces facteurs pour la synthèse de la parole audiovisuelle par DNNs. Nous analysons cet impact pour les trois aspects de la parole : la modalité acoustique, la modalité visuelle et les durées des phonèmes. Nous étudions également l'apport d'un entraînement joint et séparé des deux modalités acoustique et visuelle sur la qualité de la parole synthétique générée. Finalement, nous procédons à une validation croisée entre les résultats de la synthèse des différentes émotions. Cette validation croisée, nous a permis de vérifier la capacité des DNNs à apprendre des caractéristiques spécifiques à chaque émotion.

ABSTRACT

Comparative study of input parameters for DNN-based expressive audiovisual speech synthesis

In the past, contextual descriptors for acoustic speech synthesis have been studied for training systems based on HMMs. In this work, we study the impact of these factors for DNN-based audiovisual speech synthesis. We analyze this impact on the three aspects of speech : the acoustic modality, the visual modality and the duration of the phonemes. We also study the contribution of a joint and separate training of the acoustic and visual modalities in the quality of the generated synthetic speech. Finally, we cross-validate the results of the synthesis of the different emotions. This cross validation allowed us to analyze the ability of DNNs to learn characteristics specific to each emotion.

MOTS-CLÉS : Synthèse audiovisuelle expressive, tête parlante expressive, émotion, expression faciale, réseau de neurones profond récurrent à mémoire court-terme et long terme .

KEYWORDS: Expressive audiovisual speech synthesis, Expressive talking head, emotion, facial expression, deep bidirectional long short-term memory neural network (DBLSTM).

1 Introduction

De nos jours, l'animation automatique des têtes parlantes virtuelles expressives est en gain constant d'attention. Elle peut être utilisée dans plusieurs domaines, tel que le domaine des jeux vidéo, des films d'animation ainsi que dans le domaine médical et celui de l'éducation (Sproull *et al.*, 1996; Pandzic *et al.*, 1999; Ostermann & Millen, 2000). L'expressivité dans les systèmes de synthèse de la parole est très demandée puisqu'elle permet d'améliorer l'expérience des utilisateurs et de rendre l'interaction plus naturelle (Eyben *et al.*, 2012; Charfuelan & Steiner, 2013). La synthèse paramétrique statistique de la parole a connu des améliorations ces dernières années, notamment en terme d'intelligibilité (King, 2014), grâce aux techniques paramétriques statistiques allant des HMMs

Volume 1 : Journées d'Études sur la Parole, pages 127–135.
HAL : hal-02798526.

(Hidden Markov Models) aux réseaux de neurones (Ze *et al.*, 2013; Zen & Senior, 2014), notamment les réseaux BLSTM (Bidirectional Long Short-Term Memory) qui sont capables de prendre en compte les informations passées et futures d'une séquence. Il a été démontré que les BLSTMs donnent de meilleurs résultats de synthèse que les HMMs et les DNNs standards (Fan *et al.*, 2014, 2015; Klimkov *et al.*, 2018). Par ailleurs, le choix des paramètres et la configuration de ces réseaux sont cruciaux pour que la parole synthétique soit naturelle et intelligible. Il faut donc étudier et choisir minutieusement les différents paramètres et architectures impliqués dans l'entraînement des modèles de la parole.

Plusieurs travaux et systèmes se basent sur des descripteurs linguistiques pour des fins de synthèse vocale que ça soit avec des HMMs (Pouget, 2017; Baumann & Schlangen, 2012) ou par DNNs (Wu *et al.*, 2016; Houidhek *et al.*, 2018; Ribeiro *et al.*, 2016). Nous présentons dans cet article une étude bien nécessaire qui regroupe un ensemble d'expériences permettant d'apporter des réponses et des éclaircissements sur le comportement des DNNs face aux données linguistiques et audiovisuelles. Des études similaires ont été menées dans le passé sur des systèmes HMMs (Watts *et al.*, 2010; Le Maguer *et al.*, 2013; Cernak *et al.*, 2013) mais peu d'études se sont intéressées à l'impact des paramètres linguistiques sur un système basé sur les DNNs et encore moins leur impact sur la modélisation de la parole visuelle.

Le Maguer *et al.* (2013) ont étudié l'apport des différents paramètres linguistiques à la qualité de la synthèse du système HTS basé sur des HMMs. Cette étude menée sur un corpus acoustique de langue française a montré que l'utilisation du contexte phonétique améliore la modélisation du spectre de la parole et des durées, et que l'utilisation des informations sur les syllabes améliore la modélisation de la F0. Toutefois, le reste des facteurs contextuels ne semblent pas apporter une amélioration significative à la modélisation acoustique avec HTS. Cernak *et al.* (2013) ont également étudié les facteurs contextuels des données linguistiques pour la synthèse vocale par HMMs pour l'anglais. Cette étude confirme que le contexte syllabique fait partie des facteurs contextuels les plus importants et que le contexte relatif aux mots de la phrase a peu d'importance comme préalablement établie dans l'étude de Yu *et al.* (2010).

Pour la synthèse vocale par DNNs, Ribeiro *et al.* (2016) utilisent différents niveaux de contextes linguistiques pour entraîner un réseau de neurones à propagation vers l'avant (DNN-Feed-Forward ou DNN-FF) pour l'anglais. Les paramètres suprasegmentaux ont été traités par un DNN agissant au niveau des syllabes, et la sortie (sous forme de paramètres acoustiques) de ce dernier a été intégrée en tant qu'entrée supplémentaire à un DNN standard agissant au niveau des *frames*. Cette étude montre que l'ajout d'une représentation pré-entraînée des paramètres suprasegmentaux est bénéfique pour la modélisation acoustique. Par ailleurs, l'ajout des vecteurs de plongement (*embedding*) appris sur des mots ne montre aucune amélioration des performances du DNN. Récemment, Mametani *et al.* (2019) ont présenté une étude des paramètres contextuels appris automatiquement par un système de synthèse *End-to-End* pour l'anglais. Ce genre de système se base sur des DNNs et prend en entrée un texte brut (ou sa représentation phonétique) pour le convertir en paramètres vocaux. Les résultats expérimentaux montrent que le réseau arrive à tirer parti de l'information implicite contenue dans la représentation phonétique du texte comme la réduction des voyelles ou le stress sur les syllabes dans le mot.

Notre étude s'ajoute au travail d'exploration des paramètres d'entrée pour la modélisation de la parole dans la synthèse par DNNs pour la langue française. De plus, dans ce travail nous étudions différents aspects relatifs à la parole, partant des données linguistiques, passant par la modélisation des durées et des données acoustiques et visuelles jusqu'à la modélisation des émotions. Nous effectuons aussi une comparaison objective entre les performances d'un modèle audiovisuel entraîné sur les données

acoustiques et visuelles conjointement puis séparément. Dans le passé, Schabus *et al.* (2013) ont entraîné un système HMMs pour la modélisation audiovisuelle de la parole. Cette étude a montré que les modèles joints offrent une meilleure synchronisation entre les modalités acoustique et visuelle et que la qualité des paramètres acoustiques prédits ne subit pas de dégradation par rapport au modèle acoustique indépendant. Dans une étude similaire, effectuée sur des données audiovisuelles provenant d'une caméra, Filntisis *et al.* (2017) ont déclaré n'avoir trouvé aucune différence significative entre les résultats des deux modèles DNNs (joints et séparés) concernant le réalisme de la vidéo de synthèse. Toutefois, les résultats acoustiques du modèle séparé ont été significativement plus appréciés que ceux du modèle joint. Cette étude s'est basée sur des résultats de tests perceptifs, dans notre étude nous voulons quantifier avec un test objectif l'apport ou la dégradation de la qualité due à l'utilisation d'un modèle joint. De plus, les données visuelles que nous utilisons proviennent d'un système de capture de mouvements et contiennent les informations en 3D. Finalement, nous effectuons une validation croisée sur les résultats obtenus pour les différentes émotions, pour vérifier si les modèles des durées, acoustique et visuel, arrivent à se spécialiser dans la modélisation des différentes émotions.

2 Données utilisées

Dans ce travail nous utilisons le corpus de langue française présenté dans Dahmani *et al.* (2019). Ce corpus a été joué par une actrice semi-professionnelle et contient six émotions plus l'état neutre. Des marqueurs rétro-réflectifs ont été collés sur le visage de l'actrice pour suivre les mouvements de son visage. 2000 phrases ont été enregistrées pour l'état neutre (4h de parole). De ces 2000 phrases, un sous-ensemble de 500 phrases a été sélectionné pour chacune des six émotions basiques : joie, tristesse, colère, surprise, peur et dégoût (entre 55min et 1h 11min de parole pour chaque émotion). Le contenu linguistique est identique pour toutes les émotions et les phrases de ce corpus ont été considérées de telle sorte qu'elles offrent une couverture phonétique maximale. Le corpus neutre couvre 92% des diphones du français et le sous-corpus de 500 phrases en couvre 52%. Les données textuelles, acoustiques et visuelles ont été alignées automatiquement au niveau phonétique.

Nous utilisons un vecteur de 417 paramètres linguistiques composé de :

- 190 paramètres binaires relatifs à la nature du phonème courant et de ses contextes gauches et droits (5x38 paramètres : 36 phonèmes et 2 codes supplémentaires, un pour les pauses et le deuxième pour les silences de début et de fin),
- 195 paramètres binaires relatifs à la catégorie phonétique (voyelle, consonne, nasal, fricatif,...) du phonème courant et de ses contextes gauches et droits (5x39 paramètres),
- 2 paramètres numériques relatifs à la position du phonème courant dans la syllabe courante,
- 7 paramètres numériques relatifs à la syllabe courante précédente et suivante, le nombre de phonèmes qu'elles contiennent, la position de la syllabe courante dans la phrase et dans le mot courants,
- 18 paramètres binaires relatifs à la nature de la voyelle centrale dans la syllabe courante,
- 5 paramètres numériques relatifs aux nombres de syllabes dans le mot courant, précédent et suivant ainsi que la position du mot courant dans la phrase courante.

Ces paramètres linguistiques représentent le vecteur d'entrée pour l'entraînement des trois modèles principaux : des durées, acoustique et visuel.

La durée de chaque phonème est extraite sous forme du nombre de *frames* qu'il couvre en considérant un pas de 5ms entre deux *frames* consécutifs. Pour les paramètres acoustiques, nous avons utilisé le Vocodeur WORLD pour extraire 60 coefficients MFCC (Mel-Frequency Cepstral Coefficients), 5 paramètres BAP (Band-Aperiodicity), la fréquence fondamentale avec une échelle logarithmique (log F0) et leurs paramètres dynamiques (Δ et $\Delta\Delta$) ainsi qu'un paramètre binaire pour préciser la nature

voisée/non-voisée du son dans chaque frame. Ces paramètres ont été extraits des fichiers audio toutes les 5ms. Ils représentent la sortie du DNN qui sera entraîné pour générer des paramètres acoustiques à partir des paramètres linguistiques. Concernant l'aspect visuel, nous nous intéressons dans ce travail uniquement à l'animation de la partie inférieure du visage. Nous sélectionnons sur l'ensemble des données disponibles dans le corpus les 44 points 3D qui couvrent la région des articulateurs (lèvres, joues, mâchoire et menton), soit un vecteur de 132 valeurs. Ces données ont été également présentées avec un écart de 5ms entre deux *frames* consécutifs. Nous avons divisé le corpus en trois sous-ensembles : l'ensemble d'entraînement contenant 80% des données, l'ensemble de validation et celui de test avec 10% de données chacun.

3 Impact du contexte linguistique sur la qualité de la synthèse

Dans cette section, nous testons 4 types différents de paramètres d'entrée pour analyser leurs impacts sur l'apprentissage des modèles de durées, acoustique et visuel :

- 1_cont : Uniquement l'information sur le phonème central ;
- 3_cont : L'information sur le phonème central son contexte gauche et droit immédiats ;
- 5_cont : L'information sur le phonème central ses deux contextes gauches et ses deux droits ;
- 5_cont_p : Même informations que 5_cont en plus des informations sur la position du phonème courant dans la syllabe et la catégorie phonétique des cinq phonèmes du contexte ;
- 5_cont_p_s : Même informations que 5_cont_p en plus des informations sur les syllabes ;
- 5_cont_p_s_m : Même informations que 5_cont_p_s en plus des informations sur les mots ;

Dans cette section nous utilisons uniquement les données du corpus neutre et adoptons deux architectures : une avec des DNN-FF et une autre avec des BLSTMs. Pour ces deux architectures, un DNN à deux couches a été retenu et nous avons essayé plusieurs largeurs de DNNs pour les différents vecteurs d'entrée (256, 512, 1024 et 2048). Les trois modèles ont été entraînés séparément, et l'architecture qui donne le meilleur résultat sur l'ensemble de validation a été retenue pour chaque expérience. Les meilleurs modèles ont été sélectionnés avec la technique du *early stopping*. Les modèles ont été entraînés avec MSE comme fonction de perte. La fonction d'activation des couches cachées est TANH et une fonction d'activation linéaire pour la couche de sortie. Nous avons utilisé l'optimiseur Adam et aucun dropout, BatchNorm ou régularisation spécifique n'a été utilisée.

Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original. Nous affichons la moyenne et les intervalles de confiance pour chaque métrique.

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	8.672 (±0.018)	5.888 (±0.007)	5.532 (±0.006)	5.435 (±0.009)	5.259 (±0.008)	**5.256** (±0.007)
Corrélation	0.389 (±0.002)	0.781 (±0.0008)	0.811 (±0.0004)	0.822 (±0.0007)	0.827 (±0.0007)	**0.827** (±0.0007)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [512, 512]	5_cont_p_s [512, 512]	5_cont_p_s_m [512, 512]
RMSE (f/p)	7.237 (±0.011)	5.418 (±0.006)	5.413 (±0.006)	5.411 (±0.006)	5.301 (±0.007)	**5.247** (±0.006)
Corrélation	0.639 (±0.002)	0.821 (±0.0007)	0.824 (±0.00006)	0.826 (±0.0006)	0.827 (±0.0006)	**0.827** (±0.0006)

TABLE 1 – *Les résultats du RMSE en frames/phonème et de la corrélation de Pearson sur l'ensemble de test générés par le modèle de durées en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

Dans les tableaux 1, 2 et 3, il est très intéressant de constater qu'en utilisant une architecture DNN-FF, l'ajout de toutes les informations contextuelles améliore la qualité de la synthèse pour les trois aspects de la parole, bien que l'écart soit extrêmement serré entre les résultats avec et sans informations relatives aux mots. Toutefois, pour le réseau de type BLSTM les trois modèles n'ont pas tous le

	DNN-FF					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	6.653 (±0.026)	6.136 (±0.025)	6.132 (±0.024)	5.910 (±0.024)	5.901 (±0.024)	**5.900** (±0.024)
BAPD (dB)	0.327 (±0.004)	0.295 (±0.004)	0.292 (±0.004)	0.295 (±0.003)	0.288 (±0.003)	**0.287** (±0.003)
F0-RMSE (Hz)	35.226 (±1.105)	32.447 (±1.132)	31.334 (±1.106)	31.360 (±0.757)	30.648 (±0.733)	**30.555** (±0.743)
F0-Corrélation	0.341 (±0.021)	0.481 (±0.018)	0.529 (±0.017)	0.526 (±0.016)	0.557 (±0.016)	**0.563** (±0.015)
V/N-V (%)	14.250 (±0.424)	13.195 (±0.539)	13.090 (±0.553)	12.877 (±0.371)	12.872 (±0.375)	**12.848** (±0.371)
	BLSTM					
	1_cont [256, 256]	3_cont [512, 512]	5_cont [1024, 1024]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
MCD (dB)	5.304 (±0.029)	**5.099** (±0.023)	5.152 (±0.023)	5.103 (±0.025)	5.106 (±0.026)	5.146 (±0.024)
BAPD (dB)	0.282 (±0.004)	**0.242** (±0.003)	0.245 (±0.003)	0.242 (±0.003)	0.247 (±0.003)	0.247 (±0.003)
F0-RMSE (Hz)	32.580 (±0.818)	**27.934** (±0.622)	29.010 (±0.624)	28.460 (±0.690)	28.201 (±0.648)	28.207 (±0.865)
F0-Corrélation	0.471 (±0.016)	**0.640** (±0.013)	0.620 (±0.013)	0.628 (±0.014)	0.639 (±0.013)	0.637 (±0.014)
V/N-V (%)	10.736 (±0.369)	**8.348** (±0.262)	8.822 (±0.257)	8.571 (±0.293)	8.566 (±0.303)	8.755 (±0.292)

TABLE 2 – *Les résultats sur l'ensemble de test générés par le modèle acoustique en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

	DNN-FF					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.760 (±0.031)	1.458 (±0.029)	1.429 (±0.030)	1.427 (±0.030)	1.424 (±0.029)	**1.423** (±0.029)
Corrélation	0.574 (±0.007)	0.763 (±0.005)	0.778 (±0.006)	0.778 (±0.005)	0.779 (±0.005)	**0.779** (±0.005)
	BLSTM					
	1_cont [256, 256]	3_cont [256, 256]	5_cont [512, 512]	5_cont_p [1024, 1024]	5_cont_p_s [1024, 1024]	5_cont_p_s_m [1024, 1024]
RMSE (mm)	1.327 (±0.030)	**1.316** (±0.029)	1.328 (±0.031)	1.330 (±0.030)	1.331 (±0.031)	1.332 (±0.031)
Corrélation	0.823 (±0.005)	**0.828** (±0.005)	0.822 (±0.005)	0.822(±0.005)	0.821 (±0.005)	0.821 (±0.005)

TABLE 3 – *Les résultats sur l'ensemble de test générés par le modèle visuel en variant les paramètres linguistiques lors de l'entraînement avec une architecture de type DNN-FF et BLSTM.*

même comportement face aux informations linguistiques. Pour le modèle des durées, et de manière similaire au DNN-FF, l'ajout de l'ensemble des informations linguistiques améliore la prédiction des durées. Concernant les modèles acoustique et visuel, le réseau BLSTM atteint la meilleure qualité de synthèse avec les contextes gauche et droit immédiats uniquement. Cela peut s'expliquer par la capacité des BLSTMs à accéder automatiquement aux contextes passés et futurs de la frame courante, contrairement aux DNN-FF qui nécessitent que cette information soit explicitement donnée en entrée.

Nous remarquons qu'en utilisant le réseau BLSTM, pour le modèle acoustique, les informations sur la position et la catégorie des phonèmes ainsi que les informations sur les syllabes améliorent la modélisation de la F0, alors que l'ajout des informations relatives aux mots a un impact presque nul sur les mesures objectives. Ces constatations confirment les résultats des études précédentes (Le Maguer *et al.*, 2013; Cernak *et al.*, 2013; Yu *et al.*, 2010). Par ailleurs, force est de constater que, pour le modèle visuel, l'ajout des informations contextuelles autres que les contextes gauche et droit immédiats est néfaste pour l'apprentissage. Ce comportement peut être expliqué par la réduction du nombre d'exemples d'apprentissage avec l'augmentation du nombre de combinaisons possibles dans le vecteur d'entrée. En réalité, l'ajout de plus de contraintes contextuelles divise les données en classes de plus en plus petites, et réduit de ce fait le nombre d'exemples d'apprentissage de chaque classe. Pour le modèle des durées, ce comportement ne semble pas se produire, nous pensons que cela vient du fait que le réseau doit prédire un seul et unique paramètre, qui est une tâche plus simple et qui nécessite donc moins d'exemples d'apprentissage.

4 Entraînement joint et séparé des modèles acoustique et visuel

Dans cette section nous étudions l'apport éventuel d'un entraînement joint des modalités acoustique et visuelle sur la qualité de la synthèse audiovisuelle. Nous incluons les six catégories d'émotions dans le processus d'apprentissage et nous utilisons 3_cont comme informations linguistiques. Le

vecteur de sortie pour le modèle joint est le résultat de la concaténation des paramètres acoustiques et visuels.

Le calcul des différentes métriques a été effectué entre les paramètres prédits et ceux provenant du corpus original. Pour le modèle acoustique et visuel, les durées utilisées sont celles provenant du corpus original.

	Modèles Séparés							Modèles joints [2048, 2048]						
	Neu	Joi	Tri	Col	Sur	Peu	Dég	Neu	Joi	Tri	Col	Sur	Peu	Dég
	Acoustique [1024, 1024]							Acoustique						
MCD (dB)	**4.863**	**5.738**	**5.288**	**5.262**	**5.699**	**5.226**	**5.431**	5.305	6.135	5.740	5.691	6.157	5.669	5.844
BAPD (dB)	**0.224**	**0.312**	**0.269**	**0.268**	**0.287**	**0.231**	**0.256**	0.265	0.359	0.304	0.322	0.335	0.269	0.304
F0-RMSE (Hz)	**26.172**	**46.723**	**32.074**	**39.514**	**32.203**	**40.617**	**35.972**	32.203	47.617	37.972	45.094	45.676	46.201	44.003
F0-Corrélation	**0.687**	**0.631**	**0.518**	**0.524**	**0.702**	**0.627**	**0.535**	0.683	0.627	0.514	0.513	0.683	0.488	0.518
V/N-V (%)	**6.900**	**10.167**	**7.692**	**8.082**	**9.874**	**7.711**	**9.137**	7.851	11.879	8.955	9.587	11.864	8.814	10.560
	Visuel [1024, 1024]							Visuel						
RMSE (mm)	**1.304**	**1.572**	**1.317**	**1.466**	**1.482**	**1.424**	**2.124**	1.309	1.581	1.320	1.475	1.504	1.429	2.132
Corrélation	**0.833**	**0.777**	**0.792**	**0.810**	**0.807**	**0.826**	**0.696**	0.829	0.776	0.790	0.808	0.803	0.825	0.689

TABLE 4 – *Les résultats des paramètres acoustiques et visuels sur l'ensemble de test générés en entraînant le DNN avec les données acoustiques et visuelles séparément puis conjointement.*

Le tableau 4 montre les résultats obtenus avec les deux modèles. Nous remarquons que l'entraînement joint des deux modalités dégrade toutes les mesures objectives, que ça soit pour la modalité acoustique ou visuelle. En effectuant une écoute informelle nous avons constaté plus de distorsion et un son légèrement étouffé dans les résultats acoustiques du modèle joint, mais pour les résultats visuels, nous n'avons constaté aucune différence humainement perceptible. Ce résultat rejoint celui de Filntisis *et al.* (2017) qui a montré via des tests perceptifs que les résultats des modèles séparés sont considérés comme légèrement plus réalistes, mais qu'aucune différence d'ordre significatif n'a été trouvée entre les résultats audiovisuels des deux modèles. Cependant, les résultats acoustiques du modèle séparé ont été considérés comme significativement plus réalistes que ceux générés par le modèle joint.

5 Validation croisée des résultats de la synthèse expressive

Dans cette expérience nous utilisons des modèles acoustique et visuel séparés avec 3_cont comme vecteur d'entrée, et 5_cont_p_s_m comme entrée du modèle des durées. Sachant que dans une étude précédente (Dahmani *et al.*, 2019), utilisant le même corpus et une architecture neuronale similaire (BLSTM), il a été montré, via des tests perceptifs que les émotions synthétiques sont correctement reconnues (sauf la peur et la tristesse). Dans ce travail nous souhaitons vérifier, via une étude objective la capacité des modèles à apprendre des caractéristiques spécifiques à chaque émotion. Pour ce faire nous procédons à une validation croisée.

Dans cette expérience, nous évaluons la capacité de nos modèles à modéliser les durées et les modalités acoustique et visuelle, toutefois la prononciation des phrases peut changer d'une émotion à l'autre (plus ou moins de pauses, suppression/ajout de voyelles). Ce dernier point n'est pas étudié dans ce travail. Pour le modèle des durées, nous utilisons les informations linguistiques de l'ensemble de test d'une émotion cible pour générer les durées de toutes les autres émotions et nous avons calculé les mesures de toutes les autres émotions par rapport aux données originales de l'émotion cible. Pour les modèles acoustique et visuel, nous avons considéré les données linguistiques ainsi que les durées des données originales de l'ensemble de test de l'émotion cible. En utilisant ces informations, nous générons les paramètres acoustiques et visuels correspondants à chaque émotion et calculons les différentes mesures. Les résultats relatifs à chaque émotion traitée sont représentés dans les lignes des tableaux 5, 6 et 7. Les résultats affichés dans ces trois tableaux montrent que les trois modèles arrivent à se spécialiser dans la modélisation des différentes émotions. Pour le modèle des durées, le dégoût semble être très différent des autres émotions. Cela peut s'expliquer par les durées des

Durées								
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	RMSE (f/p)	**5.289**	6.110	6.001	5.917	5.652	6.378	13.280
	Corrélation	**0.831**	0.799	0.804	0.786	0.803	0.806	0.779
Joie	RMSE (f/p)	7.346	**7.136**	7.385	7.272	7.206	7.708	14.703
	Corrélation	0.752	**0.774**	0.756	0.760	0.769	0.751	0.720
Tristesse	RMSE (f/p)	6.886	6.881	**6.606**	7.118	7.176	6.926	13.856
	Corrélation	0.770	0.765	**0.777**	0.755	0.754	0.770	0.747
colère	RMSE (f/p)	6.879	7.130	7.222	**6.463**	7.597	6.578	15.195
	Corrélation	0.720	0.737	0.728	**0.758**	0.729	0.744	0.686
surprise	RMSE (f/p)	6.394	6.905	7.134	6.471	**6.006**	7.532	14.582
	Corrélation	0.756	0.763	0.741	0.753	**0.781**	0.749	0.708
Peur	RMSE (f/p)	7.573	7.468	7.287	7.760	7.789	**7.174**	13.578
	Corrélation	0.767	0.758	0.766	0.756	0.763	**0.781**	0.753
Dégoût	RMSE (f/p)	13.614	15.709	13.669	15.162	14.361	13.146	**9.311**
	Corrélation	0.728	0.716	0.723	0.693	0.712	0.721	**0.741**

TABLE 5 – *Les résultats du RMSE en frames/phonème et de la corrélation de Pearson pour la validation croisée sur les résultats de prédiction des durées des données expressives de l'ensemble de test.*

Visuel									
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût	Statique
Neutre	RMSE (mm)	**1.304**	2.392	1.635	2.464	1.945	2.245	2.377	2.170
	Corrélation	**0.833**	0.77	0.801	0.769	0.782	0.805	0.739	—
Joie	RMSE (mm)	2.500	**1.572**	2.125	2.703	2.605	2.814	2.732	3.217
	Corrélation	0.727	**0.777**	0.734	0.722	0.736	0.734	0.712	—
Tristesse	RMSE (mm)	1.655	2.092	**1.317**	2.378	2.241	2.221	2.325	2.364
	Corrélation	0.775	0.753	**0.792**	0.723	0.727	0.773	0.713	—
colère	RMSE (mm)	2.604	2.564	2.439	**1.466**	2.100	1.688	3.124	3.308
	Corrélation	0.732	0.735	0.714	**0.810**	0.783	0.774	0.716	—
surprise	RMSE (mm)	1.984	2.537	2.271	2.046	**1.482**	1.980	2.614	2.817
	Corrélation	0.750	0.744	0.723	0.785	**0.807**	0.771	0.727	—
Peur	RMSE (mm)	2.255	2.778	2.239	1.715	1.883	**1.424**	3.041	3.055
	Corrélation	0.794	0.772	0.791	0.795	0.790	**0.826**	0.748	—
Dégoût	RMSE (mm)	2.823	3.160	2.822	3.414	3.063	3.460	**2.124**	3.530
	Corrélation	0.651	0.647	0.641	0.644	0.651	0.649	**0.696**	—

TABLE 6 – *Les résultats de validation croisée sur les résultats de prédiction des trajectoires visuelles des données expressives de l'ensemble de test. Statique représente un visage à l'état neutre avec une bouche constamment fermée.*

émotions dans le corpus utilisé. En fait, cette émotion a été jouée avec un débit remarquablement lent. La durée du corpus du dégoût (1h 53min) représente environ le double des durées des autres émotions (entre 55min et 1h 11min). Les résultats visuels nous permettent de voir certaines ressemblances entre quelques émotions, notamment entre l'état neutre et la tristesse et entre la colère et la peur. En ce qui concerne le modèle acoustique, nous remarquons qu'il y a également une ressemblance entre l'état neutre et la tristesse puis entre la colère et le dégoût, de plus la joie et la surprise sont les émotions avec le plus grand écart de F0 par rapport au neutre et aux autres émotions.

6 Conclusion

Dans cet article, nous avons effectué une étude bien nécessaire sur la synthèse audiovisuelle expressive de la parole, afin de donner des éclaircissements sur l'apport de certains paramètres sur les résultats générés. Pour atteindre cet objectif, nous avons adopté différentes architectures neuronales pour entraîner trois modèles : le modèle des durées, le modèle acoustique et le modèle visuel. Nous avons réalisé une comparaison directe entre ces architectures en variant les paramètres linguistiques utilisés. Les résultats obtenus montrent que bien que toutes les informations linguistiques soient bénéfiques pour le modèle des durées, pour le modèle acoustique, uniquement les informations sur le contexte gauche et droit immédiats ainsi que le contexte syllabique améliore la prédiction. Toutefois pour le modèle visuel les informations autres que le contexte gauche et droit immédiats semblent être nuisibles pour l'apprentissage. Nous avons également comparé la qualité de la synthèse des modèles

		Acoustique						
		Neutre	Joie	Tristesse	colère	surprise	Peur	Dégoût
Neutre	MCD (dB)	**4.863**	6.409	5.390	5.784	6.548	5.327	5.539
	BAPD (dB)	**0.224**	0.304	0.243	0.268	0.263	0.229	0.232
	F0-RMSE (Hz)	**26.172**	97.521	33.810	42.063	108.220	35.640	38.839
	F0-Corrélation	**0.687**	0.610	0.598	0.546	0.404	0.558	0.604
	V/N-V (%)	**6.900**	7.565	7.594	7.512	7.612	7.154	7.486
Joie	MCD (dB)	7.010	**5.738**	6.696	6.417	6.132	7.227	7.045
	BAPD (dB)	0.367	**0.312**	0.347	0.330	0.334	0.377	0.371
	F0-RMSE (Hz)	103.444	**46.723**	85.568	97.438	58.942	113.167	109.806
	F0-Corrélation	0.586	**0.631**	0.552	0.547	0.455	0.526	0.540
	V/N-V (%)	11.015	**10.167**	10.792	10.751	10.547	10.812	10.908
Tristesse	MCD (dB)	5.688	6.442	**5.288**	5.825	6.642	5.660	5.817
	BAPD (dB)	0.271	0.301	**0.269**	0.284	0.284	0.271	0.271
	F0-RMSE (Hz)	33.943	82.658	**32.074**	39.353	96.357	47.107	44.815
	F0-Corrélation	0.503	0.476	**0.518**	0.496	0.284	0.509	0.514
	V/N-V (%)	8.107	8.246	**7.692**	8.167	8.258	7.984	8.023
colère	MCD (dB)	6.177	6.069	5.793	**5.262**	6.171	6.039	6.040
	BAPD (dB)	0.303	0.287	0.290	**0.268**	0.291	0.303	0.304
	F0-RMSE (Hz)	43.043	89.370	41.357	**39.514**	97.935	44.919	43.601
	F0-Corrélation	0.440	0.454	0.497	**0.524**	0.357	0.505	0.491
	V/N-V (%)	8.705	8.615	8.515	**8.082**	8.807	8.347	8.495
surprise	MCD (dB)	6.806	5.916	6.616	6.277	**5.699**	6.951	6.911
	BAPD (dB)	0.305	0.30	0.302	0.301	**0.287**	0.311	0.311
	F0-RMSE (Hz)	102.248	51.066	86.765	97.706	**32.203**	112.539	109.363
	F0-Corrélation	0.449	0.564	0.394	0.444	**0.702**	0.382	0.391
	V/N-V (%)	10.176	9.769	9.967	10.078	**9.874**	10.007	10.105
Peur	MCD (dB)	5.730	7.015	5.729	6.097	7.075	**5.252**	5.226
	BAPD (dB)	0.246	0.334	0.262	0.297	0.286	**0.234**	0.231
	F0-RMSE (Hz)	37.586	116.012	48.980	41.281	128.206	**32.505**	35.090
	F0-Corrélation	0.435	0.404	0.487	0.477	0.242	**0.494**	0.627
	V/N-V (%)	7.951	8.254	8.307	8.258	8.352	**7.649**	7.711
Dégoût	MCD (dB)	5.995	7.124	5.949	6.250	7.269	5.641	**5.431**
	BAPD (dB)	0.272	0.338	0.279	0.328	0.296	0.265	**0.256**
	F0-RMSE (Hz)	38.890	117.422	46.779	42.528	132.273	36.477	**35.972**
	F0-Corrélation	0.473	0.439	0.512	0.502	0.299	0.516	**0.535**
	V/N-V (%)	9.350	9.840	9.446	9.837	9.828	9.245	**9.137**

TABLE 7 – *Les résultats de validation croisée sur les résultats de prédiction des paramètres acoustiques des données expressives de l'ensemble de test.*

acoustique et visuel entraînés séparément puis conjointement. Nous avons trouvé que les modèles entraînés séparément atteignent une meilleure précision de reconstruction lors de la comparaison via des tests objectifs. Finalement, les résultats objectifs de la validation-croisée effectuée sur les différentes émotions, montrent que les trois modèles arrivent à se spécialiser aux différentes émotions. Ces résultats nous ont aussi permis de constater des similarités et des différences entre certaines émotions. Ces résultats viennent pour compléter une étude perceptive effectuée précédemment par Dahmani *et al.* (2019) sur le même corpus. Sachant que les résultats objectifs ne reflètent pas toujours la perception humaine, nous comptons compléter, dans nos prochains travaux, cette étude par des tests perceptifs.

Nous souhaitons que cet ensemble d'expériences apportera plus de clarté sur le comportement des DNNs face aux différentes données linguistiques, des durées et audiovisuelles, et que ça facilitera, pour les autres chercheurs, le choix de l'architecture neuronale la plus adaptée pour la synthèse audiovisuelle expressive de la parole.

Remerciements

Nous remercions la plateforme Grid'5000 de nous avoir fourni des ressources GPU pour entraîner nos modèles (Balouek *et al.*, 2012).

Références

BALOUEK D., AMARIE *et al.* (2012). Adding virtualization capabilities to the Grid'5000 testbed. In *International Conference on Cloud Computing and Services Science* : Springer.

BAUMANN T. & SCHLANGEN D. (2012). Inpro_iss : A component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*.

CERNAK M., MOTLICEK P. & GARNER P. N. (2013). On the (un) importance of the contextual factors in hmm-based speech synthesis and coding. In *ICASSP 2013* : IEEE.

CHARFUELAN M. & STEINER I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and emotionML. In *INTERSPEECH*.

DAHMANI S., COLOTTE V., GIRARD V. & OUNI S. (2019). Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. *Proc. Interspeech 2019*.

EYBEN F., BUCHHOLZ *et al.* (2012). Unsupervised clustering of emotion and voice styles for expressive TTS. In *ICASSP 2012* : IEEE.

FAN B. *et al.* (2015). Photo-real talking head with deep bidirectional lstm. In *ICASSP*.

FAN Y. *et al.* (2014). Tts synthesis with bidirectional lstm based recurrent neural networks. In *INTERSPEECH*.

FILNTISIS P. P., KATSAMANIS A., TSIAKOULIS P. & MARAGOS P. (2017). Video-realistic expressive audio-visual speech synthesis for the greek language. *Speech Communication*, **95**.

HOUIDHEK A., COLOTTE V., MNASRI Z. & JOUVET D. (2018). Dnn-based speech synthesis for arabic : modelling and evaluation. In *International Conference on Statistical Language and Speech*.

KING S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, **1**(1).

KLIMKOV V., MOINET A. *et al.* (2018). Parameter generation algorithms for text-to-speech synthesis with recurrent neural networks. In *SLT Workshop 2018* : IEEE.

LE MAGUER S. L., BARBOT N. & BOEFFARD O. (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *Eighth ISCA Workshop on Speech Synthesis*.

MAMETANI K., KATO T. & YAMAMOTO S. (2019). Investigating context features hidden in end-to-end tts. In *ICASSP 2019* : IEEE.

OSTERMANN J. & MILLEN D. (2000). Talking heads and synthetic speech : An architecture for supporting electronic commerce. In *2000 ICME2000* : IEEE.

PANDZIC I. S., OSTERMANN J. & MILLEN D. (1999). User evaluation : Synthetic talking faces for interactive services. *The visual computer*, **15**(7-8).

POUGET M. (2017). *Synthèse incrémentale de la parole à partir du texte*. Thèse de doctorat.

RIBEIRO M. S., WATTS O. & YAMAGISHI J. (2016). Syllable-level representations of suprasegmental features for dnn-based text-to-speech synthesis. In *INTERSPEECH*.

SCHABUS D., PUCHER M. & HOFER G. (2013). Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, **8**(2).

SPROULL L., *et al.* (1996). When the interface is a face. *Human-Computer Interaction*.

WATTS O. *et al.* (2010). The role of higher-level linguistic features in hmm-based speech synthesis.

WU Z. *et al.* (2016). Merlin : An open source neural network speech synthesis system. In *SSW*.

YU K. *et al.* (2010). Word-level emphasis modelling in hmm-based speech synthesis. In *ICASSP*.

ZE H. *et al.* (2013). Statistical parametric speech synthesis using deep neural networks. In *ICASSP*.

ZEN H. & SENIOR A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *ICASSP*.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Rythme et contrôle articulatoire : étude préliminaire du Human Beatbox.

Alexis Dehais Underdown[1] Paul Vignes[1], Lise Crevier-Buchman[1, 2] & Didier Demolin[1]
(1) Laboratoire de Phonétique et Phonologie, (CNRS / Sorbonne Nouvelle), 19 Rue des Bernardins, 75005 Paris, France
(2) Unité d'exploration fonctionnelle Voix-Parole – Service d'ORL &Chirurgie Cervico-Faciale, 40 rue Worth, BP 36, 92151 Suresnes, France
alexis.dehais-underdown@sorbonne-nouvelle.fr, vignes.paul@gmail.com,
lise.buchman1@gmail.com, didier.demolin@sorbonne-nouvelle.fr

RÉSUMÉ

Dans cette étude nous nous intéressons à l'analyse spectrale d'imitation de grosses caisses, de charleston et de caisse claire dans un paradigme d'augmentation de la vitesse de production. La vitesse de production a été contrôlée en utilisant un métronome à vibration paramétré à 90, 120 puis 150 battements par minute. Le Centre de Gravité spectral et le coefficient d'asymétrie ont été mesurés pour inférer la stabilité et la variabilité articulatoire des sons produits dans les patterns beatboxés. Les grosse caisses sont les sons les plus contrôlés suivi par les caisses claires puis les charlestons.

ABSTRACT

Rhythm and articulatory Control : Preliminary study of Human Beatboxing.
In this study we were interested in the spectral analysis of kick drums, hi-hat and snare drum imitations in a speed increasing production task. The production speed was controlled using a vibration metronome set at 90, 120 and 150 beat per minute. Center of Gravity and Skewness were measured in order to infer articulatory stability and variability of the sounds in the beatboxed patterns. Kick drum were the most controlled sounds followed by snare drums and hi-hats.

MOTS-CLÉS : HUMAN BEATBOX, VITESSE DE PRODUCTION, ACOUSTIQUE, RYTHME.
KEYWORDS: HUMAN BEATBOXING, SPEED RATE, ACOUSTICS, RHYTHM.

1 Introduction

La parole est un accommodement entre des principes de production et de perception pour délivrer un message linguistique. L'utilisation du Conduit Vocal (CV) humain est soumise à des contraintes qui permettent de minimiser le coût de la production afin de maximiser la perception des sons produits. La parole est dépendante d'une organisation complexe de différents niveaux linguistiques (lexical, morphosyntaxique, phonologique …). Le Human Beatbox (HBB) n'obéit pas à ses restrictions, on ne beatboxe pas pour véhiculer un message linguistique. Cela implique que la production n'est pas soumise à des contraintes liées aux langues. Cette différence entre parole et HBB a un impact fondamental sur la façon dont on utilise le conduit vocal lorsque l'on beatboxe. L'HBB n'est pas

Volume 1 : Journées d'Études sur la Parole, pages 136–144.
HAL : hal-02798527.

restreint à l'ensemble des gestes articulatoires répondant au compromis linguistique entre production et perception. Cependant, il doit aussi exister une relation entre production et perception dans le HBB dont on ignore, à ce jour, la nature.

Dans cette étude nous nous intéressons à l'analyse spectrale d'imitation d'instruments percussifs dans un paradigme d'augmentation de la vitesse de production. Il existe à ce jour peu d'études sur le HBB. Concernant le signal acoustique des sons beatboxés la littérature est encore plus restreinte. On retiendra celles de Picart, Brognaux Dupont, 2015 ; Stowell & Plumbley, 2010 s'intéressant à la classification automatique des sons, celle de Stowell & Plumbley (2008) décrivant de façon impressionniste les sons et celles de Dehais Underdown, Crevier Buchman & Demolin (2019) et Dehais Underdown et al. (2019) qui ont analysé les caractéristiques spectrales de la grosse caisse [p'] et la caisse claire [pf']. Enfin, à partir du même corpus Dehais Underdown et al. (2020) ont montré que la durée des séquences beatboxées tend à diminuer lorsque la vitesse de production augmente ($r = 0.97$) ; de plus ils ont trouvé un faible taux d'erreurs de production de 7,6% ($n=1083$).

Les sons de notre corpus sont non-voisés, il faut donc s'interroger sur l'analyse acoustique de la production d'une suite de sons sourds. En l'absence de sons vocaliques il est difficile d'utiliser certains indices acoustiques comme le VOT (Cho & Ladefoged, 1999), les équations de locus (Sussman, McCaffrey & Matthews, 1991) ainsi que d'autres corrélats destinés à l'étude des consonnes glottiques (Kingston, 1985 ; Wright et al., 2002). Nous avons donc choisi de porter notre analyse sur les bruits des sons que nous avons analysé en s'intéressant aux moments spectraux 1 et 3, c'est-à-dire le Centre de Gravité spectral (CdG) et le coefficient d'asymétrie (Forrest et al. 1988). Ces paramètres permettent de distinguer les lieux d'articulation à la fois des occlusives et des fricatives sibilantes (Forrest et al. 1988 ; Jongman et al. 2000 ; Nittrouer, 1995 ; Lousada et al. 2012). Il existe quatre moments spectraux décrivant la répartition d'énergie dans le spectre. Le premier correspond au CdG, c'est une moyenne de la composition spectrale indiquant la fréquence où l'énergie est concentrée. Le deuxième moment décrit la variance de la distribution. Le troisième moment (i.e. coefficient d'asymétrie ou « skewness » en anglais) renseigne sur la symétrie ou l'asymétrie de la distribution. Enfin le quatrième moment (i.e. coefficient d'aplatissement ou kurtosis) décrit l'aplatissement de la distribution. Les fortes corrélations entre le CdG et les paramètres articulatoires impliqués dans la production des fricatives dans l'étude de Tabain (2001) nous permettent d'inférer que le CdG est un bon indicateur pour définir le lieu d'articulation. Nous prendrons le coefficient d'asymétrie comme paramètre indiquant la variabilité de composition spectrale et indirectement les variations de configuration du CV.

L'objectif de cette étude est d'analyser l'influence de la vitesse de production sur des sons beatboxés. Nous faisons l'hypothèse qu'en augmentant la vitesse de production cela entrainera des perturbations dans la production : (1) nous pensons que les sons coronaux montreront plus de variation de CdG (i.e. indicateur du lieu d'articulation) que les vélaires et ensuite que les labiales ; (2) que les labiales montreront plus de variation du coefficient d'asymétrie (i.e. indicateur de la configuration du conduit vocal) ; enfin (3) la position du son dans une structure beatboxée est un facteur influençant la variation de ces deux paramètres spectraux.

2 Méthodologie

2.1 Corpus & protocole

Un beatboxeur professionnel de 31 ans a pris part à cette étude préliminaire. Les enregistrements se sont déroulés dans la chambre sourde du Laboratoire de Phonétique et Phonologie avec un microphone cardioïde AKG C520 avec une fréquence d'échantillonnage de 44kHz. Les séquences rythmiques (i.e. pattern de beatbox) à l'étude ont été extraites d'un corpus plus large. Le pattern de beatbox désigne une structure composée de sons beatboxés qui peut avoir un rythme binaire ou asymétrique et une organisation plus ou moins complexe en fonction du nombre de constituants de cette structure. La structure de pattern de beatbox créé pour cette étude est composée de 9 frappes (i.e. production du son) sur 4 battements. Chaque frappe a été numérotée de 1 à 9 sous forme de position dans le pattern de beatbox. Nous avons contrôlé la vitesse de production (i.e. Tempo) en utilisant un métronome à vibration paramétré à 90, 120 puis 150 battements par minute (BPM). Chaque pattern de beatbox (PB) se compose d'une même structure : Grosse Caisse, Charleston, Caisse Claire, Charleston, Grosse Caisse, Grosse Caisse, Charleston, Caisse Claire, Charleston. Nous avons choisi de transcrire les sons beatboxés à l'aide de l'API (Association Phonétique Internationale, 1999) et de ses extensions destinées à la parole pathologique (Ball, Howard & Miller, 2018).

1. [p' t͡s' ↓kɬ̠ t͡s' p'p' t͡s' ↓kɬ̠ t͡s'],
2. {ŋ [↑ʘ ↑ǀ ǂ* ↑ǀ ↑ʘ↑ʘ ↑ǀ ǂ* ↑ǀ] ŋ} ({ŋ … ŋ} indique que l'artiste peut fredonner en même temps qu'il beatboxe, [ǂ*] est un click où le relâchement est postérieur décrit par (Tuhuse & Traill, 1999)),
3. [t' t͡s' tʃ̥͡ʲ' t͡s' t't' t͡s' tʃ̥͡ʲ' t͡s'].

Chaque pattern a été répété 8 fois de suite à trois vitesses différentes (3 PB x 8 répétitions x 3 vitesses = 72 PB). Les fichiers sons ont été ré-échantillonnés à 32kHz dans le but d'observer les plus hautes fréquences.

2.2 Analyse

Nous avons utilisé Praat (Boersma & Weenink, 2006) pour segmenter les données et extraire le CdG et le coefficient d'asymétrie. Sur le textgrid nous avons noté le son tel qu'il a été produit et la cible qui était attendue. Pour chaque son, nous avons extrait les spectres FFT sur des fenêtres de 25ms afin d'observer les changements spectraux au cours du temps pour chaque son. À partir des spectres extraits, nous avons relevé les valeurs de Centre de Gravité spectral (CdG) et le coefficient d'asymétrie (i.e. Skewness) dans le but d'analyser la composition spectrale de chaque son en fonction de la position et de la vitesse de production. La durée (ms) des PB a été calculée à partir du début du premier son jusqu'à la fin du dernier son. Étant donné que la durée des PB diminue lorsque la vitesse de production augmente (Dehais Underdown, 2020) nous utiliserons cette valeur pour évaluer l'influence de la vitesse de production sur nos mesures spectrales. Dans notre analyse nous avons enlevé toutes les erreurs de production (e.g. substitutions, omissions) dans les 3 patterns à l'étude, c'est-à-dire lorsqu'un son produit ne correspond pas à la cible attendue, pour ne pas biaiser notre analyse. Nous n'analyserons pas les erreurs de production (4,3%).

Le logiciel R (Team R core, 2005) a été utilisé dans le but de générer des nuages de points et des coefficients de corrélations de Pearson des mesures spectrales (variable dépendante) et de la durée des PB (variable indépendante).

3 Résultats

3.1 Grosse caisse

Concernant le CdG des grosses caisses [p'] [↑ʘ], sur la figure 1 (gauche) les valeurs se situent en dessous de 500 Hz, pour [t'] les valeurs se trouvent en dessous de 2,5 kHz. Cela suggère qu'il y a une concentration de l'énergie qui prédomine dans les basses fréquences dans le but d'imiter au mieux le timbre grave d'une grosse caisse. Nous avons d'un côté les sons [p'] et [↑ʘ] qui, dans toutes les positions (i.e. position 1, 5 et 6), montrent peu de dispersion des valeurs de CdG alors que la vitesse augmente. Il n' y a pas de corrélation entre la vitesse et le CdG. De plus les positions 1, 5 et 6 ne semblent pas influer la production des grosses caisses labiales. La position 6 est celle qui montre le moins de variation et suggère donc que la production est stable. De l'autre côté, [t'] montre plus de dispersion des valeurs de CdG suggérant des variations subtiles de lieu d'articulation dans la zone dentale-alvéolaire.

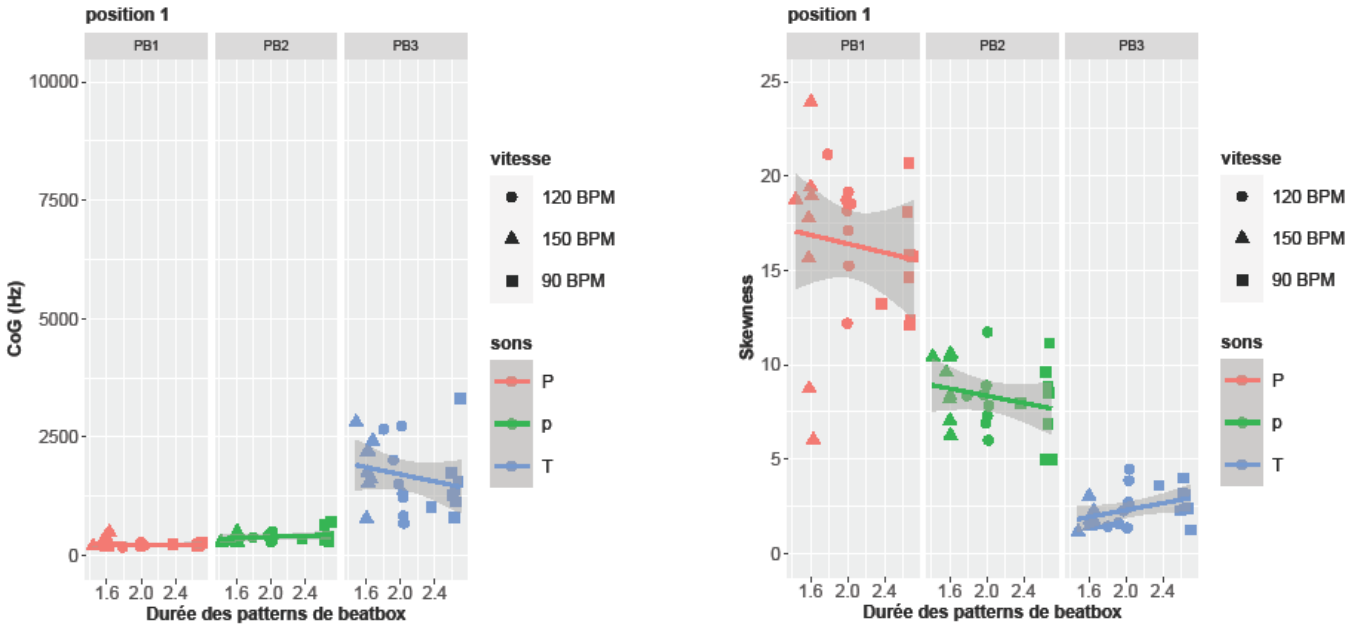


FIGURE 1 : Nuages de points du CdG (gauche) et coefficient d'asymétrie (droite) en fonction de la durée des PB de différentes imitations de grosses caisses en position 1. P = [p'], p = [↑ʘ], T = [t']. Les zones grisées indiquent l'intervalle de confiance.

L'analyse du coefficient d'asymétrie nuance quelque peu notre analyse. [p'] montre plus de dispersion des valeurs, ce qui suggère des variations de composition spectrale dues à des différences de configuration du conduit vocal lors de la production de cette labiale. Ces variations de composition ne sont pas surprenantes pour une labiale car les lèvres sont l'articulateur actif, laissant les articulateurs buccaux libres de leurs mouvements. Il est possible que ces variations soient dues à une anticipation des gestes suivants ou bien à la fin du geste précédent. La grosse caisse [↑ʘ] suit cette

tendance dans une moindre mesure car on observe moins de dispersion des valeurs du coefficient. Cela s'explique par le fait que le mécanisme d'initiation vélique de ce pattern contraint plus la configuration du conduit afin de maintenir l'occlusion postérieure du dos de la langue. Aucune corrélation entre la vitesse de production et le coefficient d'asymétrie n'a été trouvée pour les labiales. À l'inverse, [t'] montre moins de dispersion du coefficient sur les nuages de points générés. Cela suggère donc que la configuration du conduit vocal reste similaire d'une répétition à l'autre. Cette grosse caisse se trouve dans un PB uniquement composé de sons coronaux et implique donc seulement des changements de position et de la forme de l'apex et la lame de la langue dans la zone dentale-alvéolaire. Il est possible que ce pattern contraigne aussi les variations de configuration du conduit. La corrélation entre la vitesse et le coefficient d'asymétrie de la coronale est très faible ($r > 0.48$). Les positions 1, 5 et 6 ne semblent pas avoir d'influence sur la production des grosses caisses.

3.2 Charleston

À l'inverse des grosses caisses, les charlestons sont des instruments dont le timbre est aigu, les imitations de charleston produites par le participant ont des CdG allant jusqu'à 10.5 kHz et des coefficients d'asymétrie négatifs ou proches de 0 (i.e. répartition dans les hautes fréquences). Nous avons exclu de notre analyse tous les [t͡s'] du PB3 (i.e. pattern de coronales) en position 4 car ils étaient réalisés entre un [t͡s'] et un [t͡ʃi̯'].

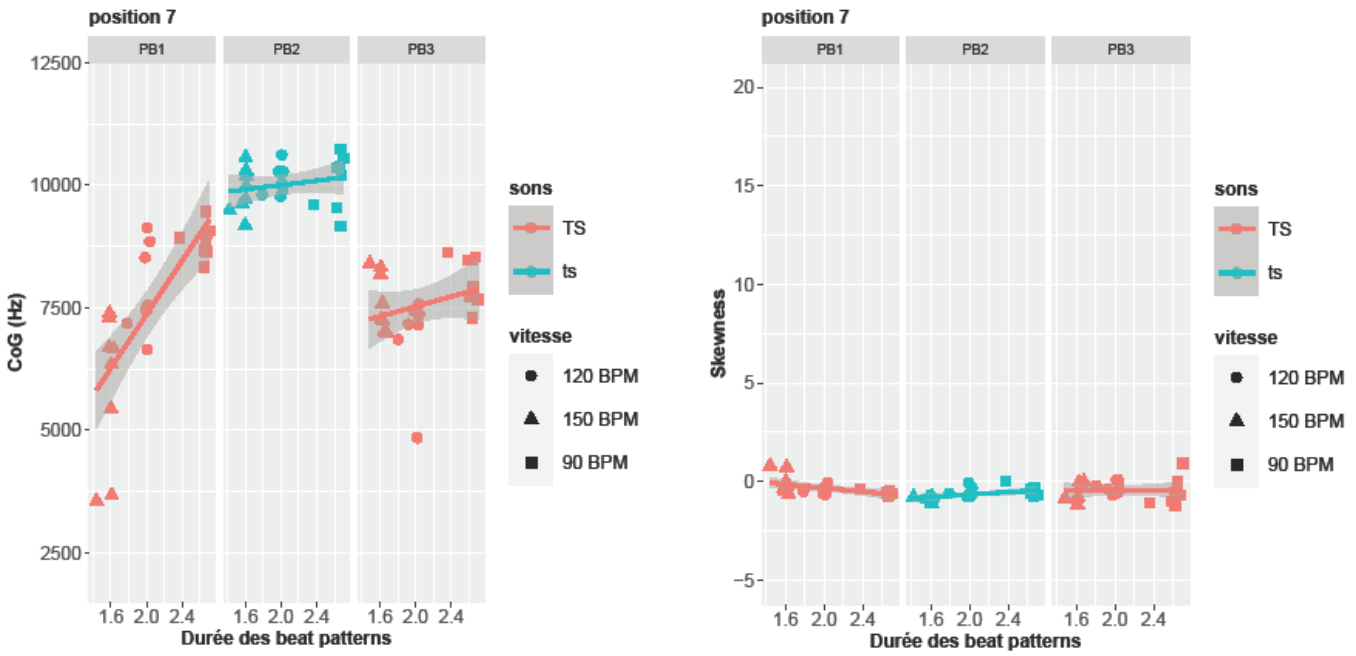


FIGURE 2 : Nuages de points du CdG (gauche) et coefficient d'asymétrie (droite) en fonction de la durée des PB de différentes imitations de Charleston en position 7. TS = [t͡s'], ts = [↑|]. Les zones grisées indiquent l'intervalle de confiance.

L'analyse des charlestons montre que ces sons sont les plus variables. De plus, selon la position ou le PB où ils se trouvent, l'impact de la perturbation induite par la vitesse est différent. Les charlestons [t͡s'] du PB1 sont les plus affectés (figure 2). En effet, on observe une importante dispersion des valeurs de CdG et une forte corrélation entre le CdG et la vitesse pour les positions 2 ($r=0.70$), 4 ($r=0.87$), 7 et 9 ($r=0.76$). Ces résultats reflètent des variations articulatoires dues à un possible recul du lieu d'articulation. Cela supposerait donc une possible interaction avec les sons adjacents. En revanche, le même son dans le 3ème pattern se comporte légèrement différemment. On observe moins

de dispersion des valeurs de CdG ce qui souligne plus de stabilité articulatoire. Nous avons trouvé une corrélation du même ordre de grandeur en position 2 (r=0.62), plus faible en position 7 (r=0.26) et plus forte en position 9 (r=0.83). Ces résultats reflètent aussi des variations articulatoires quant au lieu d'articulation. Enfin, [↑ǀ] montre moins de dispersion de CdG suggérant moins de variation de lieu d'articulation. La position la plus corrélée à l'augmentation de vitesse est la position 4 (r=0.66) et plus faiblement la position 9 (r=0.46). Ces résultats supposent encore une fois que la nature des contraintes articulatoires propres à chaque PB influence la production des sons.

Concernant le coefficient d'asymétrie on observe peu de dispersion des valeurs du Skewness quelle que soit le type de charleston. Cela indique moins de variation dans la configuration du CV. On observe cependant des corrélations entre la vitesse et le skewness. Le skewness du PB1 est négativement corrélé à la vitesse, particulièrement pour la position 4 (r=-0.77) et la position 9 (r = -0.62). Le spectre se compose donc de fréquences plus hautes à vitesse rapide. De même, dans le PB3 il est négativement corrélé ; encore une fois cette corrélation est plus forte pour les position 4 (r=-0.76) et 9 (r=-0.79). À l'inverse, on observe le phénomène contraire pour le PB2 avec une corrélation positive.

3.3 Caisse Claire

Les caisses claires de notre corpus sont des sons longs. Il y a trois fenêtres d'analyse pour ces sons (Figure 3, chaque colonne est une fenêtre de 25 ms). Pour les affriquées, la première fenêtre correspond à la partie occlusive, inversement les fenêtres 2 et 3 correspondent à la partie fricative ; pour [ǂ*] les fenêtres 2 et 3 correspondent à du bruit. Les analyses de CdG montrent des valeurs comprises entre 1 et 5 kHz pour [↓k͡ǁ], entre 3 et 4 kHz pour [ǂ*] et entre 4 et 8 kHz pour [t͡ʃi̥']. Quant aux valeurs du coefficient d'asymétrie elles sont comprises entre 1 et 6 (prédominance de moyennes et hautes fréquences) pour [↓k͡ǁ] et entre -0,5 et 1,5 (prédominance de hautes fréquences) pour [t͡ʃi̥'] et [ǂ*].La caisse claire ingressive [↓k͡ǁ] se compose d'une occlusive vélaire ingressive sourde et d'une fricative latérale ingressive sourde. La position dans le pattern n'influence pas le CdG de ces deux composantes. Le peu de dispersion des valeurs sur le nuage de points (Figure 3) et l'absence de relation forte entre la vitesse de production et le CdG (r=-0.1) montre que phase occlusive est stable. On observe un peu plus de variabilité en ce qui concerne la fricative latérale [↓ǁ] (figure 3) en terme de dispersion des valeurs des CdG. Il semble que la phase fricative soit tout aussi stable. Le coefficient d'asymétrie montre peu de dispersion ce qui nous laisse penser que la configuration du conduit vocal varie peu. Aucune corrélation n'a été faite entre les paramètres spectraux et la vitesse.

La caisse claire est en fait un clic [ǂ*] dont le relâchement se fait au niveau de la constriction postérieure (cf. Tuhuse & Traill, 1999 pour une description). Acoustiquement il se compose d'un relâchement avec la présence d'un burst acoustique et suivi de bruit. Sur la (Figure 3) on observe que la caisse claire vélique montre plus de dispersion dans les valeurs de CdG lors de la phase de relâchement (colonne 1) de la constriction vélaire que lors du bruit suivant le relâchement (colonne 2 et 3). Il n'y a pas de dispersion du coefficient d'asymétrie. Le relâchement sur la figure 3 est négativement corrélé à la vitesse concernant le CdG et positivement à la vitesse concernant le Skewness. Cependant cette corrélation vaut seulement pour ce son et la position 8 ; cela se traduit par une augmentation du CdG et une diminution du skewness à vitesse rapide.

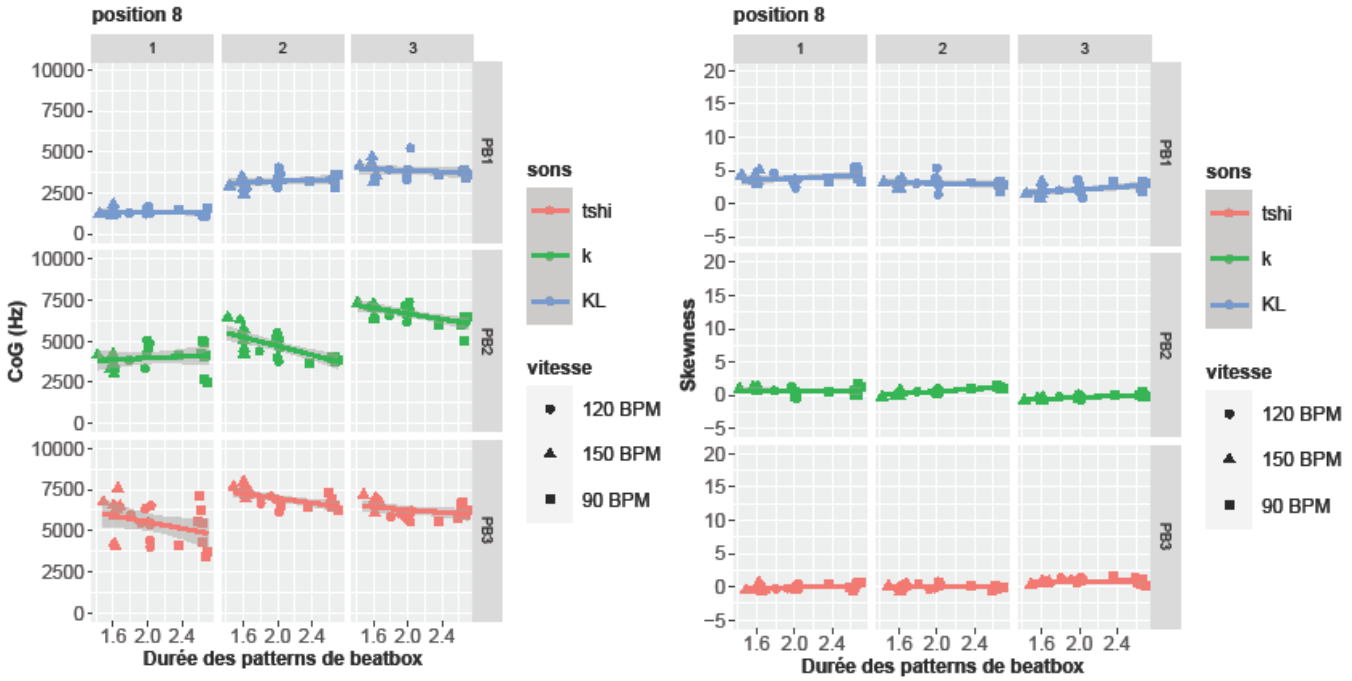


FIGURE 3 : Nuages de points du CdG (gauche) et coefficient d'asymétrie (droite) en fonction de la durée des PB de différentes imitations de caisse claire en position 8. Chaque colonne correspond à une fenêtre de 25ms. tshi = [t͡ʃʲ], k = [ǂ*], KL = [ǀ͡ɬ]. Les zones grisées indiquent l'intervalle de confiance.

Enfin la caisse claire coronale [t͡ʃʲ] est plus variable concernant le relâchement de la phase occlusive (Figure 3, fenêtre 1). Il y a une variabilité de positionnement articulatoire. Le passage à la fricative post-alvéolaire est intéressant à analyser. Le CdG sur la 2ème fenêtre d'analyse montre moins de dispersion du CdG que la partie occlusive. L'absence de relation entre la vitesse et le CdG suggère que la production de la phase fricative est plus stable que la phase occlusive. Conjointement, le coefficient d'asymétrie montre une stabilité de la configuration articulatoire. Encore une fois on note l'absence de relation entre la vitesse et le coefficient d'asymétrie. Cela nous amène à penser que l'occlusive pourrait faire office de transitoire impulsionnel par analogie à la baguette de batterie qui bat la peau de l'instrument et que la friction ferait office de transitoire d'extinction qui porte les indices du timbre acoustique. Il se peut que pour des raisons perceptives le contrôle de la production de la phase fricative soit accru.

4 Discussion

Nous avions émis l'hypothèse que la vitesse entrainerait des perturbations du CdG, du coefficient d'asymétrie en fonction des positions. Les grosses caisses labiales, quelle que soit leur position dans le pattern, montrent une grande stabilité concernant le geste d'occlusion labiale, reflété par l'analyse du CdG. En revanche les variations de skewness laissent penser que la configuration du conduit vocal diffère d'une répétition à l'autre et d'une position à l'autre. Ces variations peuvent être dues à l'anticipation du geste suivant ou bien de la fin du geste précédent. À l'inverse la grosse caisse linguale montre la tendance inverse : des variations du lieu d'articulation et une stabilité de la configuration du CV. Le CdG des charlestons est plus impacté par le changement de vitesse comme le montre les résultats. Cela suppose que le geste lingual n'atteint pas le même lieu dans la zone dentale-alvéolaire. En revanche le peu de variation de skewness indique que la configuration du CV reste similaire d'une répétition à l'autre et d'une position à l'autre. Nous pensions initialement que

les sons vélaires montreraient plus de variations articulatoires que les labiales, or les données ne vont pas dans ce sens. Le peu de variabilité du CdG de [↓k͡ɬ] et [ǂ*] suggèrent un bon contrôle de production des labiales et des vélaires, et à moindre mesure de la caisse claire linguale [tʃ͡ʃ̟ʲ]. Concernant cette dernière, notre hypothèse prédisait qu'en tant que coronale, elle serait plus sujette à la variation articulatoire, mais les données nous racontent une autre histoire. Bien que la phase occlusive soit quelque peu plus variable, l'ensemble de l'affriquée est stable ce qui suggère un bon contrôle de cette caisse claire.

La position dans le pattern pourrait bien être un facteur influençant la production. Les positions les plus variables sont celles occupées par les charlestons, plus particulièrement les positions 4 et 9. C'est en effet sur ces positions que nous avons trouvé le plus de variation en terme de stabilité articulatoire. La vitesse agit comme une perturbation sur la production et le beatboxeur doit donc faire preuve d'adaptabilité pour respecter les contraintes rythmiques des structures à produire. Dehais Underdown et al. (2020) ont montré que pour s'adapter à la perturbation induite par la vitesse, le beatboxeur réduisait la durée des pauses entre chaque son ainsi que la durée des sons en eux-mêmes, plus particulièrement celle des caisses claires. Cela implique donc que le timing entre les gestes se réduit lorsqu'on augmente la vitesse. Une possible stratégie d'adaptation, ici, serait de réduire la précision articulatoire des charlestons afin de conserver une fluidité dans sa production et garder un meilleur contrôle articulatoire. Pour renforcer cette interprétation il serait bien d'analyser les erreurs de production ; cela permettrait d'alimenter la question du chevauchement des gestes articulatoires chez les beatboxeurs et du contrôle articulatoire. Il est évidement nécessaire d'obtenir des données articulatoires pour confirmer ou réfuter cela.

5 Conclusion

L'étude du rythme et de son interaction avec des différences de vitesse de production est une question très intéressante en ce qu'elle permet d'étudier le contrôle articulatoire chez des experts de l'articulation. Dans de futures études (acoustiques et articulatoires) il faudra inclure plus de participants pour observer les tendances générales et les stratégies individuelles pour renforcer notre analyse sur les questions de contrôle et de précision articulatoire.

Remerciements

Ce travail est soutenu par le Labex EFL (ANR-10-LABX-0083)

Références

BALL, M. J., HOWARD, S. J., & MILLER, K. (2018). Revisions to the extIPA chart. *Journal of the International Phonetic Association, 48(2)*, 155-164. doi: 10.1017/S0025100317000147
BOERSMA, P. & WEENINK, D. (2006). Praat: doing phonetics by computer. Version 6.0.21, recupéré le 25 Septembre 2016 sur http://www.praat.org/
CHO, T. & LADEFOGED, P. . (1999). Variation and universals in VOT: evidence from 18 languages . *Journal of phonetics 27*, 207-229.
DEHAIS UNDERDOWN, A., CREVIER BUCHMAN, L. & DEMOLIN, D. (2019). Acoustico-Physiological coordination in the Human Beatbox: A pilot study on the beatboxed Classic Kick Drum. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp.

142-146). Canberra, Australia: Australasian Speech Science and Technology Association Inc.: Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.).
DEHAIS UNDERDOWN, A., VIGNES, P., CREVIER-BUCHMAN, L., & DEMOLIN, D. . (2019). Human beatboxing: A multi-instrumental pilot. *The Journal of the Acoustical Society of America, 146(4)*, (pp. 3082-3082).
DEHAIS UNDERDOWN, A., VIGNES, P., CREVIER BUCHMAN, L. & DEMOLIN, D. (2020). Articulatory control and beatboxing rate: A preliminary study. *12th International Seminar on Speech Production.* Providence.
FORREST, K., WEISMER, G., MILENKOVIC, P., & DOUGALL, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: preliminary data. *The Journal of the Acoustical Society of America, 84(1)*, 115-123.
INTERNATIONAL PHONETIC ASSOCIATION. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press.
JONGMAN, A., WAYLAND, R., & WONG, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America, 108(3)*, 1252-1263.
KINGSTON, J. (1985). *The Phonetics and Phonology of the Timing of Oral and Glottal Events (PhD).* Berkeley University.
LOUSADA, M. L., JESUS, L. M., & PAPE, D. (2012). Estimation of stops' spectral place cues using multitaper techniques. *Documentação de Estudos em Lingüística Teórica e Aplicada, 28(1)*, 1-26.
NITTROUER, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *The Journal of the Acoustical Society of America, 97(1)*, 520-530.
PICART, B., BROGNAUX, S., & DUPONT, S. (2015). Analysis and automatic recognition of human beatbox sounds: A comparative study. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4255-4259). IEEE. doi: 10.1109/ICASSP.2015.7178773
STOWELL, D., & PLUMBLEY, M. D. (2008). *Characteristics of the beatboxing vocal style.* Technical Report, Centre for Digital Music C4DMTR-08-01, Queen Mary, University of London, Dept. of Electronic Engineering. Récupéré sur http://c4dm.eecs.qmul.ac.uk/papers/2008/Stowell08-beatboxvocalstyle-C4DM-TR-08-01.pdf
STOWELL, D., & PLUMBLEY, M. D. (2010). Delayed decision-making in real-time beatbox percussion classification. *Journal of New Music Research, 39(3)*, 203-213. doi: 10.1080/09298215.2010.512979
SUSSMAN, H. M., MCCAFFREY, H. A., & MATTHEWS, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America, 90(3)*, 1309-1325. doi: 10.1121/1.401923
TABAIN, M. (2001). Variability in fricative production and spectra: Implications for the hyper-and hypo-and quantal theories of speech production. *Language and speech, 44(1)*, 57-93.
TEAM, R. D. (2005). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Récupéré sur http://www.R-project.org
TUHUSE, B. ǁX. & TRAILL, A. (1999). |Hán-|Hánsè, the desert Cisticola, implements an acoustic target , . *In ICPhS-14*, (pp. 1041-1042). San Francisco.
WRIGHT, R., HARGUS, S., & DAVIS, K. (2002). On the categorization of ejectives: data from Witsuwit'en. *Journal of the International Phonetic Association, 32(1)*, 43-77.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Unités prosodiques et grammaire intonative du français : vers une nouvelle approche

Elisabeth Delais-Roussarie[1] Brechtje Post[2] Hiyon Yoo [3]
(1) Université de Nantes, UMR 6310-LLING, Nantes, France
(2) University of Cambridge, Phonetics Laboratory, Cambridge, Royaume Uni
(3) Université Paris-Diderot, UMR 7110-LLF, Paris, France
Elisabeth.delais-roussarie@univ-nantes.fr, bmbp2@cam.ac.uk,
yoo@linguist.univ-paris-diderot.fr

RÉSUMÉ

Dans les travaux sur la prosodie du français sont généralement proposés deux ou trois niveaux de structuration prosodique: le syntagme accentuel, le syntagme intermédiaire et le syntagme intonatif. Alors que les auteurs sont souvent d'accord sur les modalités de construction du syntagme accentuel, il n'en est pas de même pour les deux autres niveaux.
Dans cet article, nous proposons de redéfinir le syntagme intermédiaire. Cette proposition diffère des autres travaux en deux points. Premièrement, l'extension et le statut du syntagme intermédiaire est clarifié pour en faire une unité métrique. Deuxièmement, une distinction est faite entre cette unité et deux types de syntagme intonatif.
Cette proposition se base sur l'inventaire des contours observés à la frontière droite de ces unités et sur l'étude des relations qu'elles entretiennent avec les structures morpho-syntaxique et sémantique. Elle vise à rendre compte du phrasé et du choix des contours intonatifs à un niveau phonologique sous-jacent.

ABSTRACT

Prosodic Units and Intonational Grammar in French: towards a new Approach

In most studies on French prosody, two or three distinct levels of constituency are assumed: the accentual phrase, the intermediate phrase and the intonational phrase. While there is considerable agreement on the definition of the accentual phrase, there is much controversy over the two other levels.
In this paper, a new definition of the intermediate phrase is argued for. Our proposal departs from previous work in two ways. First, the extension and status of the intermediate phrase is clarified in order to consider it as a metrically-driven unit. Second, a distinction is made between this phrase and two types of intonational phrases.
This proposal is based on (a) the inventory of the contours at the right edge of these phrases, and (b) their relation with the morpho-syntactic and semantic structures. Note that our proposal accounts for phrasing and intonation contour choice at the underlying phonological level.

MOTS-CLÉS : Structure prosodique, intonation, syntagme intermédiaire, syntagme intonatif.
KEYWORDS: Prosodic structure, intonation, intermediate phrase, intonational phrase.

Volume 1 : Journées d'Études sur la Parole, pages 145–153.
HAL : hal-02798529.

1 Introduction

Dans la plupart des études consacrées au phrasé prosodique et à l'intonation, un énoncé est considéré comme segmenté en constituants prosodiques organisés hiérarchiquement (cf., entre autres, Ladd, 2008 ; Nespor & Vogel, 1986 ; Selkirk, 1986 & 2011). Pour le français, trois niveaux de structuration sont généralement retenus dans les modèles prosodiques développés dans le cadre métrique-autosegmental (Post, 2000 ; Jun & Fougeron, 2000 ; Delais-Roussarie et al., 2016 ; Michelas, 2011) :

- le syntagme accentuel ou AP, parfois appelé mot prosodique (Delais-Roussarie, 1996) ou syntagme phonologique (Post, 2000) ;
- le syntagme intermédiaire ou *ip* (Jun & Fougeron, 2000 ; Michelas, 2011)
- le syntagme intonatif ou IP (dans presque tous les modèles)

Bien qu'un même énoncé puisse être segmenté de plusieurs manières en APs, comme indiqué sous (1), il existe un consensus assez large sur la définition et la caractérisation de cette unité prosodique.

(1) *Ces jeunes enfants apprennent à parler le français.*

a. (ces jeunes enfants)$_{AP}$ (apprennent à parler)$_{AP}$ (le français)$_{AP}$
b. (ces jeunes enfants)$_{AP}$ (apprennent)$_{AP}$ (à parler)$_{AP}$ (le français)$_{AP}$

Sur le plan métrique, un AP correspond à une séquence de syllabes s'achevant par une syllabe accentuée (ou portant un accent primaire). Du fait d'un important syncrétisme entre accentuation et intonation, un contour intonatif montant est associé à la syllabe accentuée, et un schéma tonal comparable à celui sous (2) est associé à tout syntagme accentuel (Delais-Roussarie et al., 2016).

(2) aL (Hi) L H* (aL signifiant ton bas de frontière gauche)

Pour ce qui est de l'interface avec la structure morpho-syntaxique, l'AP est souvent décrit comme contenant un mot de contenu (verbe, nom, adverbe) précédé des mots grammaticaux qui en dépendent (préposition, déterminant, auxiliaire, etc.) ; ce syntagme peut donc être dérivé de la structure morpho-syntaxique à l'aide de la contrainte d'alignement *Align R Xhead* qui stipule qu'une frontière d'AP coïncide avec la fin des têtes lexicales de constituants syntaxiques (cf. Selkirk, 1986 ; Delais-Roussarie, 1996, ; Jun & Fougeron, 2000). En revanche, il n'existe pas de réel consensus sur la façon de définir le syntagme intermédiaire ou, dans une moindre mesure, le syntagme intonatif. De plus, l'extension et les frontières de ces deux constituants peuvent se recouper.

Dans cet article, nous nous centrons sur le syntagme intermédiaire (*ip*) et sur les patrons intonatifs observés à sa frontière droite. Nous proposons une nouvelle façon de définir ce constituant et d'appréhender les différences entre syntagmes intermédiaire et intonatif. Cela apporte un éclairage nouveau sur la structure prosodique du français et permet de séparer les structures métriques et intonatives. L'article sera organisé comme suit. Dans la section 2, une synthèse des travaux existants est proposée afin de montrer où apparaissent les frontières de syntagmes intermédiaires. Cela conduira à distinguer trois types de syntagmes intermédiaires. Dans la section 3, après avoir établi un inventaire des contours mélodiques du français, nous décrivons les patrons mélodiques et les phénomènes intonatifs observables à la frontière droite des trois types d'*ip*s. Sur cette base, nous proposons une nouvelle façon d'appréhender les constituants prosodiques. Cette proposition est faite en s'appuyant sur l'analyse d'énoncés extraits de plusieurs corpus oraux du français ou de données expérimentales.

2 Le syntagme intermédiaire, un niveau problématique

Comme nous venons de le mentionner, alors que les définitions et caractéristiques du syntagme accentuel sont peu sujettes à discussion, celles des syntagmes intermédiaires, et dans une moindre mesure intonatifs, le sont davantage, certains auteurs ne reconnaissent pas l'*ip* comme niveau de structuration prosodique (Delais-Roussarie, 1996 ; Post, 2000, par exemple).

Un examen de la littérature a permis de dégager un ensemble d'énoncés faisant référence au syntagme intermédiaire comme niveau de structuration supplémentaire (Delais-Roussarie et al., 2016 ; Michelas, 2011). L'analyse des exemples cités dans ces articles permet de distinguer trois types de syntagmes intermédiaires en fonction de leur extension.

Avant de décrire chacun de ces types, il est important de garder à l'esprit qu'en plus des indices intonatifs (resetting partiel, cf. Michelas, 2011), trois types d'informations distinctes influencent généralement la distribution et la force relative des frontières prosodiques:

- la structure morpho-syntaxique, les constituants prosodiques étant parfois définis relativement à certaines unités syntaxiques (têtes de syntagmes, projections maximales, etc.). Dans certaines études, les relations entre structure syntaxique et structure prosodique sont exprimées en termes de contraintes d'alignement (cf. Selkirk, 2011 ; Delais-Roussarie, 1996 pour le français) ;
- la structure informationnelle, des frontières prosodiques particulières pouvant être réalisées en fonction du focus et du topique d'un énoncé. Voir, par exemple, les contraintes qui militent pour qu'une frontière prosodique apparaisse à droite du topique (Feldhausen, 2010) ou bien du focus informationnel (Selkirk, 2000 ; Jun & Lee, 1998 ; Fery, 2001) ;
- la structure métrique, la taille ou la structure métrique des syntagmes pouvant influencer la distribution des frontières prosodiques et le placement des accents (Post, 2000 ; Delais-Roussarie, 1996). En français, par exemple, la taille du syntagme accentuel est généralement limitée à six ou sept syllabes (Delais-Roussarie, 1996 ; Martin, 1987).

2.1 Le syntagme intermédiaire guidé par la métrique

Ce type de syntagme intermédiaire (*ip*) correspond à un SN sujet long comme dans les exemples (3) et (4) extraits respectivement de (Delais-Roussarie et al, 2016) et (Michelas, 2011).

(3) Le directeur de l'école ne voulait pas voir le guide des touristes qui attendait à la réception.
[{[le directeur)$_{AP}$ (de l'école)$_{AP}$}$_{ip}$ {(ne voulait pas voir)$_{AP}$ (le guide des touristes)$_{AP}$}$_{ip}$ {(qui attendait)$_{AP}$ (à la réception)$_{AP}$}$_{ip}$]$_{IP}$

(4) La mamie des amis de Rémi demandait l'institutrice.
[{(La mamie)$_{AP}$ (des amis)$_{AP}$ (de Rémi)$_{AP}$}$_{ip}$ {(demandait)$_{AP}$ (l'institutrice)$_{AP}$}$_{ip}$]$_{IP}$

Des frontières d'*ip* sont également réalisées à la fin d'un SN objet dans le cas des constructions avec deux compléments d'objet ou avec un complément d'objet complexe, comme sous (5).

(5) Il réglait le déchargement des casiers sur les chariots des mareyeurs.
[{(Il réglait)$_{AP}$ (le déchargement)$_{AP}$ (des casiers)$_{AP}$}$_{ip}$ {(sur les chariots)$_{AP}$ (des mareyeurs)$_{AP}$}$_{ip}$]$_{IP}$

Il est intéressant de noter que ces syntagmes intermédiaires contiennent deux ou trois syntagmes accentuels, mais jamais davantage. De plus, la frontière droite de ces *ip*s correspond généralement à

une frontière droite de projections maximales au niveau syntaxique (fin d'un syntagme nominal sujet branchant, etc.). Ce constituant est donc comparable au syntagme phonologique majeur défini à l'aide de la contrainte d'alignement *Align R* X_{Max} (Selkirk, 2000 et 2011 ; Delais-Roussarie, 1996).

2.2 Le syntagme intermédiaire incident

Dans bon nombre d'études (cf., Mertens 2008 par exemple), une frontière d'*ip* est réalisée à la fin de constituants syntaxiques remplissant une fonction discursive ou informationnelle particulière (topiques, cadratifs, parenthétiques, etc.). Ces *ip*s apparaissent dans des constructions particulières comme celles mentionnées sous (6).

(6) Constructions appelant la réalisation d'une frontière d'*ip* à leur droite

a. Dislocation
<u>A Paul</u>, je lui ai donné un livre.
[{(à Paul)$_{AP}$}$_{ip}${(je lui ai donné)$_{AP}$ (un livre)$_{AP}$}$_{ip}$]$_{IP}$

b. Ajouts à S et cadratifs
<u>Chaque lundi</u>, Paul n'est pas là.
[{(Chaque lundi)$_{AP}$}$_{ip}$ {(Paul n'est pas là)$_{AP}$}$_{ip}$]$_{IP}$

<u>Quand je vais à Toulouse</u>, je prends toujours le train.
[{(Quand je vais)$_{AP}$ (à Toulouse)$_{AP}$}$_{ip}$ {(je prends toujours le train)$_{AP}$}$_{ip}$]$_{IP}$

c. Constituants incidents, parenthétiques
François, <u>d'après ce qu'on m'a dit</u>, va partir en vacances en Grèce.
[{(François)$_{AP}$}$_{ip}$ {(d'après ce qu'on m'a dit)$_{AP}$}$_{ip}$ {(va partir)$_{AP}$ (en vacances)$_{AP}$ (en Grèce)$_{AP}$}$_{ip}$]$_{IP}$

Dans (Delais-Roussarie et al, 2016), comme dans bien d'autres études menées dans le cadre métrique-autosegmental, le syntagme intonatif englobe généralement la phrase dans son ensemble ou toute proposition ou clause indépendante. Les constituants détachés, souvent entourés de virgules, sont alors traités comme des syntagmes intermédiaires. Notons néanmoins que les contours intonatifs réalisés à la frontière droite de ces syntagmes intermédiaires varient considérablement (cf. Delais-Roussarie & Feldhausen, 2014 ; Avanzi, 2012). Remarquons cependant que ces syntagmes intermédiaires dérivés à partir de constructions particulières ne peuvent pas se restructurer avec le matériel qui suit, même lorsque le constituant syntaxique incident est très court (Delais-Roussarie & Feldhausen, 2016). Cela explique pourquoi certains auteurs considèrent que ces constructions appellent la réalisation d'une frontière de syntagme intonatif à leur droite (Mertens, 2008 ; Delais-Roussarie & Post, 2008).

Pour résumer, la variabilité observée dans le choix des contours en même temps que l'impossible restructuration prosodique peuvent expliquer les discussions sur la nature et la force relative de la frontière prosodique : syntagme intermédiaire ou syntagme intonatif ?

2.3 Frontière d'*ip*s et focus informationnnel

Si on considère que le syntagme intonatif (IP) correspond à une phrase ou une proposition indépendante (Delais-Roussarie et al., 2016), les constituants de début de phrases sont traités comme

des syntagmes intermédiaires, bien qu'ils puissent être utilisés comme des énoncés ou propositions indépendants et elliptiques (cf. les exemples sous (7) extraits de Delais-Roussarie et al., 2016.).

(7) a. A: Elle est enceinte de qui ?
B : de son mari, pardi !
[{(de son mari)$_{AP}$}$_{ip}${(pardi)$_{AP}$}$_{ip}$]$_{IP}$

b. A: Qu'est-ce que vous voulez ?
B: je voudrais des oranges, s'il vous plaît madame.
[{(je voudrais)$_{AP}$ (des oranges)$_{AP}$}$_{ip}$ {(s'il vous plait)$_{AP}$ (madame)$_{AP}$}$_{ip}$]$_{IP}$

c. A: Vous voulez des citrons ?
B: Non, ce sont des oranges que je veux.
[non]$_{IP}$ [{(ce sont des oranges)$_{AP}$}$_{ip}${(que je veux)$_{AP}$}$_{ip}$]$_{IP}$

Les frontières d'*ip*s après *mari* (7a), et *oranges* (7b et 7c) coïncident avec la frontière droite du focus informationnel, et les éléments soulignés pourraient être une réponse elliptique aux questions posées. Sur la base des contours intonatifs observés à la fin de ces séquences, certains travaux analysent cette frontière comme une frontière d'IP plutôt que d'*ip* (Martin, 1987 ; Delais-Roussarie & Post, 2008 ; Delais-Roussarie & Rialland, 2005). Il est clair en tout cas que la partie de l'énoncé analysé comme formant un *ip* comprend le focus informationnel, ce qui suit n'étant pas important pour le contenu informatif et pouvant être traité comme un appendice (ou *« tail »*).

3 Patrons intonatifs et syntagme intermédiaire

Dans la section 2, trois types distincts de syntagme intermédiaire ont été définis sur la base de la localisation de leurs frontières. Dans cette section, nous nous intéressons aux patrons intonatifs observés aux frontières des *ip*s, mais un bref rappel sur l'intonation du français sera proposé au préalable.

3.1 Inventaire tonal du français

L'intonation du français consiste en une succession de contours mélodiques réalisés à la fin de tout constituant prosodique. En position interne, ces contours sont généralement montants, quel que soit le niveau de structuration (LH* pour les syntagmes accentuels, LH*H- pour les syntagmes intermédiaires et LH*H% pour les syntagmes intonatifs, cf. Post, 2000 ; Delais-Roussarie et al, 2016 ; Delais-Roussarie & Post, 2008 ; Delattre, 1966). En revanche, une plus large variété de contours peut apparaître en fin d'énoncé ou à la fin des syntagmes prosodiques intégrant le focus informationnel.

Dans de nombreux travaux consacrés à l'intonation du français, on distingue généralement deux catégories de contours intonatifs : les contours terminaux et les non-terminaux (Post, 2000 ; Delais-Roussarie et al., 2016 ; Martin, 1987 par exemple). Pour (Martin, 1987), la forme des contours terminaux (contours $C_{i,}$) dépend de la modalité de l'énoncé (assertion *vs* question), celle des autres dépend de leur position linéaire et des degrés d'enchâssement, une fois la forme de C_i posée. Dans (Post, 2000) comme dans (Delais-Roussarie et al. 2016 et Delais-Roussarie & Rialland, 2005), une distinction est faite entre contours terminaux and non-terminaux, l'inventaire des contours non-terminaux étant plus limité, comme on le voit dans la Table 1.

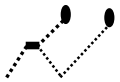
	Montant	Montant-Descendant	Descendant	Descendant après pic sur pénultième
Contours terminaux	LH*H% 	LH*L% 	L*L %, !H*L% 	H+!H*L% 
Contours non-terminaux	LH*, LH*H-, LH*H% 			

TABLE 1 : Inventaire des contours terminaux et non-terminaux du français

3.2 Quel contour à la fin des syntagmes intermédiaire ?

D'après la Table 1, des contours montants sont généralement observés à la fin des syntagmes accentuels et intermédiaires non-terminaux. Cela a été confirmé dans les analyses présentées dans (Delais-Roussarie et al., 2016) et dans (Michelas, 2011). Dans la plupart des exemples proposés, un contour mélodique montant est en effet observé à la frontière droite des syntagmes intermédiaires, qu'ils renferment un SN sujet long (cf. (4)) ou que leur frontière droite coïncide avec la fin d'un constituant topique ou disloqué comme sous (9), (cf. aussi figure 1).

(9) Les amis du mari de Valérie, je les ai appelés.
[{(les amis)$_{AP}$ (du mari)$_{AP}$ (de Valérie)$_{AP}$}$_{ip}$ {(je les ai appelés)$_{AP}$}$_{ip}$]$_{IP}$

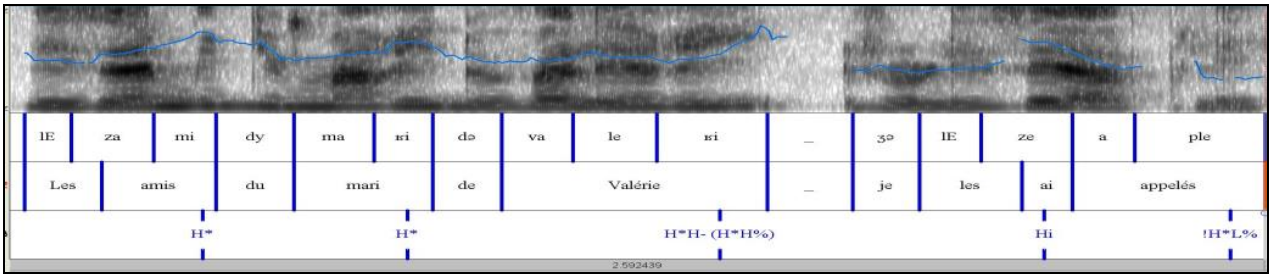


FIGURE 1 : Contour associé à (9)

Les contours intonatifs non-montants sont presque exclusivement observés dans les cas où la frontière droite de l'*ip* coïncide avec la fin du domaine focal. Dans les exemples sous (7) ont été observées en (7a) une descente après un pic sur la pénultième (H+H*L-), en (7b et c) une descente (codée !H*L- ou L*L-). Il est intéressant de noter que la forme de ces contours correspond à ce qui est considéré comme contour terminal dans la littérature (Delais-Roussarie et al., 2016 ; Post, 2000). La séquence de l'énoncé qui suit l'élément focal (et donc la frontière d'*ip*) est généralement prononcée avec une intonation relativement plate, similaire à celle des appendices (Delattre, 1966 ; Wunderli, 1987). En s'appuyant sur ces observations, une distinction peut être faite entre *ip*s métriques et incidents, d'une part, et *ip*s dont le bord droit coïncide avec la fin du domaine focal, d'autre part. Au vu des contours intonatifs observés à la frontière droite des domaines focaux (comme dans les exemples sous (7), on peut considérer qu'on a affaire à une frontière terminale, et donc une fin d'IP. Nous y reviendrons dans la section 4.

3.3 Réalisation des contours : de la phonologie à l'implémentation phonétique

Comme nous l'avons mentionné précédemment, un contour montant est généralement observé à la fin des syntagmes intermédiaires métriques et incidents. Dans certains cas cependant, un contour descendant est réalisé à la fin de ces syntagmes intermédiaires non terminaux. Une analyse de ces énoncés pousse à faire une distinction entre *ip*s métriques et *ip*s incidents.

En (9), représenté sur la figure 1, un contour montant LH*H- est réalisé à la frontière droite de l'*ip* incident qui correspond au SN disloqué *les amis du mari de Valérie* tandis que l'énoncé dans son entier s'achève par un contour descendant !H* L%. Mais, lorsque la proposition indépendante à laquelle se rattache le SN disloqué s'achève par un contour montant H*H% comme en (10) et (11), le contour à la fin de l'*ip* incident est généralement descendant (voir fig 2, mais aussi fig. 3.7 dans Delais-Roussarie et al., 2016), la chute de F0 pouvant être retardée (delayed).

(10) Les amis du mari de Valérie, vous les avez appelés ?
[{(les amis)$_{AP}$(du mari)$_{AP}$ (de Valérie)$_{AP}$↘}$_{ip}$ {.....}$_{ip}$]$_{IP}$

(11) Les amis du mari de Valérie, je les ai appelés et nous nous sommes rencontrés (fig. 2).
[{(les amis)$_{AP}$(du mari)$_{AP}$ (de Valérie)$_{AP}$↘}$_{ip}$ {(je les ai appelés)$_{AP}$}$_{ip}$]$_{IP}$ [...et nous nous sommes rencontrés..]$_{IP}$

Ces réalisations descendantes des contours de continuation ont été mentionnées par (Martin, 1987) et (Delattre, 1966). (Martin, 1987) soutient qu'elles s'expliquent par un besoin de marquer le contraste entre le contour de continuation mineure (Delattre, 1966) et les contours montants apparaissant en fin de clause (*continuation majeure*) ou de questions déclaratives (*contour de questions*). Il est cependant intéressant de noter que ces réalisations descendantes n'apparaissent quasiment jamais à la fin des *ip*s métriques. Lorsque un SN sujet long est réalisé dans une proposition indépendante non finale ou dans une question déclarative (comme sous (12)), le contour montant attendu à la droite de l'*ip* comprenant le SN sujet n'est pas clairement marqué du fait d'une accélération de débit (cf. fig. 3), mais aucune chute de F0 n'est réalisée ensuite, contrairement à ce qu'on observe en (10) et (11).

(12) Les amis du mari de Valérie m'ont appelé et nous nous sommes rencontrés.
[{(les amis)$_{AP}$(du mari)$_{AP}$ (de Valérie)$_{AP}$}$_{ip}$ {(m'ont appelé)$_{AP}$}$_{ip}$]$_{IP}$ [...et nous nous sommes rencontrés..]$_{IP}$

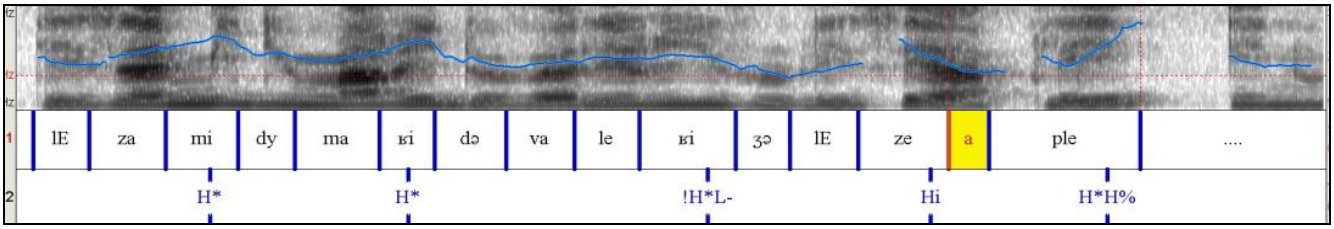


FIGURE 2 : Contour associé à (11)

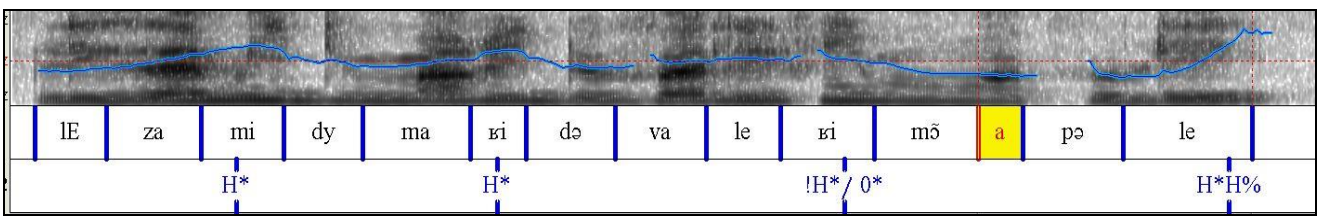


FIGURE 3 : Contour associé à (12)

Les changements dans la façon d'implémenter phonétiquement le contour réalisé sur la dernière syllabe de l'*ip* métrique dans l'énoncé sous (12) sont comparables à ceux observés par (Michelas, 2011) lorsque les sujets parlent plus vite. Cela va aussi dans le sens de l'analyse proposée par (Post, 2000) qui montre que des désaccentuations totales ou partielles peuvent apparaître au niveau du syntagme phonologique. Ces phénomènes conduisent (i) à assigner un statut métrique aux *ip*s de ce genre, et (ii) à analyser le contour mélodique de fin d'*ip* comme un accent syntagmatique.

4 Proposition et conclusion

Les différences observées dans la distribution et la réalisation des patrons intonatifs de fin d'*ip*s militent en faveur d'une distinction entre les trois types d'*ip*s proposée dans la section 2. Au vu des patrons intonatifs et des phénomènes de contrastes observés à la fin des syntagmes intermédiaires correspondant à des frontières de clauses ou de domaine focal d'une part, et à la fin des *ip*s incidents d'autre part, nous considérons ces deux types d'*ip*s comme des IPs. Pour rendre compte du fait que ces deux types de syntagmes intonatifs ne reçoivent pas les mêmes contours (terminal *vs.* non-terminal) et qu'une relation de dépendance ou de contraste existe entre eux, nous proposons de distinguer deux types d'IP: (i) l'**IP Majeur** qui correspond à des propositions ou clauses indépendantes ou à des séquences elliptiques contenant le focus informationnel de l'énoncé ; et (ii) l'**IP Mineur** dont la frontière droite correspond à la frontière droite d'un syntagme CommaP ou TopicP.

Les contours terminaux sont réalisés à la fin des IPs majeurs, et sont ensuite copiés à la fin de l'énoncé, comme on peut le voir pour (7) répété ici en (13), cela donnant naissance à une structure récursive.

(13) [[(je voudrais)$_{AP}$ (des oranges)$_{AP}$]$_{Major\text{-}IP}$ (s'il vous plait)$_{AP}$ (madame)$_{AP}$]$_{Copied\text{-}Major\text{-}IP}$

Quant aux IPs mineurs, ils ont à leur frontière droite un contour tonal dont la forme est en partie déterminée par la forme du contour terminal, même si le contour montant est le contour par défaut. Ces IPs entrent également dans une structure récursive avec les IPs Majeurs. Cf. (14a et b)

(14) a. [[(les amis)$_{AP}$ (du mari)$_{AP}$ (de Valérie)$_{AP}$]$_{Minor\text{-}IP}$ (je les ai appelés)$_{AP}$]$_{Major\text{-}IP}$

b. [[(les amis)$_{AP}$(du mari)$_{AP}$ (de Valérie)$_{AP}$↘]$_{Minor\text{-}IP}$ [(je les ai appelés)$_{AP}$] $_{Major\text{-}IP}$ [...et nous nous sommes rencontrés..] $_{Major\text{-}IP}$

Les IPs correspondant à des ajouts ou des topiques (IPs Mineurs) peuvent être analysés comme une sous-catégorie des IPs correspondant à des propositions indépendantes. Un parallèle peut être fait avec ce qui a été proposé pour distinguer les syntagmes mineurs des syntagmes majeurs (Ito & Mester, 2013), mais des recherches sont nécessaires pour les parenthétiques.

Contrairement aux *ip*s analysés comme IPs, les *ip*s métriques restent des *ip*s, mais ils peuvent être vus comme équivalant aux syntagmes phonologiques. APs et *ip*s peuvent donc être analysés comme des unités métriques équivalentes aux syntagmes phonologiques mineurs et majeurs (Nespor & Vogel, 1986 ; Selkirk, 1986). Du fait du syncrétisme entre intonation et accentuation en français, des événements intonatifs, qui pourraient être considérés comme des accents de groupe, se réalisent à la fin de ces unités. Deux arguments militent pour cette analyse : pas de réalisation descendante en cas de contraste et possibilité de restructuration. De plus amples recherches sont nécessaires pour valider cette analyse à l'aide de données de corpus et d'expériences.

Références

AVANZI, M. (2012) “La dislocation à gauche en français parlé. Etude instrumentale”, in *Le français moderne*.

DELAIS-ROUSSARIE E. (1996), “Phonological Phrasing and Accentuation in French”. in *Dam Phonology: HIL phonology papers* II. Holland Academic Graphics, The Hague, p. 1-38.

DELAIS-ROUSSARIE E & A. RIALLAND. Metrical organization, tonal association and focus in French. In S. Blaauw et F. Drijkoningen (eds.) *Romance languages and linguistic theory 2005*. John Benjamins Publishing, pages 73-98. 2005.

DELAIS-ROUSSARIE E. & POST B. (2008) “Unités prosodiques et grammaire de l’intonation : vers une nouvelle approche”, in *Actes des Journées d’étude sur la Parole JEP-TALN*.

DELAIS-ROUSSARIE E. & FELDHAUSEN I. (2014) “Variation in Boundary Strength in French.” in *Proceedings of Speech Prosody 2014*. Dublin, May 2014.

DELAIS-ROUSSARIE E., POST B., AVANZI M., BUTHKE C., DI CRISTO A., FELDHAUSEN I., JUN S.-A., MARTIN P., MEISENBURG T., RIALLAND A., SICHEL-BAZIN R., & YOO H.-Y., (2016), “Developing a ToBI system for French“, in S. Frota & P. Prieto [Eds], *Intonational Variation in Romance*, chapter 3, Oxford University Press.

DELATTRE P. Les dix intonations de base du français. *The French Review* 40 (1), pages 1-14, 1966.

FELDHAUSEN I., (2010) *Sentential Form and Prosodic Structure of Catalan*, John Benjamins.

FERY C. (2001), “Focus and Phrasing in French”. In Caroline Féry and Wolfgang Sternefeld (eds), *Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow*, Berlin. Akademie-Verlag, P. 153-181.

ITO J. & MESTER A. (2013) “Prosodic Subcategories in Japanese”. in *Lingua 124*. p.20-40.

JUN S.-A. & LEE H.-J. (1998) “Phonetic and Phonological markers of Contrastive Focus in Korean”, in *Proceedings of the 5th ICSLP, p.*4:1295-1298.

JUN SA., FOUGERON C. (2000) “A Phonological Model of French Intonation”. In: Botinis A. (eds) *Intonation. Text, Speech and Language Technology*, vol 15. Springer, Dordrecht, DOI 10.1007/978-94-011-4317-2_10

JUN S.A. & FOUGERON C. (2000). “A phonological model of French Intonation”. In *Intonation: Models, analysis and applications*. Cambridge University Press, Cambridge, p. 209-242

LADD R. D., (2008) *Intonational Phonology*, 2ème, CUP. DOI : 10.1017/CBO9780511808814

MARTIN P. (1987). Prosodic and rhythmic structures in French. in *Linguistics, 25*, p.925-949.

MERTENS P. (2008), “Syntaxe, Prosodie et structure informationnelle: une approche prédictive pour l’analyse de l’intonation dans le discours”, in *Travaux de Linguistique, 56,* p.97-124.

MICHELAS A. (2011), *Caractérisation phonétique et phonologique du syntagme intermédiaire en français: de la production à la perception*, Thèse de doctorat d'Aix-Marseille Université.

NESPOR,N. & VOGEL I., (1986, 2007), *Prosodic Phonology*, Mouton de Gruyter, (1986: Foris).

PIERREHUMBERT J., & BECKMAN M. (1988) *Japanese Tone Structure*, MIT Press

POST B. (2000) *Tonal and Phrasal Structures in French Intonation*. Thèse de Doctorat. Holland Academic Graphics, The Hague.

SELKIRK E. (1986), “On derived domains in sentence phonology”, in *Phonology 3*, p. 371-405.

SELKIRK, E. (2000) “The Interaction of Constraints on Prosodic Phrasing”, in M. Horne [Ed], *Prosody: Theory and Experiment,* Kluwer Academic Press, p. 231-261.

SELKIRK E. (2011), “The syntax-phonology interface”, in J. GOLDSMITH, J. RIGGLE & A. YU [Eds], *The Handbook of Phonological Theory*, 2ème edition, Blackwell, p. 435-484.

WUNDERLI P. (1987), *L’intonation des séquences extraposées en Français.* Gunter Narr Verlag.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Quel type de systèmes utiliser pour la transcription automatique du français ? Les HMM font de la résistance.

Paul Deléglise[1] Carole Lailler[2]
(1) LIUM, Université du Maine, avenue Olivier Messiaen, 72085 Le Mans Cedex 09, France
(2) Scribe-conseil, 9 avenue Adrien Daurelle, 05100 Briançon, France
paul.deleglise at gmail.com,c.lailler chez scribe-conseil.com

RÉSUMÉ

Forts d'une utilisation couronnée de succès en traduction automatique, les systèmes *end-to-end* dont la sortie réside en une suite de caractères, ont vu leur utilisation étendue à la transcription automatique de la parole. De nombreuses comparaisons ont alors été effectuées sur des corpus anglais libres de droits, de parole lue. Nous proposons ici de réaliser une comparaison entre deux systèmes état de l'art, non pas sur de la parole lue mais bel et bien sur un corpus d'émissions audiovisuelles françaises présentant différents degrés de spontanéité. Le premier est un *end-to-end* et le second est un système hybride (HMM/DNN). L'obtention de résultats satisfaisants pour le *end-to-end* nécessitant un lexique et modèle de langage dédiés, il est intéressant de constater qu'une meilleure intégration dans les systèmes hybrides (HMM/DNN) est source de performances supérieures, notamment en Français où le contexte est primordial pour capturer un énoncé.

ABSTRACT

What system for the automatic transcription of French in audiovisual broadcasts ?

With a successful use in machine translation, Character-based Neural Machine were naturally used in automatic speech transcription. Many comparisons were then made on English free corpora of speech read. We deal here with a comparison between two state-of-the-art systems, not on read speech but rather on a corpus of French audiovisual broadcasts with different degrees of spontaneity. The first is a *end-to-end* and the second is an hybrid system (HMM / DNN). As obtaining good results for the *end-to-end* requires a dedicated vocabulary and a language model, the goal is to see if better integration in hybrid systems (HMM / DNN) is really a source of superior performances, especially in French language where context is essential for syntax.

MOTS-CLÉS : Transcription Automatique de la parole, Évaluation, end-to-end, HMM/DNN.

KEYWORDS: ASR, Evaluation, end-to-end, HMM/DNN.

1 Introduction

Depuis leur apparition dans le domaine de la transcription automatique de la parole en proposant en 2013 une transcription en phonèmes suivie en 2014 par une transcription en mots sans connaissances phonétiques (Graves & Jaitly, 2014), les systèmes *end-to-end* s'imposent comme une alternative sérieuse aux systèmes à base de HMM. (Amodei *et al.*, 2016) rapportent une évaluation sur le test du corpus Librispeech (Panayotov *et al.*, 2015), obtenue en utilisant un corpus d'apprentissage d'une dizaine de milliers d'heures. Les résultats sont notamment comparés avec la performance d'un

Volume 1 : Journées d'Études sur la Parole, pages 154–162.
HAL : hal-02798531.

locuteur humain. La taille du corpus employé pose cependant un net problème de reproductibilité des expériences. (Zeghidour *et al.*, 2018) font état de meilleurs résultats avec une architecture de réseaux différente sur Librispeech (3,26% partie *clean* , 12,76% partie *other*) en utilisant "seulement" 960 heures qui correspondent à l'apprentissage de ce même corpus. Dans ce même article, les résultats sont comparés avec ceux d'un système hybride (HMM/DNN) (Han *et al.*, 2017) (3,51% partie *clean*, 8,58% partie *other*). Enfin, (Lüscher *et al.*, 2019) présentent deux architectures : l'une hybride (HMM/DNN), l'autre de type *end-to-end* dont les résultats sont respectivement de 5.0% et 9.3% sur la partie *other*. Toutes les modélisations *end-to-end* utilisent un modèle de langage construit sur des données textuelles en appui du DNN appris, lui, sur les corpus oraux. Il ressort que les différences avec un modèle hybride portent sur l'utilisation de connaissances phonétiques préalables comme le dictionnaire de phonétisation et le processus d'alignement signal/symbole : HMM dans un cas, neuronal (convolution, encodeur-décodeur) dans l'autre.

Comme l'ensemble des évaluations de ces systèmes *end-to-end* est présenté en langue anglaise sur des corpus de parole lue, nous proposons d'étudier ces différences sur du français, avec un corpus d'émissions audiovisuelles comprenant une part non négligeable de parole spontanée. Nous pourrons ainsi considérer l'influence de la langue, y compris et surtout dans sa dimension spontanée, sur les résultats. L'article se compose d'une description des systèmes et corpus utilisés. Nous présentons ensuite les processus d'apprentissage, les expériences de transcription et leur évaluation. Enfin, nous analysons les résultats obtenus.

2 Description des systèmes

Avant de décrire les 2 types de systèmes utilisés, nous présentons les 2 composants communs des systèmes construits, à savoir une étape de segmentation automatique préalable au décodage et une modélisation du langage.

La segmentation automatique est obligatoire puisqu'aucun des deux systèmes ne permet un décodage en flux. En outre, il n'est pas question d'utiliser une segmentation manuelle pour l'évaluation car la segmentation automatique influence les résultats. Cette segmentation est réalisée par le système Lium Spkdiarization (Barras *et al.*, 2006; Meignier & Merlin, 2010) sans utiliser la tâche de détection parole/silence/musique, car le décodeur réalise cette tâche de meilleure manière. En outre, le regroupement de locuteurs à l'aide du critère de vraisemblance croisée pour obtenir une homogénéité des classes n'est pas utile dans le décodage. Il est donc omis.

Pour réaliser les modèles de langage, nous avons opté pour une modélisation avec repli, la différence avec les modèles continus étant limitée jusqu'à l'apparition des modèles de langage à *Transformer* introduit dans (Lüscher *et al.*, 2019). Nos modèles de langage sont construits avec l'outil POCOLM [1] qui optimise le poids des différents corpus en fonction de la cible de manière plus pertinente que les autres boîtes à outils.

2.1 Système hybride

Nous avons choisi comme représentant d'un système hybride (HMM/DNN), un système utilisant la boîte à outils KALDI (Povey *et al.*, 2011) dont les performances ont fortement augmenté suite

1. https://github.com/danpovey/pocolm

à l'utilisation de DNN pour le calcul de la probabilité d'une trame (Povey *et al.*, 2016), avec un apprentissage discriminant de type LF-MMI. Un modèle de Markov caché utilisant des mélanges de gaussiennes est construit par étape. L'avant-dernière étape utilise une représentation par un vecteur de 40 paramètres issu d'une LDA/MLLT calculée sur la concaténation des paramètres de 9 trames consécutives où chaque trame est paramétrée par 13 coefficients MFCC auxquels sont ajoutées les dérivées premières et secondes. Dans la dernière étape, une adaptation au locuteur est réalisée par un processus de fMLLR (Gales, 1998). Ce modèle ne servant qu'à fournir, pour l'apprentissage du DNN, un alignement des états des HMM/signal de parole, l'apprentissage discriminant pour le modèle gaussien présent dans la recette KALDI n'est pas effectué.

Pour le DNN, nous avons choisi l'architecture tdnn7n présente dans la recette *Switchboard* de Kaldi. Elle comprend schématiquement 11 couches de TDNN intercalées avec des couches de régularisation de type ReLu pour un total de 20 millions de paramètres et une couche de sortie de taille 11500, qui correspond au nombre d'états partagés de l'étape fMLLR. Les paramètres d'entrée sont des MFFC calculés sur 40 bandes spectrales en gardant tous les coefficients cepstraux d'une part, et un i-vector de dimension 100 calculés sur une fenêtre glissante caractérisant le locuteur d'autre part.

Le décodage après le calcul des i-vectors s'effectue en 3 passes :

1. une première utilise le HMM hybride avec un modèle de langage d'ordre 2 pour calculer un treillis pour chaque segment,
2. puis, une réévaluation de ces derniers par un modèle de langage d'ordre 4 est effectuée,
3. enfin, le calcul de la meilleure séquence de mots en utilisant différentes valeurs pour le poids du LM et la pénalité d'insertion des mots à partir des treillis est opéré.

2.2 Système *End-to-end*

Pour ce système, nous avons opté pour le *wav2letter++* (Pratap *et al.*, 2019) dans sa version convolutionnelle (Zeghidour *et al.*, 2018) avec un critère d'apprentissage de type *AutoSegCriterion*. Les paramètres d'entrée du réseau sont 40 coefficients log-Mels calculés à la volée, solution qui s'est montrée plus efficace que l'utilisation de MFCC sur Librispeech. La sortie du réseau de neurones est une suite de symboles : ici des lettres. Ainsi, une liste de mots est obtenue directement en utilisant un symbole spécial indiquant la fin d'un mot. Pour obtenir de meilleures performances, un élargissement à l'utilisation d'un lexique et d'un modèle de langage est proposé. Ce dernier peut être à base de mots ou de lettres. Cela correspond à la version *conv_glu* de la recette dédiée à Librispeech. Ce réseau comprend 19 couches convolutives intercalées de couches de normalisation, dropout, et Relu. Il a 200 millions de paramètres. Par ailleurs, il faut noter l'absence de toute indication temporelle associée aux mots dans la sortie.

3 Données d'apprentissage

3.1 L'acoustique

Les données acoustiques pour l'apprentissage sont les audios associés à leur transcription fine provenant des différentes campagnes d'évaluation du français. Elles comportent des émissions de radio pour les 2 campagnes ESTER (Gravier *et al.*, 2004) dont la parole est, le plus souvent, préparée, auxquelles sont ajoutées les émissions de télévision de la campagne ETAPE (Gravier *et al.*, 2012)

qui contient davantage de parole spontanée. Des transcriptions larges de podcasts complètent ce corpus d'apprentissage. Les corpus de développement et de test sont ceux de l'évaluation finale du défi REPERE (Giraudel *et al.*, 2012). Le tableau 1 donne le nombre d'heures transcrites des différents corpus. Le corpus d'apprentissage est utilisé tel quel pour l'apprentissage des GMM. Pour l'apprentissage des DNN en revanche, une augmentation de corpus est effectuée. Pour cela, on pratique une modification de la vitesse de 0,9 et de 1,1, conjuguée à une variation aléatoire de l'amplitude par émission : cela multiplie par 3 le nombre d'heures disponibles pour l'apprentissage.

Corpus	Taille en heures	Nb émissions
Apprentissage campagne évaluation	291h	759
Apprentissage podcast	225h	1229
Total Apprentissage	617h	1988
Développement	5h30	28
Test	9h25	62

TABLE 1 – Les corpus audio

3.2 Les données textuelles et les modèles de langage

Nous regroupons dans ce paragraphe les informations sur les données textuelles et les modèles de langage utilisés par les 2 types de systèmes.

3.2.1 Les textes

Les textes proviennent de 6 sources différentes auxquelles il faut ajouter la transcription du corpus de développement qui sert de cible pour l'optimisation des LM. Le vocabulaire a une taille de 160 000 mots. Il est construit sur la réunion des mots présents dans les transcriptions des corpus des campagnes d'évaluation ESTER et des mots les plus fréquents sur les autres corpus. La table 2 donne le nombre d'occurrences de mot par corpus. Les extraits de corpus sont obtenus par une méthode d'entropie croisée entre la source originale et le corpus de développement pour ne garder que les phrases les plus pertinentes.

Corpus	Nb occurrences
Audio transcrites	8M de mots
Sites web de télévision	5M
Extraits (40%) de Google News ($\leq$ 2012)	80M
Extraits (75%) de French Gigaword ($\leq$ 2012)	753M
Extraits (84 %) du journal le monde de 1988 à 2003	316M
Sous titre de journaux télévisés (OCR et télétexte)	11M
Transcription du corpus de développement	5,2k

TABLE 2 – Les textes

3.2.2 Les modèles de langage

Deux modèles sont construits : un modèle 2-gram et un modèle 4-gram. Le modèle 2-G n'est cependant utilisé que par la première passe du modèle hybride. Les tailles de ces modèles sont présentées dans

le tableau 3. Trois modèles de langage sur les séquences de lettres sont également calculés pour le modèle *end-to-end*. Ils sont respectivement d'ordre 2, 7 et 10 et contiennent respectivement 2192, 7M et 26 M n-grams pour des perplexités respectivement de 7.9, 3.2 et 2.87.

modèle	1-gram	2-gram	3-gram	4-gram	perplexité
2G	160 k	1,70 M	-	-	152,39
4G	160 k	24,00 M	140 M	338 M	84,35

TABLE 3 – Modèles de langage

4 Apprentissage de modèles et optimisation des paramètres

4.1 Système hybride

Pour ce dernier, seul un dictionnaire de phonétisation lui est fourni. Il contient l'ensemble des mots issus des transcriptions du corpus d'apprentissage (phase apprentissage) et le vocabulaire des modèles langage (phase décodage). Ce dictionnaire est construit à l'aide de BDLEX, puis de LIAPHON contrôlé par un alignement forcé sur le corpus d'apprentissage. Ensuite, il ne reste plus qu'à lancer le script fourni dans la recette. L'apprentissage du modèle acoustique prend 70 heures sur une machine ayant 24 cœurs et de 2 cartes GPU de modèle GTX Titan X dotées chacune de 12 Go de ram.

L' optimisation des paramètres est effectuée sur le corpus de développement. Elle consiste à choisir le meilleur poids pour le modèle de langage et la pénalité d'insertion. Comme ces poids ne servent que dans le calcul des CTM à partir des treillis, un algorithme purement combinatoire est utilisé sur les 39 combinaisons les plus susceptibles de contenir l'optimal sur cette passe très rapide.

4.2 Système *end-to-end*

L'apprentissage doit commencer par une phase de préparation des données avec la création d'un fichier par segment de parole. En outre, la représentation des mots doit être spécifiée : nous avons tout d'abord choisi de les représenter, classiquement, par la suite des lettres en testant divers regroupements que nous détaillerons dans la partie 5.1.1. Nous avons rajouté 2 mots et 2 symboles spécifiques pour représenter d'une part les silences et d'autre part, les inspirations et autres bruits. Pour lancer l'algorithme sur la machine dont on dispose, nous devons limiter la taille des minibatchs à 4. Par ailleurs, l'algorithme d'apprentissage ne progresse pas si le corpus comprend des segments trop longs. Le plus usuel est de commencer en restreignant le corpus à des segments de petite taille comme dans (Lüscher *et al.*, 2019). Toutefois, cela peut conduire à l'introduction d'un biais dans l'initialisation du modèle. Nous avons donc préféré utiliser les alignements effectués lors de l'apprentissage du système hybride pour redécouper le corpus en segment de moins de 8 secondes, grâce aux silences présents dans le signal. Lors de cet apprentissage, le corpus de développement sert à contrôler le nombre d'itérations. La convergence prend une vingtaine de jours sur la machine décrite dans 4.1. Cette durée importante ne nous a pas permis de tester les métaparamètres de l'apprentissage du modèle.

Une phase d'optimisation est aussi réalisée pour les paramètres du décodage : poids du modèle de langage, probabilité du silence et des autres bruits, probabilité de production d'un mot. Cette étape est faite avec l'algorithme du simplex en quelques jours de calculs.

5 Expériences et évaluation

5.1 Expériences

En plus de la comparaison entre les 2 types de systèmes, nous avons aussi testé différentes variantes dans la représentation des mots et dans l'algorithme de décodage pour le système *end-to-end*.

5.1.1 Représentation de mots

Nous avons testé 3 solutions pour cette représentation :

1. Représentation par la suite de lettres suivies d'un symbole de fin de mot : cette solution produit 44 lettres correspondant aux 26 lettres de l'alphabet et à leurs versions accentuées ainsi que 4 symboles spécifiques « `|` `'` `S` `B` ». Cette représentation sera notée *Base*.
2. Dans le but de diminuer la couche de sortie et de tenir compte des faibles différences de prononciation de certaines lettres accentuées, nous avons regroupé l'ensemble des ces lettres sur leur version non-accentuée à l'exception des lettres *é* et *è* réunies ensemble. Ceci réduit l'ensemble à un total de 35 symboles. Cette représentation sera notée *Regr*.
3. Comme une des spécificités du français est une plus grande coarticulation entre les mots allant jusqu'à la liaison, nous avons voulu tester une représentation sans symbole de fin de mots. Elle sera notée *NoSep*.

5.1.2 Algorithme de décodage

Plusieurs possibilités existent pour guider la recherche en faisceau du système *end-to-end*. En plus d'une transcription contrainte par un lexique, nous avons testé pour la solution *Base*, la comparaison entre un modèle de mots 4-grammes et des modèles sur les lettres d'ordre 2, 7 et 10 notés *Lettre2*, *Lettre7* et *Lettre10*. Nous n'avons pas été jusqu'à un ordre de 15, comme dans (Likhomanenko *et al.*, 2019), car le peu de variabilité de nos corpus de textes entrainait des problèmes de convergence numérique dans les algorithmes d'élagage du modèle de langage. Pour les solutions *Regr* et *NoSep*, nous avons utilisé seulement le modèle en mots.

5.2 Évaluation

Pour cette dernière phase de travail, nous avons utilisé les outils du défi REPERE (Giraudel *et al.*, 2012) qui demandent un étiquetage temporel des mots. Pour obtenir ce dernier pour les sorties du système *end-to-end*, nous avons dû procéder en plusieurs temps :

— une équi-répartition temporelle des mots dans le segment de la transcription automatique,
— suivie d'une première évaluation qui fournit 2 suites *sentences* parallèles ; l'une avec le texte de la référence, l'autre avec celui de l'hypothèse.
— l'utilisation de *mwerSegmenter*[2] a permis de répartir les mots de l'hypothèse sur chaque *sentence* de la référence (et par là-même l'intervalle de temps desdites *sentences*). Cette réaffectation s'est faite à la marge, après la première évaluation.
— enfin, une nouvelle équi-répartition des mots a été réalisée en utilisant les temps de l'étape précédente.

Nous présentons ci-après, dans le tableau 4, les résultats de l'ensemble des systèmes sur les corpus de développement et de test.

2. https://www- i6.informatik.rwth- aachen.de/web/ Software/mwerSegmenter.tar.gz

Corpus	Système	Cor.	Sub.	Sup.	Ins.	WER
Dev.	Kaldi	88,18	7,83	3,63	3,14	**14,60**
Dev.	Base	76,84	9,72	12,95	4,14	26,81
Dev.	Regr	74,69	7,27	17,54	0,83	**25,63**
Dev.	NoSep	72,38	17,20	9,93	3,17	30,31
Dev.	Lettre2	60,86	23,91	14,74	5,54	44,19
Dev.	Lettre7	73,59	18,17	7,76	6,61	32,53
Dev.	Lettre10	75,52	16,82	7,17	6,55	30,54
Test	Kaldi	88,63	7,17	3,58	2,80	**13,55**
Test	Base	80,19	10,21	8,95	4,41	**23,58**
Test	Regr	74,55	6,69	18,07	0,63	25,40
Test	NoSep	73,65	16,68	9,02	2,96	28,65
Test	Lettre2	59,49	24,63	15,22	5,31	45,16
Test	Lettre7	73,74	17,71	7,90	6,36	31,97
Test	Lettre10	75,72	16,25	7,37	6,48	30,11

TABLE 4 – Résultats de l'évaluation

6 Discussions

6.1 Les *end-to-end*

Avant même de comparer avec le système KALDI, un des premiers enseignements à tirer réside dans le fait que le système *end-to-end* a besoin d'un séparateur de mots bien que celui-ci n'ait aucune réalité en français (*Base* vs *Regr*). Pour le regroupement des lettres, la réduction de nombre de paramètres (11830 sur 200 millions) ne semble pas jouer. La différence entre les résultats du corpus de test par rapport au corpus de développement nécessite une analyse des résultats par type d'émissions. Concernant les décodages avec LM lettres, ils s'améliorent certes avec l'ordre mais les temps de calcul deviennent aussi importants que pour le LM classique. On s'aperçoit rapidement que le modèle d'ordre 2 est effectivement intéressant dans sa vitesse d'exécution, notamment pour une recherche en mots clés avec tolérance à l'orthographe, ce qui peut présenter un véritable intérêt dans un couplage avec des *embedings*.

6.2 KALDI vs Wav2letter

Le système KALDI présente une amélioration d'environ 40% en relatif par rapport au meilleur *end-to-end* : *Base*. Ce taux est similaire à celui présenté dans (Lüscher *et al.*, 2019). Cet écart important a deux explications principales. La première réside dans l'utilisation de connaissances phonétiques à travers le dictionnaire. Celles-ci ont l'avantage de mentionner les variations de prononciation des homographes hétérophones : « *Les poules du couvent couvent.* ». Le deuxième point concerne un problème algorithmique : avec les seuils utilisés dans la recherche en faisceau, le temps de décodage pour KALDI est de 7 minutes pour 5h30 de parole *vs* 1h30 pour le *end-to-end* en utilisant tous les coeurs de la machine. Il semble donc difficile d'élargir le champ de recherche pour le *end-to-end*. On pourrait sans doute augmenter son efficacité en faisant une première recherche avec un ML sur les lettres, puis en utilisant un ML sur les mots. Toutefois, l'efficacité de KALDI réside principalement dans la réunion, en un seul réseau de type FST (Mohri *et al.*, 2002), de l'ensemble

des connaissances acoustiques, phonétiques, lexicales, langagières. Après compilation de ce réseau, les poids des arcs sont poussés au maximum vers le début du réseau, ainsi les procédures d'élagage peuvent fonctionner au niveau de la trame avec le maximum d'informations. Il n'y a pas de "coupure" entre l'acoustique, le lexique, et le modèle de langage. Or, dans le cas du *end-to-end*, ce sont bel et bien trois entités distinctes lors de la recherche en faisceau. Pour une étude plus fine, une comparaison entre les 2 systèmes est présentée dans le tableau 5 sur les différents types d'émissions du corpus de test. L'émission LCP_topQuestions est une émission très facile à décoder, seulement ce type de langage relativement contraint dialogiquement est faiblement présent dans le corpus d'apprentissage acoustique. L'ajout du ML en fin de processus avec la limitation de la recherche en faisceau ne permet pas de faire survivre la bonne solution. Ce phénomène est présent, dans une moindre mesure, pour LCPActu. L'émission Culture et Vous, qui se caractérise par un français très spontané, décousu et bruité, présente, tout en ayant des écarts de scores en relatif stables, des variations en absolu allant jusqu'à rendre la transcription presque inutilisable, dans le cas du *end-to-end*.

Émissions	Kaldi	Base
BFMTV_BFMStory	14,24	23,03
BFMTV_CultureEtVous	21,61	41,14
BFMTV_RuthElkrief	16,29	24,95
LCP_CaVousRegarde	12,26	22,34
LCP_EntreLesLignes	11,57	22,15
LCP_LCPActu	9,06	17,61
LCP_LCPInfo	13,78	22,43
LCP_PileEtFace	10,75	21,59
LCP_TopQuestions	7,83	18,06

TABLE 5 – WER sur le corpus de test par type de systèmes

7 Conclusion

Dans cet article, nous avons présenté une comparaison entre un système hybride et un système *end-to-end*. Le peu de disponibilité de corpus textuels par rapport aux enregistrements audio implique la nécessaire utilisation, pour chacun, d'un modèle de langage. Or, l'intégration de ce dernier dans le processus de décodage du système hybride permet un élagage performant au niveau de la trame, ce qui rend ce dernier réellement plus efficace. Il faut noter aussi que, compte tenu du découplage entre langage et réseaux de neurones, une meilleure adaptation est réalisable en ne travaillant que sur un ML dédié (pour les émissions TOP_QUESTION par exemple). Ce phénomène confère au système hybride un avantage des plus sérieux, notamment si l'on souhaite l'intégrer en première brique à une chaîne de traitements complète comme celles des *chatbots*. Par ailleurs, ce système hybride présente l'intérêt d'un apprentissage nettement plus écologique : on constate ainsi un rapport de 1 à 10.

Références

AMODEI D., ANUBHAI R., BATTENBERG E. & ALL (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, p. 173–182 : JMLR.org.

BARRAS C., ZHU X., MEIGNIER S. & GAUVAIN J.-L. (2006). Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, **14**(5). DOI : 10.1109/TASL.2006.878261, HAL : hal-01434241.

GALES M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Comp. Speech and Lang.*, **12**, 75–98.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The REPERE corpus : a multimodal corpus for person recognition. In *The Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1102–1107, Istanbul, Turkey : European Language Resources Association (ELRA).

GRAVES A. & JAITLY N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *The 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II–1764–II–1772 : JMLR.org.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *The Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 114–118, Istanbul, Turkey : European Language Resources Association (ELRA).

GRAVIER G., BONASTRE J.-F., GEOFFROIS E. & ALL (2004). The ESTER evaluation campaign for the rich transcription of French broadcast news. In *The Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

HAN K. J., CHANDRASHEKARAN A., KIM J. & LANE I. (2017). The capio 2017 conversational speech recognition system. *Preprint :* arXiv:1801.00059 [cs.CL].

LIKHOMANENKO T., SYNNAEVE G. & COLLOBERT R. (2019). Who needs words ? lexicon-free speech recognition. In *Interspeech 2019* : ISCA. DOI : 10.21437/interspeech.2019-3107.

LÜSCHER C., BECK E., IRIE K., KITZA M., MICHEL W., ZEYER A., SCHLÜTER R. & NEY H. (2019). Rwth asr systems for librispeech : Hybrid vs attention. *Interspeech 2019*. DOI : 10.21437/interspeech.2019-1780.

MEIGNIER S. & MERLIN T. (2010). LIUM SPKDIARIZATION : AN OPEN SOURCE TOOLKIT FOR DIARIZATION. In *CMU SPUD Workshop*, Dallas, United States. HAL : hal-01433518.

MOHRI M., PEREIRA F. & RILEY M. (2002). Weighted finite-state transducers in speech recognition. *Comput. Speech Lang.*, **16**(1), 69–88. DOI : 10.1006/csla.2001.0184.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

POVEY D., GHOSHAL & ALL (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

POVEY D., PEDDINTI V., GALVEZ D. & ALL (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech 2016*, p. 2751–2755. DOI : 10.21437/Interspeech.2016-595.

PRATAP V., HANNUN A., XU Q., CAI J., KAHN J., SYNNAEVE G., LIPTCHINSKY V. & COLLOBERT R. (2019). Wav2letter++ : A fast open-source speech recognition system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6460–6464.

ZEGHIDOUR N., XU Q., LIPTCHINSKY V., USUNIER N., SYNNAEVE G. & COLLOBERT R. (2018). Fully convolutional speech recognition. *Preprint :* arXiv:1812.06864 [cs.CL].

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Adaptations sur le F1 et le débit en réponse à diverses perturbations

Ivana Didirková[1,2], Leonardo Lancia[2], Cécile Fougeron[2]
(1) UR 1569 TransCrit, Université Paris 8
(2) UMR7018 Laboratoire de phonétique et phonologie & CNRS, Université Paris 3
ivana.didirkova@univ-paris8.fr, {leonardo.lancia,
cecile.fougeron}@sorbonne-nouvelle.fr

RESUME

Nous comparons les effets de deux conditions induisant des stratégies d'adaptation différentes (retour auditif masqué et bite-block) avec les effets des modifications intentionnelles du débit de parole. Nous examinons ces effets en termes de F1 et en termes de débit articulatoire. Nous comparons ensuite les effets de ces mêmes perturbations sur les mêmes locuteurs, afin de définir si les stratégies inter- et intra-individuelles varient en fonction de la boucle perturbée (auditive ou somatosensorielle). Cinq locutrices ont été enregistrées sans perturbation, avec un retour auditif masqué, avec un bite-block et avec des changements de tempo instruits (lent et rapide). Les résultats montrent une augmentation du débit en parallèle d'une augmentation du F1 des voyelles ouvertes, ce qui permet de supposer que les modifications spectrales et les modifications de débit ne seraient pas corrélées. La même augmentation de F1 est observée lors d'une modification intentionnelle du débit de parole.

ABSTRACT

F1 and speech rate adaptations in response to various perturbations.
We compare the effects of two conditions inducing different adaptation strategies (masked auditory feedback and bite-block) with the effects of intentional speech rate modifications. We examine these effects in terms of F1 and in terms of speech rate. We then study the effects of these perturbations on same speakers, in order to determine whether intra- and inter-speaker strategies vary depending on the perturbed loop (auditory or somatosensory). Five speakers were recorded in five conditions: without any perturbation, with masked auditory feedback, with a bite-block and with instructed speech rate modification (slow and fast). Our results show an increase of speech rate concurrently with an increase of open vowels' F1, which allows us to suppose that spectral modifications and speech rate modifications would not be in correlation. The same F1 increase is observed during intentional speech rate modification.

MOTS-CLÉS : Retour auditif masqué, bite-block, vitesse d'articulation, F1, perturbation
KEYWORDS: Masked auditory feedback, bite-block, speech rate, F1, perturbation

Volume 1 : Journées d'Études sur la Parole, pages 163–171.
HAL : hal-02798532.

1 Introduction

Dans la production de la parole, les boucles de retours auditif et somatosensoriel jouent le rôle de mécanismes de contrôle qui vérifient de manière continue les deux sorties du processus de planification de la parole, à savoir les mouvements articulatoires et les patrons acoustiques qui en résultent (Houde et Nagarajan, 2011; Guenther, 2016 ; Parrell et al., 2019). Ces mécanismes permettent de corriger la sortie du système de production lorsque celle-ci est décalée par rapport aux intentions du locuteur. L'une des manières d'étudier expérimentalement cette adaptation consiste en une perturbation artificielle des niveaux auditif et / ou somatosensoriel. Ainsi, ces paradigmes expérimentaux de perturbation ont pour objectif de déstabiliser le fonctionnement typique du système, allant du masquage du feedback (masked auditory feedback, MAF) à l'introduction d'une sortie modifiée aux boucles de feedback avec par exemple des modifications de f0 ou de formants (frequency-shifted feedback, FSF), en passant par un décalage temporel entre les gestes utilisés pour rendre la parole audible et la sortie acoustique effective (delayed auditory feedback, DAF). D'autres s'intéressent davantage aux perturbations du retour somatosensoriel. Ces perturbations peuvent être induites à l'aide d'un bite-block (BB) devant être gardé dans la cavité buccale et stabilisant la mandibule (Hoole, 1987), d'un tube tenu entre les lèvres (lip tube) utilisé par exemple par Ménard et al. (2016), des perturbations mécaniques de la langue (Ito et al., 2019), voire des produits anesthésiants (Larson et al., 2008). Le principal postulat de ces études suppose un décalage entre le feedback attendu et réel, menant à une redéfinition des commandes motrices. Les réponses des participants à ces perturbations de feedback font ressortir des réactions variables. Par exemple, les études utilisant des modifications de f0 ou de structures formantiques ont, pour la plupart, démontré trois sortes de réactions possibles aux perturbations. Tandis que certains locuteurs ont tendance à compenser les perturbations induites, répondant ainsi à la perturbation en contrant cette dernière (si la perturbation consiste à augmenter la fréquence fondamentale, ces locuteurs vont l'abaisser), chez d'autres, la perturbation induit une réponse allant dans le sens de la perturbation (dans l'exemple, à augmenter la f0). Enfin, un troisième groupe de locuteurs ne réagira pas (ou peu) aux perturbations (Burnett et al., 1998; Jones & Munhall, 2000; MacDonald et al., 2011). Notons toutefois que la perturbation d'un paramètre n'entraîne pas uniquement les modifications de ce même paramètre et, inversement, les modifications d'un paramètre peuvent être la résultante d'une perturbation ne touchant pas directement ce paramètre. Il a par exemple été démontré que la simple absence de retour auditif perturbe le contrôle du f0 (Mallard et al., 1978) et fait augmenter le F1 des voyelles (Kirchhuebel, 2010). De plus, les réactions aux perturbations comme le MAF ou DAF ont également un impact sur certaines caractéristiques globales comme le débit de parole (Jacks & Haley, 2015; Maruta et al., 2014).

Ici, nous proposons de poursuivre ces recherches à travers la comparaison entre les effets des perturbations du retour auditif et ceux des perturbations du retour somatosensoriel. Plus concrètement, nous nous intéressons aux changements induits par le MAF et le BB et ce, en termes de caractéristiques spectrales et en termes de débit. La revue de la littérature montre en effet que ces deux paradigmes seraient à l'origine de stratégies adaptatives différentes de la part des locuteurs. Si l'on suppose que le MAF induit l'effet Lombard (Kirchhuebel, 2010), cette perturbation entraîne une augmentation de l'amplitude du signal acoustique visant à surmonter le bruit de fond et un

changement des caractéristiques spectrales visant à les rendre plus saillantes (Garnier et al., 2006). De l'autre côté, la parole avec un BB forcerait le locuteur à modifier ses commandes motrices, afin de garder les caractéristiques spectrales constantes (Fowler & Turvey, 1981). Ainsi, ces deux paradigmes de perturbation permettent de se pencher sur la fonction et le rôle des phénomènes adaptatifs en production de la parole : le MAF induirait une modification du spectre ayant une fonction communicative, ce qui n'est pas le cas pour les modifications du spectre dues à la présence du BB. Par ailleurs, le débit de parole sous MAF est aussi plus lent par rapport à celui de la parole pas perturbée (Garnier et al., 2010). Il s'agit dans ce cas de comprendre si cette réduction de débit est une conséquence des changements spectraux et d'intensité (supposant que l'hyperarticulation est corrélée avec un ralentissement de débit), ou s'il a la fonction de rendre plus saillants les sons de parole dans le but de préserver leur fonction communicative. Pour mieux caractériser la relation entre les modifications de débit et la saillance spectrale, nous allons comparer les effets du MAF et du BB avec les effets des modifications du débit de parole intentionnellement produit par les locuteurs. Un second objectif de cette étude est de comparer les effets de ces perturbations sur les mêmes locuteurs, afin de définir si les stratégies inter- et intra-individuelles varient selon que la perturbation concerne la boucle auditive ou la boucle somatosensorielle.

2 Méthodologie

Cinq locutrices de langue maternelle française, appariées en âge et en catégorie socio-professionnelle, ont été enregistrées durant plusieurs perturbations de leurs retours auditif et somatosensoriel (retour auditif masqué par un bruit, retour auditif retardé, retour auditif avec modification de fréquence fondamentale, *bite-block*), ainsi que durant une modification de débit de parole avec débit imposé. Chacune des perturbations a été administrée séparément des autres.

Les locutrices ont participé à trois séances d'approximativement 45 minutes chacune. Les séances se déroulaient dans une pièce calme. La première consistait en un enregistrement de la *baseline* (sans perturbation), suivi de la condition MAF, elle-même suivie d'une condition de lecture à débit rapide[1]. Pendant la condition de MAF, les locutrices étaient équipées d'un casque transmettant un bruit de « cocktail party », tandis que leur voix était transmise par voie aérienne. Avant de commencer cette phase de l'expérience, il leur a été demandé de compter en boucle de 1 à 10 pendant qu'elles entendaient leur voix transmise par le casque (sans bruit) et que le volume de ce signal était augmenté graduellement. Les locutrices devaient s'arrêter lorsque l'intensité de la voix véhiculée par le casque était égale à celle de leur voix transmise par voie aérienne. Cela a permis de régler le volume du bruit de « *cocktail party* » de façon à obtenir une différence de 12 dB entre son intensité et celle de la voix des locutrices. La deuxième session comprenait les modifications de f0 et l'utilisation de DAF, conditions qui ne feront pas l'objet de cette étude. Enfin, la troisième session contenait la lecture des phrases avec le *bite-block* (parallélépipède en plexiglass, dont les arrêtes ont été limées, percé d'un trou servant à l'attache d'un fil de nylon pour éviter les risques de déglutition ; le fil reste à l'extérieur de la bouche lors de la production sans qu'il perturbe la fermeture des lèvres) et dans une condition

[1] Nous appellerons « tempo rapide » et « tempo lent » les conditions de, respectivement, débit rapide instruit et débit lent instruit.

où on leur demandait d'adopter un tempo lent. Les détails concernant les réglages des conditions sont décrits dans le Table 1.

Condition	**Précisions**
Bite-block	10*15*5mm
MAF	Bruit de « cocktail party » +12dB
Tempo	Guidé par curseur à 1,4 syll/sec (lent) ou 6syll/sec (rapide)

TABLE 1 : Valeurs des perturbations pour les conditions utilisées

1.	La pita d'Arabelle était sirupeuse.
2.	Le papi de Taschri loue des skis de fond.
3.	Les Beschki invitaient souvent tata Louise.
4.	Le ch'ti, c'était comme le chinois pour p**a**pa.
5.	Le bâti parisien était très coûteux.

TABLE 2: Corpus de phrases à lire

Dans chaque condition, il a été demandé aux locutrices de lire une phrase apparaissant à l'écran. Un total de cinq phrases a été utilisé (TABLE *2*). Chaque phrase était lue à 20 reprises par la locutrice avant de passer à une nouvelle phrase. Les phrases apparaissaient dans un ordre aléatoire, si bien que leur ordre changeait à chaque condition et pour chaque locutrice. Au total, chaque locutrice a donc lu 900 phrases (5 phrases * 20 répétitions par phrase * 9 conditions).

Les données ont été segmentées semi-automatiquement utilisant EasyAlign (Goldman, 2011) avant une vérification manuelle des segmentations. Le débit en syllabes par seconde a ensuite été mesuré pour chaque phrase. Cette mesure a notamment permis de tester l'effet de la condition / de la perturbation sur la vitesse d'articulation, le MAF étant connu pour induire des variations de débit. Nous avons ensuite mesuré la fréquence de F1 des premiers /a/ de « p**a**pa » (cf. phrase 4 du corpus). Cette mesure a été prise sur le milieu de la voyelle et représente une moyenne des mesures à 40%, 50% et 60% de la durée totale de la voyelle. Les données sur le F1 ont été retenues afin de comparer les caractéristiques spectrales de la voyelle liées à l'aperture, notamment du fait de l'utilisation du *bite-block*. De plus, selon Kirchhuebel (2010), le MAF fait augmenter le F1 des voyelles. L'effet de chacune des conditions (baseline, MAF, BB, tempo lent, tempo rapide) sur (1) le débit et (2) le F1 du /a/ a été testé dans R (R Core Team, 2017) en deux étapes. D'abord, l'effet global de la condition (5 niveaux) a été testé avec le package *lme4* (Bates et al., 2019) en utilisant un modèle linéaire mixte. Les variations liées aux locuteurs sont modélisées par une structure aléatoire par locuteur. Dans un deuxième temps, des comparaisons par paires avec le package *emmeans* (Lenth et al., 2020) ont été effectuées par condition et ce, pour chaque locuteur et avec correction de Bonferroni pour comparaisons multiples. Enfin, nous avons calculé, pour le débit et pour le F1, le coefficient de variation (ET / moyenne) des 20 répétitions par locutrice et par condition.

3 Résultats

3.1 Débit de parole

Pour le groupe entier, la condition montre un effet significatif sur le débit ($p < 0,001$). Cet effet se traduit par une augmentation de vitesse d'articulation en BB ($t = 5,347$; $p < 0,001$), tempo rapide

(t = 35,68 ; p < 0,001) et MAF (t = 12,053 ; p < 0,001) et une baisse de cette même vitesse en tempo lent (t = -116,793 ; p < 0,001).

Les comparaisons par paires montrent un effet des conditions sur le débit de la parole chez toutes les locutrices (Figure 1). La locutrice S01 augmente sa vitesse d'articulation de manière significative en condition BB (t = -3,378 ; p = 0,008), MAF (t = -3,993 ; p = 0,0008) et en tâche de tempo rapide (t = -34,849 ; p < 0,001) par rapport à la baseline. Elle abaisse son débit en tempo lent comparé à la baseline (t = 54,642 ; p < 0,001). La locutrice S02 augmente son débit en situation de MAF (t = -3,015 ; p = 0,02) et tempo rapide (t = -4,773 ; p < 0,001). Son débit est plus lent que la baseline en tempo lent (t = 81,917 ; p < 0,001). La vitesse d'articulation de la locutrice S03 est significativement plus rapide en BB (t = -7,826 ; p < 0,001), MAF (t = -18,444 ; p < 0,0001) et tempo rapide (t = -28,030 ; p < 0,001) qu'en baseline. Son débit est plus lent en tempo lent (t = 47,343 ; p < 0,001). Quant à S04, on observe une augmentation de débit en MAF (t = -4,175 ; p = 0,0004) et en tempo rapide (t = -18,428 ; p < 0,001), ainsi qu'une baisse de débit en tempo lent (t = 71,325 ; p < 0,001) par rapport à la baseline. Enfin, chez S05, le BB (t = -5,245 ; p < 0,001), le MAF (t = -4,46 ; p = 0,001) et le tempo rapide (t = -13,539 ; p < 0,001) font augmenter sa vitesse d'articulation comparé à la baseline et le tempo lent le fait baisser (t = 66,844 ; p < 0,001).

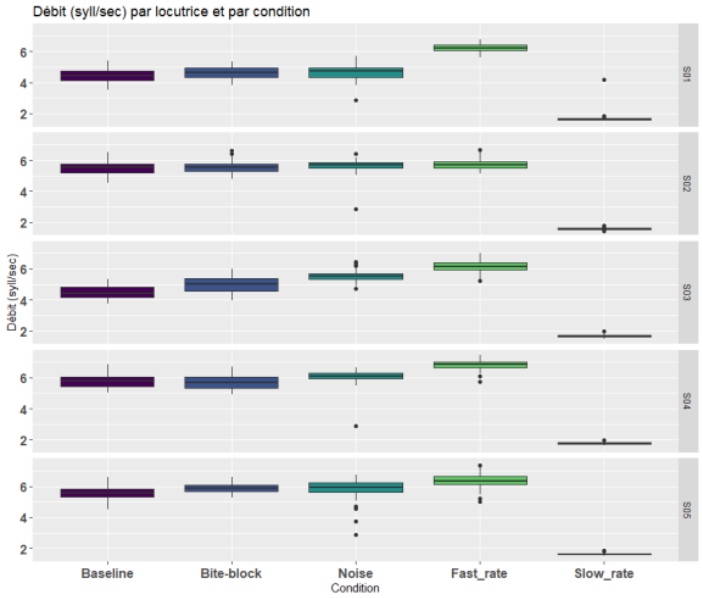


FIGURE 1: Débit d'articulation (syll/sec) par locuteur et par condition (noise = MAF, fast_rate = tempo rapide, slow_rate = tempo lent).

Nous nous sommes ensuite intéressés aux effets des conditions sur la variabilité du débit (FIGURE *2*) en calculant le coefficient de variation (écart type / moyenne), dans l'objectif d'étudier la stabilité des réactions aux conditions. Les résultats montrent que la vitesse d'articulation est plutôt stable à travers les conditions pour la plupart des locutrices. Il est toutefois à noter que la locutrice S03 – et elle seule – présente davantage de variation en condition de tempo rapide par rapport à la baseline (p = 0,04),

par rapport au MAF (p = 0,01) et par rapport au BB (p = 0,005). De même, sa vitesse d'articulation est plus variable en tempo lent par rapport au BB (p = 0,01).

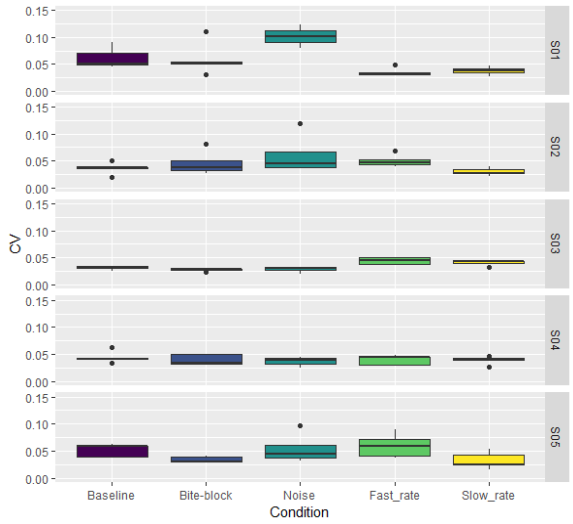


FIGURE 2: Coefficient de variation (ET / moyenne) de la vitesse d'articulation (syll/sec) par condition et par locuteur

3.2 Premier formant (F1)

La condition montre un effet sur le F1 (en Hz) (p < 0,001), avec une augmentation de F1 dans toutes les conditions : en BB (t = 5,429 ; p < 0,001), MAF (t = 3,735 ; p < 0,001), tempo rapide (t = 3,502 ; p < 0,001), et en tempo lent (t = 16,831 ; p < 0,001).

Les réactions individuelles sont illustrées sur la FIGURE 3. Le sujet S01 augmente son F1 en BB (t = -5,431 ; p < 0,001), MAF (t = -3984 ; p = 0,0011), tempo rapide (t = -4,383 ; p < 0,001) et tempo lent (t = -15,111 ; p < 0,001) comparé à la baseline. La locutrice S02 modifie son F1 en tempo lent uniquement par rapport à la baseline (t = -9,245 ; p < 0,001) et ce, en l'augmentant. Chez S03, une augmentation du F1 est observée en BB (t = 12,509 ; p < 0,001), en MAF (t = -8,263 ; p < 0,001), en tempo rapide (t = -9,544 ; p < 0,001) et en tempo lent (t = -26,257 ; p < 0,001). S04 augmente le F1 en BB (t = -4,547 ; p < 0,001), en MAF (t = -5,079 ; p < 0,001) et en tempo lent (t = -21,363 ; p < 0,001) par rapport à la baseline. Enfin, la locutrice S05 ne modifie son F1 de manière significative dans aucune condition.

L'examen du coefficient de variabilité du F1 (en Hz) entre toutes les répétitions par condition renforce les différences individuelles, dans la mesure où la variabilité intra-locuteur n'est pas systématiquement la plus élevée dans une tâche particulière mais change en fonction de la locutrice. Ainsi, tandis que S01 et S04 sont les plus variables dans la baseline, chez S02 et S05 c'est surtout dans la condition de tempo lent qu'elles sont variables, avec également une variabilité plus importante dans les conditions BB et MAF pour S05. Enfin, S03 présente une certaine stabilité à travers les conditions (FIGURE *4*).

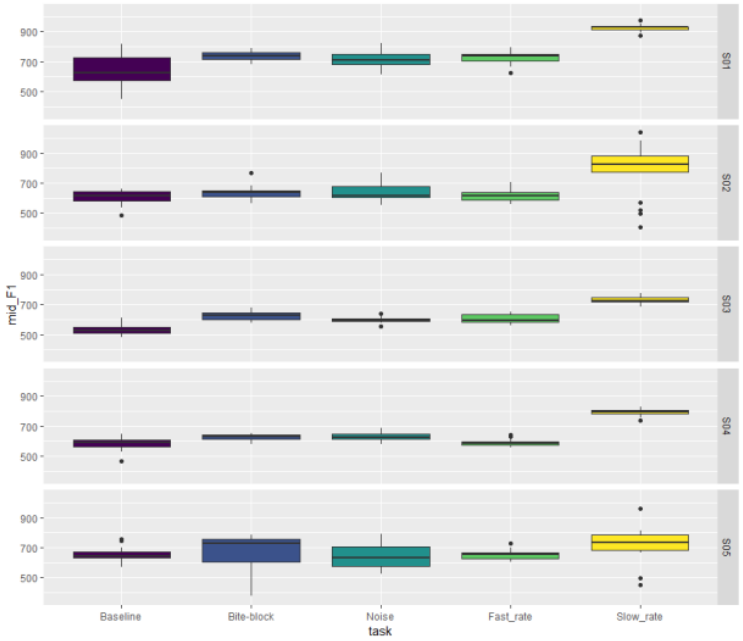


FIGURE 3 : Valeur moyenne de F1 (moyenne entre 40%, 50% et 60% de la durée totale de la voyelle) par locutrice et par condition

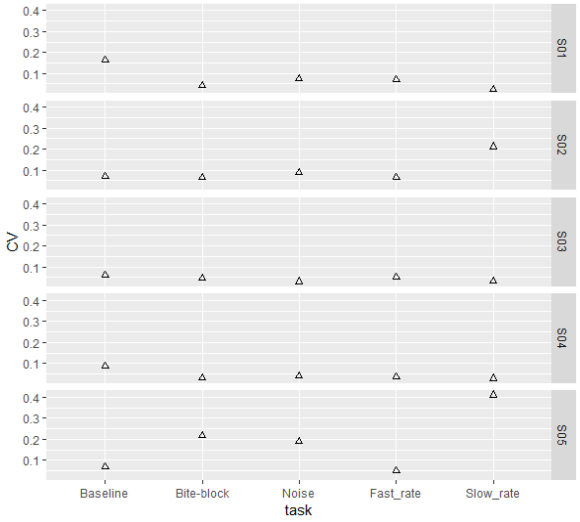


FIGURE 4: Coefficient de variation (ET / moyenne) du F1 par condition et par locutrice

4 Discussion

Nous avons comparé la production de la parole dans différentes conditions d'élocution : lorsque le débit de parole est imposé, lorsque le retour auditif est masqué et lorsque des contraintes articulatoires sont modifiées par la présence d'un bite-block affectant la configuration de la cavité buccale.

L'objectif de l'étude était d'examiner les réactions à ces perturbations en termes de débit de parole et du premier formant, ainsi que les variations intra- et inter-individuelles dans ces réactions.

Nos résultats montrent que les réactions aux modifications ne sont pas particulièrement variables quant au débit de parole : toutes les locutrices ont réagi à l'absence de retour auditif par une augmentation de leur vitesse d'articulation et, à l'exception de deux d'entre elles, elles ont également augmenté leur vitesse d'articulation en parole avec le BB. Aucune locutrice n'a diminué son débit durant ces deux perturbations. L'augmentation du débit de parole sous le MAF est un résultat intéressant qui pourrait s'expliquer en partie par le fait que les locutrices tentent de mettre terme le plus rapidement possible aux perturbations, inconfortables, en augmentant leur vitesse d'articulation (Jacks & Haley, 2015). Quant à la modification des caractéristiques spectrales, les résultats font état d'une stratégie variable. Tandis que trois locutrices augmentent leur F1 en parole avec BB, en situation de MAF et avec le tempo modifié par rapport à la baseline, une locutrice ne modifie jamais son F1 et une dernière ne l'augmente qu'en situation de tempo lent. Donc, si modification il y a, elle va dans le sens de l'augmentation du F1 car aucune locutrice n'abaisse ce formant de manière significative en condition de perturbation. Dans le cas du BB, la variété des comportements observés peut s'expliquer en faisant l'hypothèse que les deux locutrices qui ne montrent pas de différences par rapport à la condition de baseline aient prêté davantage d'attention au retour somatosensoriel (et que, par conséquent, elles aient compensé la perturbation de la forme de la cavité buccale), tandis que les locutrices qui ont montré une augmentation du F1 auraient prêté davantage d'attention au retour acoustique. En effet, le bite block augmente l'aperture de la cavité buccale. Du moment qu'une aperture plus importante ne rend pas le /a/ moins distinct des autres phonèmes, ces locutrices n'ont pas de raison de compenser cette modification. Concernant le MAF, ce dernier induit une augmentation du F1 chez 3 locutrices également. Cette réaction est conforme aux résultats obtenus dans la littérature (Kirchhuebel, 2010). Ce résultat, combiné à l'augmentation de débit, indique que les locutrices en question décorrèlent le débit et l'aperture vocalique, dans la mesure où l'on pourrait s'attendre à ce que le F1 baisse dans un souci d'économie de geste lorsque le débit augmente. Or, ce n'est pas ce que l'on observe ici et ce, aussi bien avec le MAF qu'en tempo rapide sans autre perturbation. On peut en déduire qu'une réduction de la vitesse d'élocution en réponse au MAF (cf. par ex. Garnier et al., 2006) a une fonction communicative et qu'elle n'est pas une simple conséquence d'une augmentation de la précision articulatoire / de l'hyperarticulation des voyelles. De manière intéressante, les locutrices ne modifiant pas leur F1 en MAF ne le modifient pas en BB non plus. Cela pourrait être interprété en support de l'hypothèse que le biais vers le retour somatosensoriel ou vers le retour acoustique est indépendant du type de perturbation. Bien évidemment, cette hypothèse doit être confirmée en comparant les résultats obtenus avec ceux issus de l'analyse des réponses à d'autres types de perturbation.

Remerciements

Cette étude a été financée par le projet MoSpeeDi - CRSII5_173711/1 du Fond National Suisse de la Recherche Scientifique et par le programme "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL).

Références

BATES, D. M., MAECHLER, M., BOLKER, B., WALKER, S., CHRISTENSEN, R. H. B., SINGMANN, H., DAI, B., SCHEIPL, F., GROTHENDIECK, G., GREEN, P., & FOX, J. (2019). *Package « lme4 »* (1.1-21) [R; R].

BURNETT, T. A., FREEDLAND, M. B., LARSON, C. R., & HAIN, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, *103*(6), 3153-3161. DOI : 10.1121/1.423073

FOWLER, C., & TURVEY, M. (1981). Immediate compensation in bite-block speech. *Phonetica*, *37*(5-6), 306-326.

GARNIER, M., BAILLY, L., DOHEN, M., WELBY, P., & LŒVENBRUCK, H. (2006). An Acoustic and Articulatory Study of Lombard Speech : Global Effects on the Utterance. *INTERSPEECH-2006*, 2246-2249.

GARNIER, M., HENRICH, N., & DUBOIS, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language, and Hearing Research*, *53*(3), 588-608. https://doi.org/10.1044/1092-4388(2009/08-0138).

GOLDMAN, J.-P. (2011). *EasyAlign : An automatic phonetic alignment tool under Praat*.

GUENTHER, F. H. (2016). *Neural Control of Speech*. The MIT Press.

HOOLE, P. (1987). *Bite-block speech in the absence of oral sensibility*. *4*, 16-19.

ITO, T., CAILLET, J.-L., & PERRIER, P. (2019). Posture stabilization of the tongue for speech : Responses to mechanical perturbation. *Proceedings of the 19th International Congress of Phonetic Sciences*, 1838-1842.

JACKS, A., & HALEY, K. L. (2015). Auditory Masking Effects on Speech Fluency in Apraxia of Speech and Aphasia : Comparison to Altered Auditory Feedback. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1670-1686. DOI : 10.1044/2015_JSLHR-S-14-0277

JONES, J. A., & MUNHALL, K. G. (2000). Perceptual calibration of F0 production : Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246. DOI :10.1121/1.1288414

KIRCHHUEBEL, C. (2010). The effects of Lombard speech on vowel formant measurements. *The Journal of the Acoustical Society of America*, *128*(4), 2395-.

LARSON, C. R., ALTMAN, K. W., LIU, H., & HAIN, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F 0 control. *Experimental Brain Research*, *187*(4), 613-621. DOI : 10.1007/s00221-008-1330-z

LENTH, R., SINGMANN, H., LOVE, J., BUERKNER, P., & HERVE, M. (2020). *Package « emmeans »* (Version 1.4.4) [R].

MACDONALD, E. N., PURCELL, D. W., & MUNHALL, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *The Journal of the Acoustical Society of America*, *129*(2), 955-965. DOI : 10.1121/1.3531932

MALLARD, A., RINGEL, R., & HORII, Y. (1978). Sensory contributions to control of fundamental frequency of phonation. *Folia Phoniatrica*, *30*, 199-213.

MARUTA, C., MAKHMOOD, S., DOWNEY, L. E., GOLDEN, H. L., FLETCHER, P. D., WITOONPANICH, P., ROHRER, J. D., & WARREN, J. D. (2014). Delayed auditory feedback simulates features of nonfluent primary progressive aphasia. *Journal of the Neurological Sciences*, *347*(1-2), 345-348. DOI : 10.1016/j.jns.2014.09.039

MÉNARD, L., PERRIER, P., & AUBIN, J. (2016). Compensation for a lip-tube perturbation in 4-year-olds : Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *The Journal of the Acoustical Society of America*, *139*(5), 2514-2531.

R CORE TEAM. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Perception des consonnes dans la dysarthrie parkinsonienne : effets du contexte phonémique, prosodique et lexical

Danielle Duez [1] , Alain Ghio [1], François Viallet [1, 2]

(1) Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(2) Centre Hospitalier du Pays d'Aix, Service de Neurologie, Aix-en-Provence, France

danielle.duez@lpl-aix.fr, alain.ghio@lpl-aix.fr, fviallet@ch-aix.fr

RÉSUMÉ

Les patients atteints de la maladie de Parkinson (MDP) présentent généralement des déficits dans la production de la parole. Nous avons examiné l'identification perceptive des consonnes intervocaliques produites par 10 locuteurs avec MDP et 10 locuteurs sains en lecture de texte. Pour neutraliser le contenu sémantique, toutes les consonnes intervocaliques ont été isolées avec la moitié des voyelles précédente et suivante.
20 adultes natifs francophones ont été chargés de transcrire les séquences du corpus. La consonne rapportée a été examinée par rapport à la consonne prototypique; le score de distorsion est le nombre de traits phonétiques différents par rapport à la consonne prototypique. Les résultats ont été examinés en fonction des facteurs linguistiques suivants: nature de la consonne, contexte vocalique oral / nasal, classe de mot (fonction ou contenu) et position dans les syntagmes.
L'imprécision de la consonne a été confirmée dans la parole des locuteurs MDP.

MOTS-CLÉS : perception, dysarthrie parkinsonienne, phonétique clinique, traits phonétiques

ABSTRACT

Perception of consonants in parkinsonian dysarthria: effects of the phonetic, prosodic and lexical context

Patients with Parkinson's Diseases (PD) typically exhibit deficits in speech production. We examined the perceptual identification of intervocalic consonants produced by 10 speakers with PD and 10 healthy speakers reading a text. To neutralize the semantic effect, all the intervocalic consonants were excised with half the preceding and following vowels.

20 adults native speakers of French were instructed to transcribe the sequences they heard. The reported consonant was examined in relation to the expected consonant ; the score of distorsion was the number of phonetic features differing from the prototypical consonant. The results were examined as a function of the following/or preceding linguistic factors : consonant nature, oral/nasal vocalic context, class of word (function or content) and position within sentences. Consonant imprecision was confirmed in the speech of PD speakers.

KEYWORDS: perception, parkinsonian dysarthria, clinical phonetics, phonetic features

Volume 1 : Journées d'Études sur la Parole, pages 172–180.
HAL : hal-02798533.

1 Introduction

1.1 Etat de l'art

La maladie de Parkinson (MDP) est caractérisée par une perte progressive de neurones dopaminergiques au sein de la substantia negra (pars compacta); ses manifestations externes sont des déficits de mouvement comprenant la rigidité ou la raideur (muscles résistants au mouvement), l'akinésie (incapacité à initier le mouvement), la bradykinésie (lenteur du mouvement) et le tremblement de repos. Dans la production de la parole, ces symptômes affectent la respiration, la phonation, la prosodie et l'articulation (Darley et al., 1969), ce qui se manifeste au niveau acoustique par des changements et des anomalies de la fréquence fondamentale (F0), de l'intensité, du débit de parole, de la durée/distribution des pauses et de l'imprécision consonantique.

L'imprécision des consonnes dans la MDP a été examinée dans un certain nombre d'études perceptives, acoustiques et physiologiques. Par exemple, dans une étude utilisant le test de Fisher-Logemann sur la compétence articulatoire, les auteurs ont examiné la transcription phonétique effectuée par deux experts de parole produite par 200 patients atteints de la MDP et ont observé que les distorsions prédominantes touchent les occlusives, les affriquées et les fricatives (Logemann et al., 1981). En termes de traits phonétiques, les occlusives et les affriquées, qui sont normalement [-continu] ont été produites comme [+ continu] ; les fricatives qui sont [+ stridentes] ont été produites comme [-stridentes]. Une imprécision de l'articulation des consonnes a été également rapportée perceptivement dans (Chenery et al., 1988), dans (Ho et al., 1999) et dans (Plowman et al., 2009).

Ces résultats perceptifs ont été confirmés par des études acoustiques, notamment sur les imprécisions de voisement dans les travaux de Kent et Rosenbek (1982) ou Weismer (1984) pour l'anglais. Des anomalies similaires ont été signalées dans des études sur le français (Gremy, 1958 ; Uziel et al., 1975 ; Duez, 2014). Comme il est suggéré par Kent et Netsell (1971), la persévérance de voisement peut représenter un comportement compensatoire qui permet au locuteur d'éviter d'initier et arrêter des gestes articulatoires difficiles. La spirantisation s'est également révélée être une caractéristique saillante de l'imprécision des consonnes dans la parole parkinsonienne, c'est-à-dire le remplacement d'une tenue d'occlusive par une frication de faible intensité, ce qui reflète l'échec de la fermeture orale complète (Weismer, 1984 ; Kent et al., 1982). Dans une étude acoustique de la parole produite par 12 patients allemands, Ackermann et Ziegler (1992) ont observé une capacité réduite à fermer complètement le conduit vocal dans le cas d'occlusives et ont interprété cela comme une réduction de l'amplitude de mouvement des articulateurs. Fait intéressant, cet «undershoot» articulatoire n'était pas uniforme mais influencé par les exigences linguistiques, les occlusions associées aux syllabes accentuées étant effectuées beaucoup mieux que dans les syllabes non accentuées.

1.2 Objectifs de l'étude

Notre étude a pour objectif d'examiner la perception des consonnes intervocaliques contenues dans un paragraphe lu par des locuteurs MDP et des locuteurs issus d'un groupe témoin. Choisir d'utiliser de la parole lue nous permet d'examiner une grande variété de consonnes produites dans divers contextes phonémiques, dans différents types de mots et dans différentes positions prosodiques. Pour neutraliser l'effet de restauration qui aide à rétablir les informations manquantes dans les consonnes imprécises ou réduites (Warren et al., 1970 ; Duez, 2001), toutes les consonnes « stimuli » ont été isolées de façon à ne pas distinguer le mot dans lequel elles apparaissent et éviter de reconnaitre la consonne par identification lexicale. Mais pour éviter une écoute trop courte sur l'unique segment consonantique, nous avons découpé la consonne en ajoutant la moitié des voyelles précédentes et suivantes. Notre objectif est double:

1) comparer chaque consonne transcrite avec chaque consonne attendue dans divers contextes, à la fois dans la parole parkinsonienne et dans le groupe contrôle. Nous estimons que la comparaison des caractéristiques perçues dans les deux groupes nous permettra de séparer les changements résultant de la réduction et des processus d'assimilation normaux de ceux dus aux effets de la maladie de Parkinson.

2) tester la procédure et voir dans quelle mesure les séquences VCV peuvent être utilisées dans l'analyse de l'intelligibilité de la parole pathologique.

2 Matériel et méthode

2.1 Locuteurs

Dans cette étude, nous avons sélectionné les locuteurs à partir de la base de données AHN du service de neurologie du Centre Hospitalier du Pays d'Aix (Ghio et al., 2012). Nous avons choisi 10 hommes diagnostiqués avec la maladie de Parkinson et 10 locuteurs témoins de même sexe et âge que les patients. Les patients sélectionnés avaient entre 6 et 26 ans de maladie (moy=13.1 ans). Ils étaient tous traités par L-dopa de façon usuelle mais pour observer de façon plus nette les effets de la MDP, tous les patients avaient été sevrés de médicament pendant plus de 12 heures, délai usuel pour annuler les effets pharmacologiques. Avant l'enregistrement, l'évaluation motrice de chaque patient a été évaluée par un neurologue à l'aide de l'échelle UPDRS (Unified Parkinson's Disease Rating Score). L'item n°18 de cette échelle est particulièrement informatif car il indique la sévérité de la dysarthrie dans une approche subjective clinique avec les conventions suivantes: 0 ⇔ normal; 1⇔ légère perte d'expression, de diction et / ou de volume; 2⇔ monotone, flou, mais compréhensible, modérément altéré; 3⇔ déficience marquée, difficile à comprendre; 4⇔ inintelligible. Les caractéristiques de chaque patient sont indiquées en Table 1.

Table 1 : Caractéristiques des patients

	Age du diagnostic	Durée de la maladie	UPDRS III	Sévérité de la dysarthrie
P1	48	20	61	3
P2	45	12	34	2
P3	59	6	40	3
P4	31	13	30	1
P5	48	26	53	3
P6	39	11	30	3
P7	45	8	42	2
P8	52	8	44	1
P9	54	15	40	1
P10	55	11	35	1

2.2 Corpus

Le corpus utilisé pour fabriquer les stimuli est un paragraphe de « La chèvre de Monsieur Seguin » lu par les 20 locuteurs sélectionnés. La consigne était une lecture à vitesse et intensité confortable. Les enregistrements ont été réalisés dans une salle insonorisée du service de neurologie du Centre Hospitalier du Pays d'Aix à Aix-en-Provence par le biais d'un microphone serre-tête AKG C420 connecté au dispositif EVA2 (Ghio et al.,2012). La distribution des consonnes utilisées comme stimuli est détaillée en Table 2. Une moyenne de 114 séquences VCV par locuteur a été obtenue. Par conséquent, notre corpus était composé de 2280 stimuli (20 locuteurs * 114).

Table 2 : Nombre d'occurrences des consonnes dans le corpus

occlusives sourdes			occlusives voisées			Fricatives sourdes			fricatives voisées			Sonorantes				
p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	r	Cons
77	110	77	67	72	30	30	112	68	28	62	45	57	85	187	71	Control
75	104	75	67	73	28	29	93	69	22	54	44	54	78	172	54	MDP

2.3 Auditeurs

Les 2280 stimuli ont été divisés aléatoirement en 20 blocs. Un bloc était composé de 114 éléments qui pouvaient provenir de tous les locuteurs. Ces stimuli ont été soumis à 20 auditeurs francophones naïfs (non spécialistes dans l'écoute des troubles de la parole) recrutés pour l'expérience. La consigne était la suivante : « Vous allez entendre des extraits de parole assez courts (ex: ati). En respectant les règles de l'orthographe du français, vous devrez transcrire ce que vous entendrez. Certaines prononciations seront difficiles à identifier mais dans tous les cas, vous devrez proposer une transcription ». Ces tests ont eu lieu au Centre d'Expérimentation sur la parole (http://cep.lpl-aix.fr/) au Laboratoire Parole et Langage à Aix-en-Provence. La présentation des stimuli et le recueil des réponses étaient automatisés grâce au dispositif Perceval-Lancelot (André et al., 2003).

Chaque auditeur, portant un casque audiophonique Superlux HD 681B, a transcrit 3 blocs de 114 éléments, soit 342 stimuli. L'intensité de lecture du son a été préréglée par l'auditeur pour être confortable et optimale pour la tâche. Chaque test a commencé avec quatre stimuli d'entraînement. Chaque élément était présenté une fois automatiquement mais l'auditeur pouvait répéter la lecture deux fois. L'auditeur a eu une pause de 5 minutes entre les blocs. Un total de 6840 réponses a été collectée car chaque bloc a été soumis à 3 auditeurs différents.

2.4 Prétraitement des données

Une fois les transcriptions recueillies, les réponses orthographiques ont été analysées manuellement afin d'obtenir la structure phonotactique de la réponse et surtout d'identifier la consonne perçue. En termes de structure, 67% des VCV ont été rapportés comme VCV, 13% des VCV ont été signalés comme CV, 2% comme VCCV (dans ce cas, le / d / ou / t / a été palatalisé dz ou ts) et 2% comme VCVC. Dans certains cas, les auditeurs percevaient les VCV comme des mots. Par exemple, la séquence [eʃɛ] était signalée comme [eʃɛl] ("échelle"), la séquence [ãʒe] comme [mãʒe] ("manger "). Dans d'autres cas, les auditeurs n'ont pas pu identifier la consonne et n'ont rien écrit.

Nous n'avons pas analysé la transcription des voyelles. Nous nous sommes concentrés uniquement sur les consonnes. La forme orthographique a été simplifiée en tant que forme phonétique. Par exemple, la séquence orthographique "ph" a été phonétisée comme / f /, "g" + "e" a été phonétisée comme / ʒ /. La conversion graphème/phonème en français étant cohérente, nous ne gardons que les données où il n'y avait pas d'ambiguïté (nous effaçons seulement 18 réponses ambiguës sur 6819).

La consonne transcrite a été examinée par rapport à la consonne cible. Nous appelons « score de déviation phonétique perçue » (Perceived Phonological Deviation) le nombre de traits phonétiques qui diffèrent de la consonne prototypique à la réponse. Un score de 0 signifie que la consonne a été correctement identifiée. Un score de *N* signifie qu'il y avait *N* traits phonétiques mal identifiées. La décomposition en traits que nous avons choisie est celle publiée dans (Ghio et al., 2018). Ainsi, une consonne cible /p/ perçue /b/ induira un score PPD de 1 (trait de voisement) ; un /p/ perçu /m/ fournira un score PPD de 2 (voisement + nasalité). Plus le score PPD est faible, meilleure peut être considérée l'intelligibilité (taux d'identification des consonnes) du locuteur.

3 Résultats

Le score PPD d'un locuteur est la moyenne des scores obtenue sur les 114 séquences VCV produite en moyenne par chaque locuteur et évaluées perceptivement. Tous les tests statistiques ont été effectués dans l'environnement logiciel R version 3.4.4 Des modèles linéaires à effets mixtes (package lme) ont été utilisés pour analyser les scores PPD considérés comme une variable continue.

3.1 Résultats généraux

Comme prévu, le score PPD est significativement plus faible (p <0,01) pour les locuteurs du groupe contrôle (CTRL) que pour les locuteurs parkinsoniens (PARK). Le score moyen pour les CTRL est de 0.72 avec un écart-type de 0.21, les valeurs correspondantes pour les PARK étant de 1.18 et 0.42 d'écart-type (Figure 1).

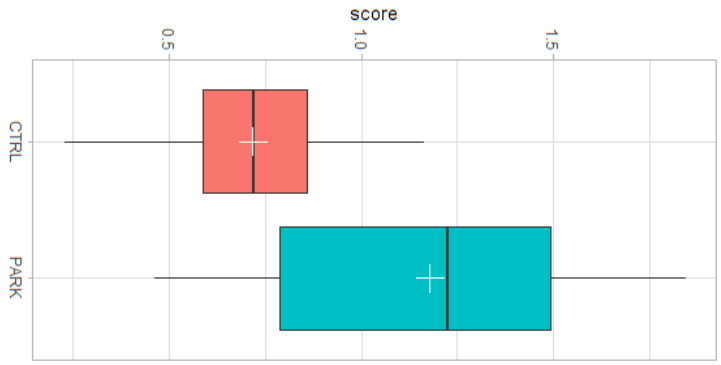


Figure 1 : score PPD (perceived phonological deviation) global pour les sujets contrôle (CTRL) vs parkinsoniens (PARK))

3.2 Les résultats par locuteur

Comme on peut le voir sur la Figure 2, les scores PPD obtenus pour les locuteurs témoins CTRL sont regroupés et inférieurs à 1; en revanche, les patients PARK se divisent en deux groupes. Les quatre premiers patients ont un score proche de celui des témoins (inférieur à 1) tandis que les six derniers patients ont un score PPD élevé (supérieur à 1,25). Fait intéressant, il existe un lien entre le premier groupe qui correspond aux patients avec un score de sévérité de la dysarthrie égal à 1 (voir Table 1) alors que ceux dont le score PPD est supérieur à 1.25 correspond aux patients avec un score de sévérité clinique de 2 ou 3 (voir Table 1).

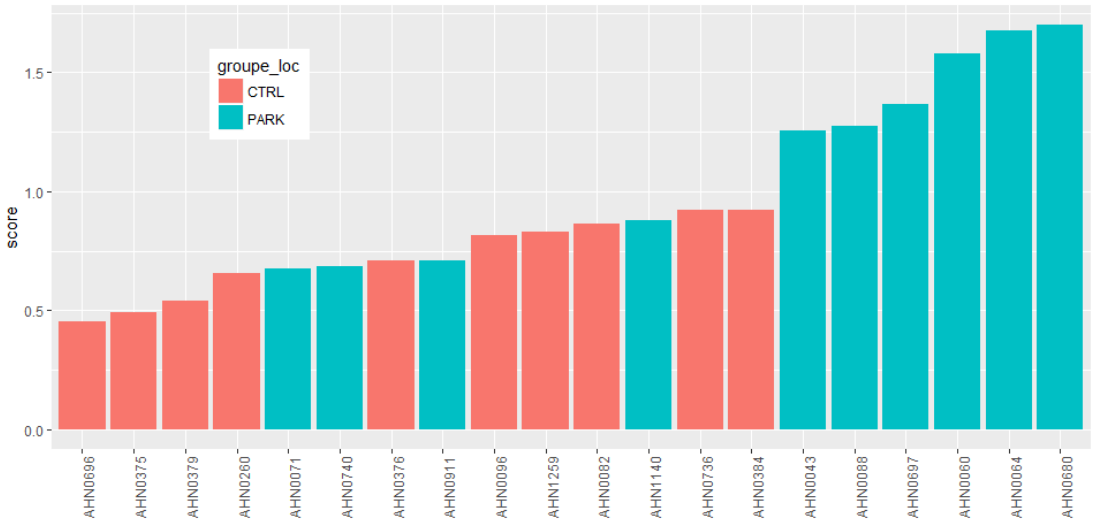


Figure 2 : score PPD pour chaque locuteur

3.3 Les résultats par consonnes cible

Comme le montre la Figure 3, il existe une hétérogénéité entre les consonnes. Les consonnes sourdes telles que /t, s, ʃ/ ont le meilleur taux d'identification tandis que les sonorantes /m, n, l, r/ ont des scores de déviation importants. La labiovelaire /v/ et la sonorante /r/ ont les taux d'identification les plus bas. Si les patients ont des scores de distorsion supérieurs à ceux des témoins, il est intéressant de noter que cette distribution est à peu près la même pour les témoins et les patients.

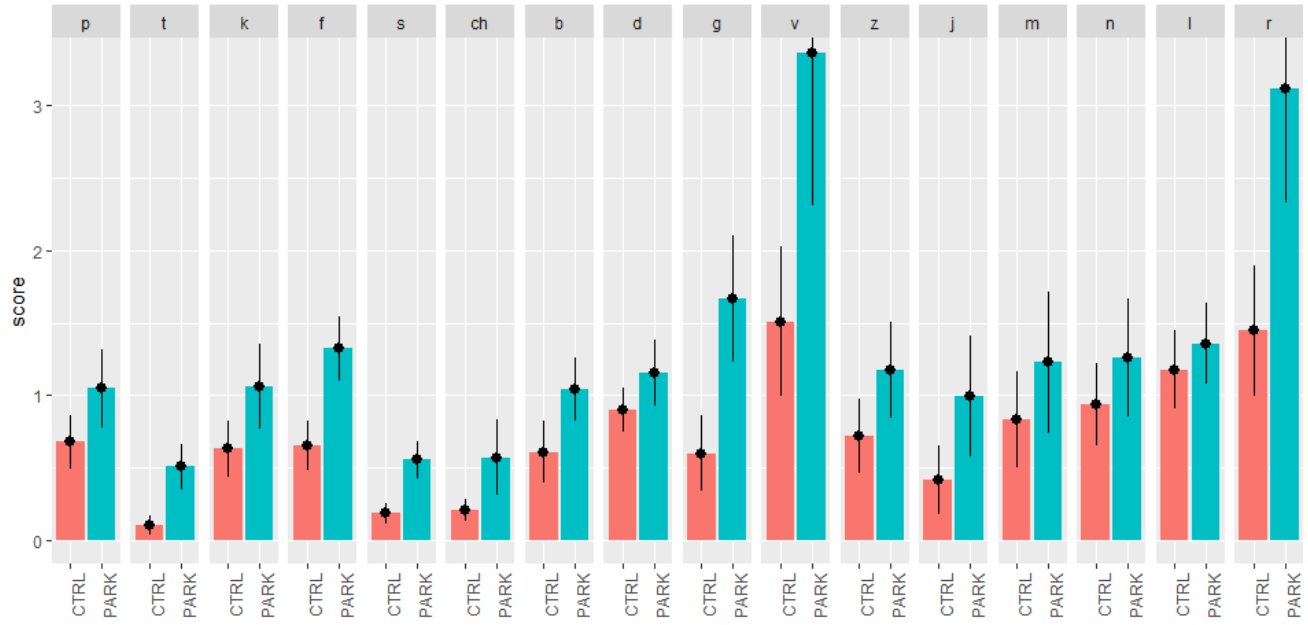


Figure 3 : score PPD en fonction de chaque consonne cible

3.4 L'effet du contexte prosodique

La Figure 4 présente les scores PPD obtenus en fonction de la position prosodique des consonnes dans l'énoncé. On observe que la plus mauvaise identification concerne les consonnes situées en position interne (NF) du groupe prosodique que ce soit pour le groupe témoin ($\mu = 0{,}92$; sd = 0,30) et pour le groupe de patients ($\mu = 1{,}35$, sd = 0,48). Lorsque les consonnes sont dans les syllabes finales (FF ou FP), les scores sont meilleurs en cas de pause finale (FP ; CTRL: $\mu = 0{,}41$, sd = 0,30; PD: $\mu = 0{,}82$, sd = 0,53) qu'en l'absence de pause (FF ; pour CTRL $\mu = 0{,}51$, sd = 0,36; pour PD $\mu = 1{,}01$, sd = 0,59). Les scores obtenus pour les consonnes en position initiale (IP) sont intermédiaires pour les témoins ($\mu = 0{,}62$, sd = 0,32) et les patients ($\mu = 1{,}16$, sd = 0,60).

Pour résumer, deux tendances émergent des résultats actuels: 1) les consonnes situées en position de frontière (IP, FF, FP) ont un score d'identification significativement plus élevé ($p < 0{,}0001$) que les consonnes situées dans les phrases et 2) les scores d'identification sont significativement plus élevés pour le groupe témoin que pour le groupe de patients ($p = 0{,}0035$). Il n'y a pas d'interaction ($p = 0.7653$).

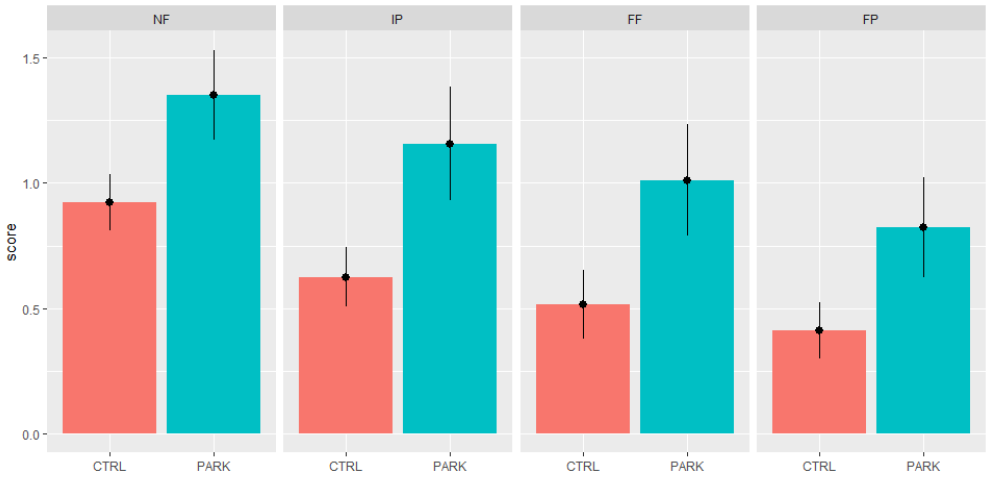


Figure 4 : score PPD en fonction de la position prosodique de la consonne cible

3.5 L'effet du contexte lié à la classe du mot contenant la consonne

Les consonnes situées dans les mots lexicaux (CW) ont un score d'identification significativement meilleur que celles situées dans les mots grammaticaux (FW) dans les deux groupes (p = 0,0003). Dans le groupe témoin, les scores PPD moyens sont plus faibles pour les consonnes situées dans les mots lexicaux (μ = 0,65, sd = 0,20) que pour celles situées dans les mots grammaticaux (μ = 0,86, sd = 0,35). Dans le groupe de patients, les scores moyens sont significativement plus élevés (p = 0,0054), les scores correspondants pour les mots lexicaux et les mots grammaticaux sont respectivement de 1,28 (sd = 0,5) et 1,12 (sd = 0,45). Il n'y a pas d'interaction (p = 0,55)

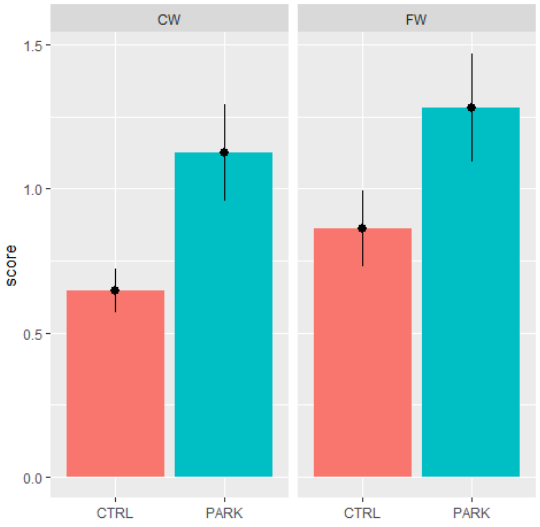


Figure 5 : score PPD en fonction de la classe du mot contenant la consonne

4 Discussion

Les résultats obtenus sont totalement conformes aux résultats rapportés dans la littérature sur l'imprécision des consonnes: les différences reflètent les problèmes articulatoires qui caractérisent la maladie de Parkinson. Le deuxième résultat est l'hétérogénéité des scores selon la consonne: certaines d'entre elles ont un score d'identification élevé, d'autres un score faible. Différentes hiérarchies ont été proposées avec une force consonantique en relation inverse avec l'échelle de sonorité, comme la hiérarchie élaborée suivante avec des occlusives et fricatives sourdes en tête suivies par les obstruantes voisées, les nasales les latérales et le R au plus bas.

Fait intéressant observé sur les locuteurs du groupe contrôle, certaines consonnes ont des caractéristiques mal identifiées. Il est bien connu que les locuteurs adaptent leurs gestes phonétiques de manière adaptative aux différents besoins des situations de parole et qu'ils maintiennent un contraste perceptif suffisant (c'est-à-dire un contraste suffisamment riche) pour être compris. Dans la présente étude, les consonnes ont été extraites de leur contexte et les informations conservées dans la

consonne peuvent ne pas avoir été suffisamment riches pour l'identification de toutes les caractéristiques. De plus, il a été démontré que la coarticulation est fréquente dans les séquences VCV et que de nombreux locuteurs témoins sains spirantisent parfois les occlusives. Par conséquent, on peut supposer que certaines caractéristiques mal identifiées sont le résultat d'une assimilation contextuelle. Concernant la parole MPD, il y a probablement des anomalies dues à la rigidité des muscles ce qui conduit à une diminution de l'amplitude et / ou de la force des mouvements articulatoires.

Le 3ième point examiné était l'impact de la structure prosodique sur la perception des consonnes dans les deux groupes de locuteur. Il a été montré que les consonnes situées dans les syllabes à l'intérieur d'un groupe prosodique avaient un score d'identification significativement plus bas que les consonnes situées dans les syllabes finales pré-pausales ou pas, suggérant ainsi que les premières étaient produites moins clairement que les secondes. Il a été démontré que les patients atteints de MP produisent un allongement final normal de la même manière que les locuteurs sains. On peut supposer que, comme les locuteurs contrôle, les patients articulent plus clairement les segments des syllabes terminales qui sont des points clés de la structure prosodique et ont un rôle important dans la perception. Fait intéressant, un effet de la proéminence initiale sur l'identification des consonnes a également été observé. La proéminence initiale a été étudiée acoustiquement et s'est révélée être caractérisée par un allongement plus important de l'onset par rapport à la proéminence finale qui présente un noyau et une rime plus longue. Cela explique pourquoi les consonnes initiales des mots ont un meilleur score d'identification que les consonnes situées dans les phrases. Dans tous les cas, ces résultats indiquent que la fonction syntaxique et démarcative de la prosodie est maintenue chez les patients souffrant de la maladie de Parkinson.

L'impact de la classe de mots sur l'identification des consonnes a été le quatrième point principal examiné. Il a été constaté que dans le discours de contrôle sain ainsi que dans le discours parkinsonien, les consonnes avaient plus de caractéristiques identifiées dans les mots de contenu que dans les mots de fonction. Ceci est en accord avec les études sur la phonétique des mots de contenu et de fonction. Ce fait est intéressant car il suggère que les patients, tout comme les locuteurs sains de contrôle, ont tendance à conserver les informations sémantiques contenues dans les mots lexicaux.

5 Conclusion

La présente étude est basée sur la perception des consonnes extraites de la parole lue par dix hommes souffrant de la maladie de Parkinson et par dix témoins. L'imprécision des consonnes a été confirmée dans la production des locuteurs MDP. De plus, deux groupes de patients ont été observés: le premier avec un faible degré de sévérité de la dysarthrie et des scores d'identification des consonnes proches de ceux des témoins; le second groupe avec un degré moyen ou élevé de sévérité de la dysarthrie et un faible score d'identification. Il semble donc que le score PPD puisse être un indicateur du degré de gravité de la maladie. Des études longitudinales sur la parole parkinsonienne pourraient nous indiquer si la production de consonnes est sensible à la progression de la maladie et pourrait être utilisée comme un outil d'évaluation.

La méthode utilisée dans cette étude nous a permis de prendre en compte l'effet des facteurs linguistiques et de montrer comment ils affectent la production et la perception des consonnes. L'intelligibilité semble optimale lorsque les consonnes appartiennent à des mots de contenu, à des syllabes initiales de mots ou à des syllabes de phrases finales.

Remerciements

Cette recherche a été financée en partie par un BQR labo (LPL) et par l'Agence nationale de la recherche ANR-18-CE45-0008. Les auteurs tiennent à remercier Alain Purson et Ludovic Jankowksi pour leur assistance lors de l'enregistrement des patients ainsi que le personnel du CEP, Laura Reynaud et Carine André pour leur assistance dans l'expérience perceptive.

Références

ACKERMANN H AND ZIEGLER W (1992) Articulatory deficits in parkinsonian dysarthria: an acoustic analysis, J Neurol Neurosurgery Psychiatry, 54 (12) 1093-1098.

ANDRÉ C, GHIO A, CAVÉ C, TESTON B . (2003) PERCEVAL: a Computer-Driven Systemfor Experimentation on Auditory and Visual Perception. International Congress of Phonetic Sciences (ICPhS), Barcelona, Spain. pp.1421-1424.

CHENERY H, MURDOCH B. AND INGRAM J (1988) Studies in Parkinson's disease : Perceptual speech analyses, Australian Journal of Human Communication disorders; 16(2), 17-29.

DARLEY FL, ARONSON AE, BROWN J (1969) Differential diagnostic patterns of dysarthria, Journal of speech and hearing research; 246-269.

DUEZ D (2001) Restoration of deleted or assimilated consonant sequences in conversational French speech : effects of preceding and following contexts, Journal of International Phonetic Association; 31, 101-114.

DUEZ D (2014) Some segmental and prosodic aspects of speech motor disorders. In Motor Speech Disorders : a Cross-language perspective.; Eds N.Miller and AL Lowit, Mulitlingual matters, Bristo, Buffalo, Toronto, 168-195

GHIO A, POUCHOULIN G, TESTON B, PINTO S, FREDOUILLE C, DE LOOZE C, ROBERT D, VIALLET F, GIOVANNI A (2012) How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? Speech Communication, Elsevier: North-Holland, 54 (5), pp.664-679.

GHIO A, LALAIN M, GIUSTI L, POUCHOULIN G, ROBERT D, REBOURG M et al. (2018) Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique, XXXIIe JEP. pp.285-293, https://dx.doi.org/10.21437/jep.2018-33

GREMY F (1958) Contribution à l'étude oscillographique de certaines dysarthries. Thèse de Médecine 1958 ; Paris.

HO AAK, IANSEK R, MARIGLIANI C, BRADSHAW JL, GATES S (1999) Speech Impairment in a large sample of patients with Parkinson's disease. Behavioural Neurology; 11 (3), 131-137.

KENT RD. AND NETSELL (1971) Effects of stress contrasts on certain articulatory parameters. Phonetica 1971 ; 24, 23-44.

KENT RD, ROSENBEK JC (1982) Prosodic disturbance and neurologic lesion, Brain and Language; 15(2) 259-291.

LOGEMANN JA, FISHER HB (1981) Vocal tract control in Parkinson's disease:phonetic feature analysis of misarticulations, J Speech Hear Disord ; 46 (4), 348-52

PLOWMAN-PRINE EK, OKUNA MS., C.M. SAPIENZA CM, R. SHRIVASTAV R, et al (2009) Perceptual characteristics of Parkinsonian speech: A comparison of the pharmacological effects of levodopa across speech and non-speech motor systems, NeuroRehabilitation; 131-144.

UZIEL A, BOHE M, CADILHAC J AND PASSOUANT P (1975): Les troubles de la voix et de la parole dans les syndromes parkinsoniens, Folia Phoniatrica Logopaedica ; 27, 166-176

WARREN RM., WARREN RP. (1970), Auditory illusions and confusions. *Sci. Am.*; 223, 30-36

WEISMER G (1984) Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination. In M.R. McNeil, J.C. Rosenbek & A.E. Aronson (Eds.), The dysarthrias: Physiology, acoustics, perception, management; (pp. 101–130). San Diego: College-Hill Press.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Statistiques des sons naturels et hypothèse du codage efficace pour la perception de la musique et de la parole : Mise en place d'une méthodologie d'évaluation.

Agnieszka Duniec[1] Olivier Crouzet[1,2] Elisabeth Delais-Roussarie[1]
(1) Laboratoire de Linguistique de Nantes, LLING – UMR6310, Université de Nantes / CNRS
chemin de la Censive du Tertre, 44312 Nantes Cedex, France
(2) ENT Department - University Medical Center Groningen, Rijksuniversiteit Groningen, Pays-Bas
agnieszka.duniec@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr,
elisabeth.delais-roussarie@univ-nantes.fr

RÉSUMÉ

L'hypothèse du *codage efficace* prédit que les systèmes perceptifs sont optimalement adaptés aux propriétés statistiques des signaux naturels. Ce caractère optimal a été récemment évalué sur la base d'analyses statistiques réalisées sur des décompositions spectrales de signaux de parole représentés comme des modulations d'énergie. Ces travaux pourraient trouver des applications directes dans l'amélioration du codage des signaux acoustiques par des implants cochléaires. Cependant, les recherches sur la perception de la musique par des personnes sourdes portant un implant cochléaire mettent en avant des limites qui semblent discordantes avec les performances observées concernant certaines propriétés fondamentales de la parole. Nous comparons les résultats d'analyses statistiques de signaux musicaux avec ceux qui ont été réalisés sur de la parole dans le but d'évaluer les impacts respectifs de ces deux gammes de signaux sonores pour évaluer leurs contributions à cette proposition théorique. Des résultats préliminaires et les perspectives futures sont discutés.

ABSTRACT

Natural sound statistics and the efficient coding hypothesis for music and speech perception : setting-up an evaluation methodology.

The *efficient coding* hypothesis predicts that perceptual systems are optimally adapted to natural signal statistics. Such optimal characterization has recently been evaluated on the basis of statistical analyses that were performed on spectral decompositions of speech signals. Speech signals were decomposed into energy envelopes and these results may find applications in the improvement of acoustic signal coding for cochlear implants. However, research on music perception in cochlear implanted deaf listeners sheds light on potential limits associated with music perception that seem to be in contradiction with how some of the fundamental properties of speech sounds are processed. Our aim is to compare the statistical analysis of natural music signals with previous work on speech in order to evaluate their respective contributions to this theoretical proposal. Preliminary results along with future perspectives are discussed.

MOTS-CLÉS : perception, implants cochléaires, statistiques des signaux naturels, hypothèse du codage efficace.

KEYWORDS: perception, cochlear implants, natural signal statistics, efficient coding hypothesis.

Volume 1 : Journées d'Études sur la Parole, pages 181–189.
HAL : hal-02798534.

1 Introduction

La perception de l'environnement sonore est un mécanisme particulièrement complexe. Les travaux princeps de Bregman (1994) ont mis en évidence que les mécanismes impliqués dans l'analyse des scènes auditives reposent en partie sur des processus cognitifs généraux. Si l'on considère en général que des mécanismes fondamentaux d'analyse auditive doivent être impliqués dans les traitements liés aux différents systèmes de communication (langage oral et musique par exemple), il semble par contre raisonnable de considérer que ces deux catégories de signaux restent au moins en partie fondamentalement différentes du point de vue de leur organisation sonore et des informations acoustiques qui les composent (rythmique, mélodique, harmonique).

En outre, on observe des performances très différentes entre la perception de la parole et celle de la musique par des auditeurs sourds portant un implant cochléaire. Si une grande partie des informations associées aux sons de parole peut donner lieu à des performances de reconnaissance satisfaisantes dans des environnements non-dégradés (Bouton *et al.*, 2012), la capacité de ces auditeurs à apprécier ou à identifier un certain nombre d'aspects liés à la musique (mélodie, accords par exemple) reste limitée (Galvin *et al.*, 2009). Ainsi, parole et musique semblent porter des informations qui ne sont pas propices à donner lieu à des traitements perceptifs similaires lorsqu'on les étudie du point de vue de personnes sourdes portant un implant cochléaire.

1.1 L'hypothèse du codage efficace

Proposée initialement par Smith & Lewicki (2006) pour le système auditif sur la base de travaux ayant plutôt porté leur attention sur le système visuel (Simoncelli & Olshausen, 2001), la théorie du « codage efficace » (*efficient coding hypothesis*) postule que les systèmes perceptifs sont optimalement adaptés aux propriétés des signaux naturels de manière à transmettre un maximum d'information en recourant à une consommation minimale de ressources. Cette hypothèse a des implications fortes concernant (1) la nature des représentations sensorielles impliquées dans la perception de notre environnement et (2) les mécanismes perceptifs qui sont mis en œuvre : On s'attend à trouver des correspondances entre propriétés physiques des signaux et caractéristiques des systèmes sensoriels.

Même si cette hypothèse peut paraître aller de soi, elle débouche sur une prédiction qui n'est pas si évidente, laquelle consiste à envisager une réduction maximale de la « granularité » des représentations perceptives permettant aux observateurs d'être parfaitement efficaces sans mettre en oeuvre de représentations « trop » fines qui demanderaient plus de complexité de traitement que nécessaire. Ainsi, les systèmes sensoriels pourraient procéder à une analyse peu précise si elle est « optimale ». L'une des prédictions qui découlent de l'hypothèse du codage efficace est que les signaux naturels développés par les espèces animales auraient évolué de manière à être compatibles avec cette caractéristique et ne requerraient donc pas de contenir plus d'information que nécessaire. De ce point de vue, aussi bien les stimuli naturels que les systèmes sensoriels seraient *optimalement* économes en information, tant que cette économie garantit une analyse perceptive efficace. C'est de ces prédictions que l'*hypothèse du codage efficace* tire son nom.

Si l'on considère cette hypothèse à l'aune des travaux réalisés ces 30 dernières années sur la perception de la parole par des personnes sourdes portant un implant cochléaire aussi bien que par des personnes normo-entendantes soumises à des simulations d'implants cochléaires (parole vocodée à canaux), la plupart des résultats observés semblent aller dans ce sens. Ainsi, certaines informations linguistiques

sont accessibles avec une résolution spectrale très limitée (4 bandes de fréquence suffisent à percevoir avec une performance élevée le voisement ou le mode d'articulation d'une consonne, Shannon *et al.*, 1995). Par contre, d'autres informations acoustiques semblent beaucoup plus problématiques : informations tonales associées à la prosodie de la phrase ou aux tons phonémiques (Milczynski *et al.*, 2012; Gaudrain *et al.*, 2008; Everhardt *et al.*, 2020), genre du locuteur (Fuller *et al.*, 2014), nasalité vocalique (Borel, 2015) par exemple. De même, la perception de la musique, du genre du locuteur, ou de la parole en environnement bruité semblent résister à des conditions plus fines de résolution spectrale (Galvin *et al.*, 2009; Fuller *et al.*, 2014). Les travaux actuels sur la parole vocodée montrent de manière générale qu'un accroissement du nombre de canaux spectraux n'est pas suffisant pour améliorer significativement les performances de perception de ces informations et une grande partie des enjeux théoriques actuels est sous-tendue par cette limite.

1.2 Analyse statistique de signaux naturels

Les signaux naturels présentent des propriétés statistiques qui ont conduit certains auteurs à explorer plus précisément le rôle que jouent ces régularités dans leur reconnaissance. Par exemple, les « textures sonores » (bruit du vent, vol d'insectes…) sont associées à des propriétés statistiques spécifiques qui sont corrélées entre les bandes de fréquence (McDermott & Simoncelli, 2011). La synthèse de sons qui respectent ces régularités corrélées conduit à une reconnaissance de ces textures, ce qui semble indiquer que certains types de « statistiques des signaux naturels » jouent un rôle crucial dans les mécanismes d'identification perceptive.

Certains travaux ont également décrit l'existence de similarités entre des propriétés statistiques observées dans des langues orales et dans de la musique. Ces similitudes porteraient notamment sur la structure énergétique du spectre des harmoniques (Schwartz *et al.*, 2003) ainsi que sur la taille des intervalles tonaux (Han *et al.*, 2011). Ces observations tendent à argumenter en faveur de l'existence de propriétés structurelles parallèles dans la parole et la musique.

Les travaux qui reposent sur l'*hypothèse du codage efficace* s'inspirent de principes relativement proches tirés des travaux sur les *statistiques des signaux naturels* : les signaux de communication seraient caractérisés par des propriétés sonores statistiques qui seraient régulières malgré la diversité apparente des réalisations acoustiques. Ces travaux se sont notamment centrés sur le caractère optimal (1) du nombre de canaux spectraux pour représenter des langues orales mais aussi (2) de la localisation des frontières entre ces canaux.

Ming & Holt (2009) ont montré que, sans changer le nombre de canaux spectraux (6 en l'occurrence) les changements de localisation des frontières spectrales en parole vocodée ont des effets massifs sur les taux de reconnaissance de mots et de segments phonétiques. Ueda & Nakajima (2017), capitalisant sur ces résultats, ont développé une méthode d'analyse inspirée des travaux de Plomp *et al.* (1967) sur les voyelles : ils étendent cette approche à l'étude d'un corpus de phrases. Ils procèdent, sur la base de signaux acoustiques de parole codés sur environ 100 canaux de représentation spectrale répartis en « bandes critiques » étroites, à diverses Analyses en Composantes Principales (ACP, en anglais *PCA*) portant sur les enveloppes d'énergie de ces canaux et varient le nombre de facteurs associés à la sortie de l'ACP (2, 3, 4, 5, 6). Leur travail aboutit à la conclusion que 4 facteurs suffiraient à représenter optimalement des signaux de parole, et ce pour chacune des 8 langues de leur échantillon. Ils constatent par ailleurs que les 3 frontières fréquentielles découlant de chacune des ACP à 4 facteurs réalisées sur ces 8 langues sont parfaitement appariées (env. 540, 1720, 3300 Hz), ce qui les amène à conclure que les langues seraient de manière générale fondées sur des indices qui seraient

parfaitement adaptés à un traitement perceptif « parcimonieux » de la parole.

Récemment, Grange & Culling (2018) ont répliqué l'étude de Ueda & Nakajima (2017) en la mettant en rapport avec des données de perception de parole vocodée (simulations d'implants cochléaires) et ont abouti à des conclusions assez similaires. Leurs résultats suggèrent néanmoins que, pour rendre compte de manière appropriée des propriétés acoustiques de la parole vocodée, il faudrait 6 à 7 canaux spectraux pour représenter optimalement ces signaux. Cette limite correspond dans leurs données, à un point d'inflexion au-delà duquel la performance de reconnaissance mesurée chez les auditeurs ne s'améliore plus.

Si Ming & Holt (2009) se positionnent en faveur d'un traitement efficace relevant de représentations équivalentes quels que soient les signaux envisagés (parole, musique, sons de l'environnement), les données de la littérature concernant les patients sourds qui utilisent un implant cochléaire pourraient amener à nuancer cette position. Ainsi, les performances observées aussi bien chez des auditeurs normo-entendants écoutant des signaux vocodés que chez des patients sourds portant un implant cochléaire sont systématiquement meilleures pour de la parole que pour de la musique (Galvin *et al.*, 2009; Crew *et al.*, 2015), notamment si l'on compare les performances mesurées dans le silence en environnement non-réverbérant. Du point de vue de l'hypothèse du codage efficace, on pourrait être amené à envisager que parole et musique requièrent des niveaux de résolution spectrale très différents pour que leur analyse perceptive soit appropriée. Si tel était le cas, une telle constatation aurait un impact crucial sur les fondements ou la compréhension de cette hypothèse du codage efficace.

L'objet de notre travail est d'évaluer cette contradiction potentielle en mettant en place une série d'analyses qui chercheront dans un premier temps à évaluer les *propriétés statistiques de signaux naturels* de musique et à les comparer à des répliques des analyses réalisées par Ueda & Nakajima (2017) sur de la parole naturelle.

2 Analyse des propriétés statistiques de signaux de musique

2.1 Méthode

L'ensemble des analyses acoustiques et statistiques est réalisé dans l'environnement Matlab. Les scripts d'analyse sont disponibles sur un dépôt github (https://github.com/crouzet-o/naturalSignalStats).

2.1.1 Base de données d'enregistrements musicaux

Dans un souci de réplicabilité des analyses et des résultats qui en découleront, nous avons choisi d'utiliser des extraits musicaux issus d'une base de données en *open source* : *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017). FMA offre la possibilité d'accéder légalement à une bibliothèque d'enregistrements de musique sous licence libre. Elle est constituée de 4 versions, lesquelles contiennent de 8000 à 106574 morceaux musicaux qui sont disponibles soit en extraits de 30 s (les 3 premières versions) soit dans leur intégralité (la 4ème version).

Tous les morceaux musicaux sont au format MP3 et sont associés à des informations qualitatives (*tags*) : numéro d'identification du morceau, titre, artiste, genre (et sous-genres), ainsi qu'à des traits

musicaux (*features*) automatiquement déterminés par la bibliothèque librosa (McFee *et al.*, 2015, traits spectraux, rythmiques…).

Les données présentées ici concernent la base de données la moins volumineuse. Elle est composée de 8000 extraits de 30 secondes de 8 genres musicaux différents en format MP3 (taux de compression entre 128 et 256 kbits/s, fréq. d'échantillonnage 44.1 kHz). Les 8 genres musicaux sont en proportions équilibrées dans cette version de la base.

2.1.2 Paramètrage acoustique des signaux

Préalablement à l'analyse statistique des signaux, nous procédons à une paramétrisation acoustique équivalente à celles qui ont été utilisées dans les travaux précédents (Ueda & Nakajima, 2017; Grange & Culling, 2018). On notera que les travaux antérieurs ayant porté sur de la parole, ils se sont restreints à des fréquences supérieures d'environ 8000 Hz. En ce qui nous concerne, nous manipulons ce paramètre afin de comparer les résultats obtenus en fonction de la limite supérieure de fréquence, laquelle pourrait comporter des informations acoustiques essentielles pour les signaux de musique.

Nous avons analysé une durée totale de signal audio équivalente à celle qui a été étudiée pour les langues les plus fournies de l'échantillon étudié par Ueda & Nakajima (2017, env. 4000 s). Pour celà, nous extrayons pour chaque enregistrement musical disponible les 10 premières secondes. Au total 471 stimuli musicaux ont été exploités, parmi lesquels 71 n'étaient pas lisibles par l'algorithme de décompression MP3 utilisé. L'échantillon final est composé de 400 enregistrements audio fournissant une durée totale de 4000 s (soit environ 1h) d'audio.

Les enregistrements sélectionnés sont ensuite convertis en monophonique par combinaison des deux canaux stéréophoniques et concaténés les uns aux autres. Les enveloppes de modulation temporelle des signaux sont extraites à partir d'un banc de filtres dont la largeur croît de manière logarithmique avec la fréquence centrale (canaux de largeur $\frac{1}{4}$ d'ERB, ce qui correspond à 106 canaux spectraux allant jusqu'à la fréquence supérieure maximale de 8000 Hz et à 129 canaux pour une fréquence maximale de 22000 Hz). Ces enveloppes subissent une rectification demi-onde puis un filtrage passe-bas avec une fréquence de coupure de 50 Hz. Les signaux d'enveloppe résultants sont ensuite élevés au carré et convertis en notes centrées réduites (*z-scores*). Cette chaîne de paramétrage permet de procéder à une analyse des co-modulations entres les bandes de fréquence sur une base d'analyse des corrélations entre les informations d'enveloppe.

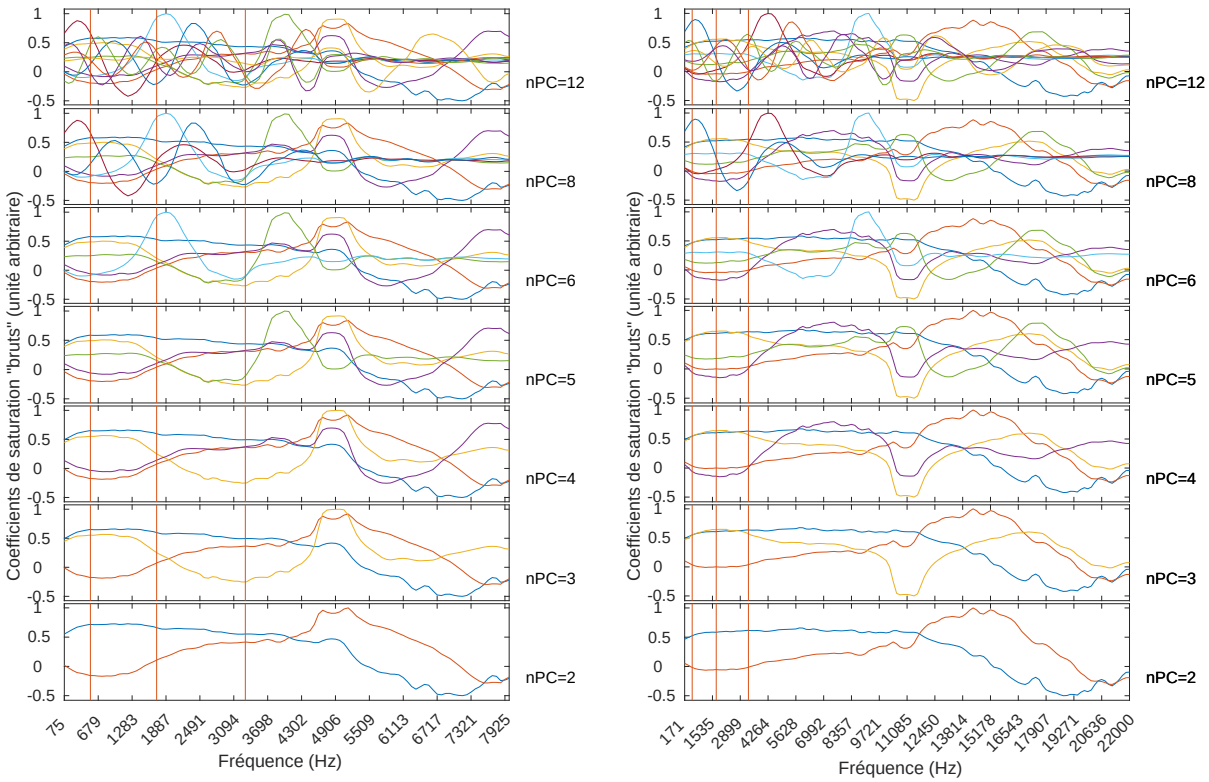
Le signal résultant, composé de 106 / 129 canaux en fonction de la fréquence maximale, correspond aux modulations temporelles de l'enveloppe de chaque canal au cours du temps. Cette matrice de modulations d'amplitude est alors transférée vers un outil statistique d'analyse en composantes principales (*Principal Components Analysis*).

2.1.3 Analyse en Composantes Principales

L'Analyse en Composantes Principales est une méthode descriptive d'analyse de données qui permet une étude simultanée de plus de 2 dimensions (analyse multivariée). L'objectif est de représenter l'essentiel de l'information contenue dans un tableau de données quantitatif en réduisant le nombre de facteurs explicatifs. Le principe est de transformer des variables liées (ayant des corrélations statistiques) en nouvelles variables synthétiques (composantes principales) en perdant le moins d'information possible. Cette analyse permet non seulement de réduire le nombre de variables à

mesurer, et ainsi améliorer la caractérisation des données, mais aussi d'identifier les facteurs non corrélés utiles pour procéder à une analyse discriminante. Concrètement, les variables initiales sont représentées dans un nouvel espace de facteurs définis par les vecteurs propres de la matrice de corrélations. L'hypothèse sous-jacente à l'application de cette méthode sur des signaux sonores est que certains canaux spectraux contiendraient des informations redondantes et qu'il serait alors économe de restreindre l'analyse perceptive à une séparation en zones de fréquences étant maximalement informatives (donc minimalement redondantes). En celà, l'ACP nous permettrait d'identifier les canaux de fréquence optimaux pour différencier de manière parcimonieuse les enregistrements d'un corpus.

L'analyse préliminaire des résultats porte essentiellement sur la description des graphiques indiquant les valeurs des coefficients de saturation pour chaque composante principale en fonction du canal de fréquence (Fig. 1), ce qui permet de décrire l'empan de fréquences qui corrèle avec une même composante.



(a) Fréq. Max. : 8000 Hz

(b) Fréq. Max. : 22000 Hz

FIGURE 1 – Coefficients de saturation (*factor loadings*) issus des ACP réalisées sur un échantillon des signaux de musique de la base de données *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017), durée totale 4000 s, fréquence maximale supérieure de (a) 8000 Hz, (b) 22000 Hz

2.2 Résultats

Les travaux antérieurs ayant porté sur de la parole, ils se sont restreints à des fréquences supérieures autour de 8000 Hz. En ce qui nous concerne, nous manipulons ce paramètre afin de comparer les

résultats obtenus en fonction de la limite supérieure de fréquence, à savoir 22000 Hz car cette gamme supérieure pourrait comporter des informations acoustiques essentielles pour les signaux de musique.

On peut dans un premier temps observer que la prise en compte des fréquences supérieures à 8000 Hz se justifie totalement puisque la deuxième composante principale (en rouge sur la fig. 1) s'étend approximativement sur la gamme 12000 à 16000 Hz. La figure 1a représente la distribution des coefficients de saturation des composantes principales sur la gamme de fréquences s'étendant jusqu'à 8000 Hz en fonction du nombre de facteurs étudiés. Le nombre de facteurs augmente de bas en haut (de 2 à 12). La figure 1b représente l'analyse réalisée sur une plus large gamme de fréquences mais pour les mêmes signaux et le même nombre de composantes principales.

Dans l'analyse réalisée par Ueda & Nakajima (2017), les frontières entre canaux optimaux sont placées aux croisements des courbes de saturation, délimitant ainsi des zones de fréquences. Les frontières que nous avons utilisées dans nos analyses sont celles de Ueda & Nakajima (2017, les traits verticaux de couleur orange sur les graphiques de la fig. 1). Dans l'analyse réalisée par les auteurs, ces frontières délimitent quatre zones de fréquence principales : '*low*' centrée à 300 Hz, '*mid-low*' centrée autour de 1000 Hz, '*mid-high*' centrée à 2200 Hz et '*high*' centrée autour de 4500 Hz. Dans l'analyse que nous avons réalisée jusqu'à 22000 Hz, les 3 zones inférieures identifiées par les auteurs ('*low*', '*mid-low*' et '*mid-high*') semblent peu représentées. Même la zone supérieure ('*high*') autour de 4500 Hz n'apparaît qu'à partir de la 8ème CP même si la zone de hautes fréquences de notre analyse (à partir de 5000 Hz) est largement représentée par les 5 à 6 premières CP. Néanmoins, la 3ème CP peut-être décrite comme s'étendant sur les 3 zones inférieures identifiées par Ueda & Nakajima (2017). Ces observations reflètent également les données obtenues en se limitant à la gamme de fréquences inférieure à 8000 Hz. Dans nos analyses, un découpage plus fin associé aux 3 zones de basse fréquence identifiées par Ueda & Nakajima (2017) apparaît seulement à partir de la huitième CP, particulièrement si l'on s'intéresse à l'analyse jusqu'à 22000 Hz.

3 Discussion

À la lueur de ces premières explorations, ont peut identifier deux points centraux : (1) si les données musicales ne coincident pas avec les données de parole décrites dans la littérature, certains principes similaires semblent ressortir (répartition des coefficients de saturation sur des gammes de fréquence spécifiques ; correspondance possible mais encore à l'état de conjecture entre les zones de fréquence décrites en parole et en musique, cependant sur des composantes très différentes –composantes de plus grande contribution / de plus bas niveau pour la parole que pour la musique). (2) il semble probable (et pas surprenant) que les signaux de musique requièrent un plus grand nombre de composantes principales pour être caractérisés que les signaux de parole mais il restera à identifier dans quelle mesure les frontières fréquentielles pertinentes pourraient entrer en correspondance.

Les résultats présentés ici sont préliminaires et constituent seulement une ébauche des analyses à venir. Nous développons actuellement une amélioration des scripts d'analyse afin de mettre en place une rotation orthogonale des facteurs qui permettra de faciliter la description des résultats issus des ACP. En outre, à partir des principes d'extraction de données présentés ici, nous nous orienterons vers des analyses plus objectives (1) des composantes principales à prendre en compte à partir des mesures d'inertie (2) ainsi que des frontières de fréquence correspondantes.

Du point de vue des perspectives générales de ce travail, l'hypothèse du codage efficace prédit qu'un

nombre réduit de dimensions devrait permettre de caractériser les signaux de musique au même titre que les signaux de parole. Cette hypothèse pourra donc être évaluée de manière *interne* : en étudiant les résultats obtenus pour des signaux de musique indépendamment des résultats antérieurs sur la parole. Nous chercherons à établir dans quelle mesure l'analyse statistique de signaux musicaux permet de retrouver certains *patterns* observés pour la parole : nombre réduit de canaux optimaux par rapport au nombre total de canaux de codage en enveloppe d'énergie, correspondance des résultats quel que soit le genre musical, comparaison des résultats en fonction de certains paramètres qualitatifs (présence majoritaire d'informations à large bande comme des percussions vs. caractéristiques tonales).

Parallèlement à cette évaluation interne, certains résultats devront nécessairement être discutés dans une perspective comparative articulant les données observées sur la parole et sur la musique : il est tout à fait possible que les détails de ces analyses fassent ressortir (1) que les signaux de musique requièrent un plus grand nombre de canaux spectraux que la parole, et (2) que les frontières entre les canaux (à nombre de canaux équivalent ou pas) soient divergentes si l'on compare les analyses qui reposent sur la parole et celles qui reposent sur la musique.

Afin de pouvoir s'assurer que d'éventuelles divergences de résultats entre d'une part nos analyses réalisées sur de la musique et, d'autre part ceux de Ueda & Nakajima (2017) et Grange & Culling (2018), ne seraient pas le fait de différences fines qui interviendraient dans la mise en œuvre des algorithmes, nous comparerons nos résultats obtenus sur des signaux de musique avec une base de données de parole. Ceci permettra de s'assurer que nous répliquons les observations des travaux antérieurs (Ueda & Nakajima, 2017; Grange & Culling, 2018). Dans le souci de s'approcher au mieux des conditions des études précédentes, nous utiliserons une base de données de phrases collectées en laboratoire et contenant un nombre suffisament large de phrases différentes pour chaque locuteur.

Les résultats observés constitueront une source précieuse d'information, d'une part pour évaluer les fondements de l'hypothèse du codage efficace et ses impacts sur la modélisation perceptive des signaux naturels, d'autre part pour envisager dans quelle mesure cette hypothèse pourrait conduire à développer des solutions permettant d'améliorer le codage de signaux sonores par des dispositifs comme les implants cochléaires.

Remerciements

Ce travail a reçu le soutien du programme Recherche – Formation – Innovation « Ouest Industries Créatives » (RFI-OIC, Région Pays de la Loire) par une allocation doctorale attribuée à AD.

Références

BOREL S. (2015). *Perception auditive, visuelle et audiovisuelle des voyelles nasales par les adultes devenus sourds. Lecture labiale, implant cochléaire, implant du tronc cérébral.* Thèse de doctorat, Université de la Sorbonne Nouvelle – Paris 3.

BOUTON S., SERNICLAES W., BERTONCINI J. & COLÉ P. (2012). Perception of Speech Features by French-Speaking Children With Cochlear Implants. *Journal of Speech Language and Hearing Research*, **55**(1), 139–153. DOI : 10.1044/1092-4388(2011/10-0330).

BREGMAN A. S. (1994). *Audirory scene analysis : the perceptual organization of sound.* A Bradford book, Cambridge, Mass. : MIT Press., 2nd édition.

CREW J. D., GALVIN J. J. & FU Q.-J. (2015). Melodic contour identification and sentence recognition using sung speech. *The Journal of the Acoustical Society of America*, **3**(138).

DEFFERRARD M., BENZI K., VANDERGHEYNST P. & BRESSON X. (2017). Fma : Dataset for music analysis. *18th International Society for Music Information Retrieval Conference*.

EVERHARDT M. K., SARAMPALIS A., COLER M., BAŞKENT D. & LOWIE W. (2020). Meta-Analysis on the Identification of Linguistic and Emotional Prosody in Cochlear Implant Users and Vocoder Simulations :. *Ear and Hearing*, p. in press. DOI : 10.1097/AUD.0000000000000863.

FULLER C. D., GAUDRAIN E., CLARKE J. N., GALVIN J. J., FU Q.-J., FREE R. H. & BAŞKENT D. (2014). Gender categorization is abnormal in cochlear implant users. *Journal of the Association for Research in Otolaryngology,*, **6**(15).

GALVIN J. J., FU Q.-J. & SHANNON R. V. (2009). Melodic contour identification and music perception by cochlear implant users. *Annals of the New York Academy of Sciences*, **1**(1169), 518–533.

GAUDRAIN E., GRIMAULT N., HEALY E. W. & BÉRA J.-C. (2008). Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation. *The Journal of the Acoustical Society of America*, **124**(5), 3076–87. DOI : 10.1121/1.2988289.

GRANGE J. & CULLING J. (2018). The factor analysis of speech : Limitations and opportunities for cochlear implants. *Acta Acustica united with Acustica*, **104**, 835–838.

HAN S. E., SUNDARARAJAN J., BOWLING D. L., LAKE J. & PURVES D. (2011). Co-Variation of Tonality in the Music and Speech of Different Cultures. *PLOS ONE*, **6**(5), e20160. DOI : 10.1371/journal.pone.0020160.

MCDERMOTT J. H. & SIMONCELLI E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery : Evidence from Sound Synthesis. *Neuron*, **71**(5), 926–940. DOI : 10.1016/j.neuron.2011.06.032.

MCFEE B., RAFFEL C., LIANG D., ELLIS D., MCVICAR M., BATTENBERG E. & NIETO O. (2015). librosa : Audio and Music Signal Analysis in Python. In *The 14th Python in Science Conference (scipy 2015)*, p. 18–24, Austin, Texas. DOI : 10.25080/Majora-7b98e3ed-003.

MILCZYNSKI M., CHANG J. E., WOUTERS J. & VAN WIERINGEN A. (2012). Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues. *Hearing Research*, **285**(1–2), 1–12. DOI : 10.1016/j.heares.2012.02.006.

MING V. L. & HOLT L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, **3**(126), 1312–1320.

PLOMP R., POLS L. C. W. & VAN DE GEER J. P. (1967). Dimensional analysis of vowel spectra. *The Journal of the Acoustical Society of America*, **3**(41), 707–712.

SCHWARTZ D. A., HOWE C. Q. & PURVES D. (2003). The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *Journal of Neuroscience*, **23**(18), 7160–7168. DOI : 10.1523/JNEUROSCI.23-18-07160.2003.

SHANNON R., ZENG F., KAMATH V., WYGONSKI J. & EKELID M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.

SIMONCELLI E. P. & OLSHAUSEN B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, **24**(1), 1193–1216. DOI : 10.1146/annurev.neuro.24.1.1193.

SMITH E. C. & LEWICKI M. S. (2006). Efficient auditory coding. *Nature*, **7079**, 978–982.

UEDA K. & NAKAJIMA Y. (2017). An acoustic key to eight languages/dialects : Factor analyses of critical-band-filtered speech. *Scientific Reports*, **7**, 42468. DOI : 10.1038/srep42468.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Adaptation de domaine non supervisée pour la reconnaissance de la langue par régularisation d'un réseau de neurones

Raphaël Duroselle[1] Denis Jouvet[1] Irina Illina [1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
raphael.duroselle@loria.fr denis.jouvet@inria.fr irina.illina@loria.fr

RÉSUMÉ

Les systèmes automatiques d'identification de la langue subissent une dégradation importante de leurs performances quand les caractéristiques acoustiques des signaux de test diffèrent fortement des caractéristiques des données d'entraînement. Dans cet article, nous étudions l'adaptation de domaine non supervisée d'un système entraîné sur des conversations téléphoniques à des transmissions radio. Nous présentons une méthode de régularisation d'un réseau de neurones consistant à ajouter à la fonction de coût un terme mesurant la divergence entre les deux domaines. Des expériences sur le corpus OpenSAD15 nous permettent de sélectionner la *Maximum Mean Discrepancy* pour réaliser cette mesure. Cette approche est ensuite appliquée à un système moderne d'identification de la langue reposant sur des *x-vectors*. Sur le corpus RATS, pour sept des huit canaux radio étudiés, l'approche permet, sans utiliser de données annotées du domaine cible, de surpasser la performance d'un système entraîné de façon supervisée avec des données annotées de ce domaine.

ABSTRACT

Unsupervised domain adaptation for language identification by regularization of a neural network

Automatic spoken language identification systems suffer from a performance drop when acoustic characterics of the test signal differ in a significative way from the characteristics of the training data. In this paper, we study the unsupervised domain adaptation of a system trained on conversational telephone speech to radio transmission channels. We present a regularization method for a neural network which consists in adding to the cost function a term that measures the discrepancy between domains. Based on experiments on the corpus OpenSAD15, we select the *Maximum Mean Discrepancy* loss to perform this measure. This approach is then applied to a state-of-the-art x-vector system. On the RATS corpus, for seven of the eight studied radio channels, our approach achieves a better performance on the target domain than a system trained in a supervised way using labelled data from this domain.

MOTS-CLÉS : adaptation de domaine non supervisée, identification de la langue, régularisation, maximum mean discrepancy, robustesse.

KEYWORDS: unsupervised domain adaptation, language identification, regularization, maximum mean discrepancy, robustness.

Volume 1 : Journées d'Études sur la Parole, pages 190–198.
HAL : hal-02798536.

1 Introduction

Un système d'identification de la langue est habituellement entraîné sur un corpus d'apprentissage spécifique à un environnement. Si les données de test ne proviennent pas de la même distribution que les données d'entraînement (et ont donc des caractéristiques différentes), les performances du système peuvent chuter significativement. Dans cet article, nous étudions l'effet du changement de canal de transmission entre les données d'entraînement et de test. Les données d'entraînement sont des conversations téléphoniques. Nous voulons appliquer un tel système à des communications radio, pour lesquelles nous ne disposons pas de données annotées. Ce problème est appelé adaptation de domaine non supervisée.

La possibilité d'adapter un système de classification fonctionnant sur un domaine source à un domaine cible repose sur l'hypothèse que les distributions des données des deux domaines partagent des caractéristiques communes pouvant être utilisées pour la classification (Ben-David *et al.*, 2010). Par conséquent l'adaptation de domaine peut être réalisée en utilisant des représentations invariantes entre les domaines. Dans ce but, deux types d'approche ont émergé (Bousquet & Rouvier, 2019) : les méthodes *feature-based*, qui transforment les représentations des données du domaine source afin de les rendre similaires au domaine cible, et les méthodes *model-based*.

Lors d'une adaptation *model-based*, les paramètres du modèle sont déterminés en prenant en compte l'objectif de généralisation au domaine cible. Nous proposons une approche *model-based* s'appliquant à un réseau de neurones dont les paramètres sont obtenus par minimisation d'une fonction de coût. Un terme de régularisation est ajouté à la fonction de coût afin de prendre en compte la contrainte d'invariance entre les domaines. Différentes fonctions de régularisation ont été proposées dans la littérature en traitement de l'image et analyse de texte : *deep CORAL* (Sun & Saenko, 2016) *Maximum Mean Discrepancy* (Long *et al.*, 2015), des fonctions de coût antagonistes (Ganin *et al.*, 2016). Jusqu'à présent aucune de ces approches n'a été appliquée à la reconnaissance de la langue.

Dans ce travail, nous comparons d'abord trois fonctions de coût pour l'adaptation à des canaux radio d'un réseau de neurones entraîné pour la tâche d'identification de la langue : la distance entre les moyennes des distributions, *deep CORAL* et la *Maximum Mean Discrepancy*. Nous montrons que cette dernière permet d'annuler la baisse de performance due à l'absence de données annotées sur le domaine cible.

Dans un second temps, nous étudions un système d'identification de la langue correspondant à l'état de l'art (Snyder *et al.*, 2018; Plchot *et al.*, 2018), constitué d'un extracteur de *features*, d'un extracteur de vecteurs représentatifs des segments audio et d'un classifieur final. Un tel système subit bien une dégradation importante de performance due au changement de canal de transmission entre les données d'entraînement et de test. Notre approche appliquée au module d'extraction de vecteurs représentatifs du segment audio permet de réduire cette dégradation et conduit même à des performances meilleures que celles d'un système entraîné de façon supervisée sur le domaine cible.

2 Méthode d'adaptation de domaine non supervisée d'un réseau de neurones

Nous nous plaçons dans le cadre d'une adaptation de domaine non supervisée. Nous disposons de données annotées (x_S, y_S) provenant d'un domaine source défini par sa distribution $\mathcal{D}_S$, et de

données non annotées x_T d'un domaine cible défini par sa distribution $\mathcal{D}_T$. Les x_S et x_T sont les données audio et les y_S les étiquettes de langue associées. L'objectif de la tâche d'adaptation de domaine est l'entraînement d'un système d'identification de langue performant sur le domaine cible.

2.1 Régularisation de la fonction de coût

Nous nous intéressons à un modèle de classification pour la tâche d'identification de la langue. C'est un réseau de neurones f_θ de paramètres θ. Ses paramètres sont appris de façon supervisée en minimisant l'entropie croisée L_{CE} sur le domaine source :

$$\min_{\theta} \mathbb{E}_{(x_S,y_S)\sim\mathcal{D}_S}[L_{CE}(f_\theta, x_S, y_S)] \tag{1}$$

Fondée sur le constat que l'erreur du modèle sur le domaine cible peut être contrôlée par la somme de l'erreur sur le domaine source et d'une mesure de divergence entre les domaines (Ben-David *et al.*, 2010), notre méthode *model-based* d'adaptation de domaine non supervisée consiste à ajouter une fonction de régularisation L_R à la fonction de coût. Le problème d'optimisation devient :

$$\min_{\theta} \mathbb{E}_{(x_S,y_S)\sim\mathcal{D}_S}[L_{CE}(f_\theta, x_S, y_S)] + \lambda L_R(f_\theta, \mathcal{D}_S, \mathcal{D}_T) \tag{2}$$

L_R est une mesure de la divergence des représentations du réseau entre les distributions $\mathcal{D}_S$ et $\mathcal{D}_T$. λ est un paramètre représentant le compromis entre bonne performance de classification sur le domaine source et invariance des représentations entre les domaines. En pratique, on choisit une couche du réseau et la fonction de coût L_R est mesurée pour les activations de celle-ci. Nous utilisons la notation $\Phi_f(x)$ pour les valeurs des activations de cette couche pour un réseau f et une donnée d'entrée x. Dans nos expériences, nous nous plaçons sur la couche de sortie du réseau.

Différentes fonctions de régularisation L_R ont été introduites : *deep CORAL* (Sun & Saenko, 2016), *Maximum Mean Discrepancy* (Long *et al.*, 2015; Lin *et al.*, 2018), ainsi que des fonctions de coût antagonistes (*adversarial*) (Ganin *et al.*, 2016). Dans ce travail, nous comparons trois fonctions de régularisation, basées sur la distance entre les moyennes des distributions, sur la distance entre les seconds moments (*deep CORAL*) et sur la *Maximum Mean Discrepancy*.

2.2 Fonctions de régularisation

Une correction simple à appliquer à deux distributions de probabilité pour les rapprocher serait de leur faire partager la même moyenne. Par conséquent, notre première fonction de régularisation est le carré de la **distance euclidienne entre les moyennes** des distributions des deux domaines :

$$L_{moy} = || \mathbb{E}_{x_S\sim\mathcal{D}_S}[\Phi_f(x_S)] - \mathbb{E}_{x_T\sim\mathcal{D}_T}[\Phi_f(x_T)] ||_2^2 \tag{3}$$

Dans le même esprit, la fonction de coût **deep CORAL** (Sun & Saenko, 2016) vise à aligner les seconds moments des deux distributions. Elle correspond au carré de la distance euclidienne entre les matrices de covariance des distributions de chacun des deux domaines :

$$L_{CORAL} = ||C_S - C_T||_2^2 \tag{4}$$

où C_S et C_T sont les matrices de covariance des activations $\Phi_f(x)$ sur les domaines $\mathcal{D}_S$ et $\mathcal{D}_T$.

Enfin, la ***Maximum Mean Discrepancy*** (MMD) est une mesure de divergence entre les domaines basée sur une mesure de similarité entre paires d'échantillons définie par un noyau semi-défini positif k. Elle prend la valeur :

$$L_{MMD} = \mathop{\mathbb{E}}_{x_S, x'_S \sim \mathcal{D}_S}[k(\Phi_f(x_S), \Phi_f(x'_S))] + \mathop{\mathbb{E}}_{x_T, x'_T \sim \mathcal{D}_T}[k(\Phi_f(x_T), \Phi_f(x'_T))] - 2\mathop{\mathbb{E}}_{x_S \sim \mathcal{D}_S, x_T \sim \mathcal{D}_T}[k(\Phi_f(x_S), \Phi_f(x_T))] \quad (5)$$

Lorsque le noyau est le produit scalaire usuel alors L_{MMD} est équivalente à L_{moy}, présentée précédemment. Pour prendre en compte de façon plus fine l'écart entre les distributions, nous utilisons un noyau gaussien, de variance notée σ^2 :

$$k(\Phi_f(x), \Phi_f(x')) = \exp(-\frac{||\Phi_f(x) - \Phi_f(x')||_2^2}{2\,\sigma^2}) \quad (6)$$

La régularisation *MMD* est une mesure de divergence entre deux distributions de probabilité, pouvant être estimée à partir d'un nombre fini d'échantillons, y compris dans des espaces de haute dimension (Peyré & Cuturi, 2019). Dans le domaine du traitement de la parole, elle a été utilisée pour l'adaptation de domaine *feature-based* d'un système de reconnaissance du locuteur (Lin *et al.*, 2018). De plus, l'estimation de la *Maximum Mean Discrepancy* peut être réalisée de façon efficace sur un GPU (Feydy *et al.*, 2019).

Au cours de l'apprentissage, ces trois fonctions de régularisation seront simplement estimées par moyenne empirique sur chaque *minibatch*, puis ajoutées au coût de classification, voir Équation (2).

3 Sélection de la fonction de régularisation

Pour isoler l'effet de la méthode de régularisation proposée, nous comparons les trois fonctions de régularisation sur un système *end-to-end* constitué d'un seul réseau de neurones, avec le corpus OpenSAD15.

3.1 Architecture du système *end-to-end*

L'identification de la langue peut être directement réalisée avec un réseau de neurones convolutionnel (Lozano-Diez *et al.*, 2015). Nous utilisons une architecture similaire décrite dans le Tableau 1. Les *features* d'entrée de notre système sont des MFCC de dimension 12, calculés pour des trames de 10 ms. Nous réalisons la classification en utilisant directement les probabilités *a posteriori* renvoyées par la couche de sortie pour chaque langue. Le système est entraîné et évalué avec des segments de parole de trois secondes.

3.2 Le corpus OpenSAD 2015

Pour étudier l'effet du changement de canal de transmission, nous avons réalisé nos expériences préliminaires sur quatre langues du corpus OpenSAD15 (NIST, 2016) : anglais, arabe, pashto et urdu.

TABLE 1 – Architecture du réseau de neurones convolutionnel

Convolutions selon l'axe temporel			
nom de la couche	taille du noyau / du *max pooling*	nombre de filtres	fonction d'activation
conv. 1	5 / 2	1024	ReLU
conv. 2	5 / 2	1024	ReLU
conv. 3	5 / 2	128	ReLU
Agrégation statistique des moyennes et écarts-types (pooling)			
dimension de sortie : $2 \times 128 = 256$			
Couches connectées			
nom de la couche	dimension		fonction d'activation
fc. 1	256×128		ReLU
fc. 2	128×4		Softmax

Il s'agit d'un corpus créé à partir de conversations téléphoniques, canal *src* du corpus, qui ont ensuite été transmises par six systèmes radio différents : B, F, G (UHF), E (VHF), D et H (HF).

Afin d'éviter un biais lors de l'apprentissage, nous n'utilisons que la moitié des données d'entraînement. La moitié des fichiers audio d'origine sont utilisés pour le domaine source (canal *src*) et l'autre moitié, correspondant à des phrases différentes, est utilisée pour les domaines cibles (canaux radio). De cette façon, le même contenu linguistique n'est pas présent sur les deux domaines lors de l'apprentissage.

3.3 Résultats des expériences préliminaires

Nous réalisons différents entraînements du réseau de neurones convolutionnel sur le corpus OpenSAD15. La performance de chacun des systèmes sur les canaux d'intérêt est présentée dans le Tableau 2. Les performances sont mesurées avec un *Equal Error Rate* (EER) moyen pour des segments de parole de trois secondes. Un EER est calculé pour chacune des quatre langues du corpus et le score obtenu est la moyenne arithmétique des taux de chaque langue.

TABLE 2 – Résultats en taux d'égale erreur de différentes méthodes d'entraînement du réseau convolutionnel pour le corpus OpenSAD15 (segments de 3 secondes). Le domaine source est le canal téléphonique (*src*).

Méthode d'apprentissage	EER sur le domaine cible (%)					
	B	*D*	*E*	*F*	*G*	*H*
supervisé sur *source*	57	52	48	51	30	50
supervisé sur *cible*	**18**	15	19	15	14	22
distance moyennes	53	44	38	35	12	41
deep CORAL	32	32	26	18	11	20
MMD	19	**11**	**16**	**13**	**9**	**18**

Le réseau est d'abord entraîné avec des données du canal téléphonique. Ce système, qui obtient un EER moyen de 8% sur le canal *src* est totalement inopérant sur les canaux cibles. Cependant, lorsqu'un système est entraîné de façon supervisée sur chacun des canaux cibles, alors nous obtenons un EER moyen compris entre 14% et 22%.

Les trois dernières lignes du Tableau 2 présentent les performances de l'apprentissage avec chacune des fonctions de régularisation proposées, en utilisant des données annotées du canal *src* et des données non annotées sur le domaine cible. Le paramètre λ (ainsi que σ^2 pour la *MMD*) est sélectionné pour chaque fonction de coût en fonction de la performance obtenue sur un ensemble de validation. Les résultats de tous les domaines cibles sont cohérents : les méthodes de régularisation permettent d'améliorer les EER moyens sur le domaine cible par rapport à un apprentissage sur le domaine source. Une hiérarchie claire apparaît : la *MMD* avec noyau gaussien est plus efficace que *deep CORAL*, qui est elle-même supérieure à la contrainte sur la distance entre les moyennes. Ce résultat signifie que, pour supprimer la distorsion due au changement de canal, le système ne peut se limiter aux deux premiers moments des distributions mais doit prendre en compte une géométrie plus complexe.

Pour cinq des six domaines cibles testés, la régularisation avec la fonction de coût *MMD* permet d'obtenir une meilleure performance sur le domaine cible que l'apprentissage supervisé sur ce domaine, alors même que l'entraînement n'a pas utilisé de données annotées du domaine cible.

4 Application à un système à l'état de l'art

Les expériences préliminaires ont permis de sélectionner la fonction de régularisation basée sur la *Maximum Mean Discrepancy* pour l'adaptation d'un réseau de neurones convolutionnel. Nous appliquons donc cette méthode d'apprentissage à un système d'identification de la langue correspondant à l'état de l'art pour cette tâche.

4.1 Architecture du système

Un système moderne de reconnaissance de la langue (Snyder *et al.*, 2018; Plchot *et al.*, 2018) est en règle générale formé de trois modules : un extracteur de représentations pour des trames localisées dans le temps, un extracteur de représentations pour l'ensemble du segment audio et un classifieur final.

Dans notre système, le premier module extrait des *stacked multilingual bottleneck features*. Il s'agit des activations d'une couche intermédiaire (*bottleneck*) d'un réseau de neurones ayant été entraîné à reconnaître des triphones pour dix-sept langues du corpus Babel. Nous utilisons les réseaux entraînés *BUT/PHONEXIA bottleneck feature extractor* (Fer *et al.*, 2017), ayant donné de bons résultats pour l'évaluation NIST LRE 2017 (Plchot *et al.*, 2018). Ils génèrent des *bottleneck features* de dimension 80, pour chaque trame de 10 ms.

Le deuxième module extrait un vecteur représentatif par segment. C'est un réseau de neurones prenant en entrée la séquence des *bottleneck features* et entraîné de façon supervisée à prédire la langue utilisée dans le segment. Nous utilisons l'architecture du système *x-vector* (Snyder *et al.*, 2018), constituée de cinq couches procédant à des traitements par trame, suivies d'une couche d'agrégation (*pooling*) statistique et de trois couches pleines. Les *x-vectors* issus de cette architecture sont des vecteurs de dimension 512.

Enfin le classifieur final prend en entrée un *x-vector* et produit un score pour chacune des langues cibles. Notre classifieur final est composé d'une *LDA* (*Linear Discriminant Analysis*), utilisée pour réduire la dimension, d'un blanchiment par multiplication matricielle et d'un *SVM* (*Support Vector Machine*).

Nous appliquons la méthode de régularisation basée sur la *Maximum Mean Discrepancy* au réseau *x-vector*, dans le but de produire des *x-vectors* invariants au changement de canal. Pour des systèmes similaires consacrés à la tâche de reconnaissance du locuteur, l'adaptation *model-based* du réseau *x-vector* a permis de réduire la distorsion due à la langue (Rohdin *et al.*, 2019) et aux conditions acoustiques (Bhattacharya *et al.*, 2019), avec des fonctions de coût antagonistes.

4.2 Le corpus RATS

Nous entraînons ce système sur le corpus RATS (Walker & Strassel, 2012). Nous utilisons les livraisons LDC2015S02 et LDC2017S20 qui comptent cinq langues : anglais, arabe, farsi, pashto et urdu. Ce corpus présente les mêmes caractéristiques que le corpus OpenSAD15 qui en est un sous-ensemble. Il contient deux canaux UHF supplémentaires : A et C. Comme pour le corpus OpenSAD15, nous n'utilisons que la moitié du corpus afin qu'un même contenu linguistique ne soit pas présent à la fois sur les domaines source et cible.

L'identification de la langue a été étudiée sur le corpus RATS (Matějka *et al.*, 2014; Lei *et al.*, 2014; Han *et al.*, 2013) avec la livraison LDC2018S10, contenant également cinq langues : arabe, dari, farsi, pashto et urdu. Pour des segments de trois secondes et pour tous les canaux, le meilleur EER moyen obtenu est de 9.59% (Matějka *et al.*, 2014).

4.3 Expériences

Tout d'abord, nous entraînons le système sur tous les canaux, nous obtenons un EER moyen de 9.36% comparable à l'état de l'art sur le corpus RATS.

Ensuite nous procédons à l'entraînement du réseau *x-vector* de façon supervisée sur le domaine source (canal téléphonique) puis sur chacun des canaux cibles. Enfin, pour chacun des domaines cibles, nous appliquons la régularisation basée sur la *Maximum Mean Discrepancy*. Rappelons que le réseau de neurones n'est pas utilisé directement pour réaliser la classification mais pour extraire un *x-vector* représentatif du segment audio. Pour évaluer les propriétés des *x-vectors* ainsi extraits, nous réalisons plusieurs systèmes en entraînant le classifieur final avec des données annotées, soit du domaine source, soit du domaine cible. Les performances de chacun de ces systèmes sont présentées dans le Tableau 3.

TABLE 3 – Résultats en taux d'égale erreur de différentes méthodes d'entraînement du réseau *x-vector* et du classifieur final pour huit canaux radio du corpus RATS (segments de 3 secondes)

Méthode d'entraînement		EER moyen sur le domaine cible (%)							
x-vector	classifieur final	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
supervisé sur *source*	supervisé sur *source*	50,2	42,3	34,4	39,6	48,5	45,1	17,4	43,6
supervisé sur *source*	supervisé sur *cible*	15,8	15,0	14,1	14,3	21,8	20,5	9,8	18,8
supervisé sur *cible*	supervisé sur *cible*	14,6	12,5	12,6	6,7	13,6	13,5	8,6	14,2
MMD	supervisé sur *source*	12,7	10,6	11,7	7,6	13,3	11,9	5,5	12,2
MMD	supervisé sur *cible*	**10,2**	**9,2**	**11,3**	**6,0**	**11,8**	**10,3**	**5,1**	**10,0**

Notons d'abord qu'un système entraîné sur le domaine source, qui obtient pourtant un EER moyen de 6.0% sur ce domaine, atteint une très mauvaise performance sur les canaux radio. À l'opposé,

l'entraînement supervisé du système sur le domaine cible atteint des EER moyens compris entre 6.7% et 14.6%. D'autre part, nous observons que l'entraînement supervisé du classifieur final sur le domaine cible avec des *x-vectors* entraînés sur le domaine source (ligne 2 du Tableau 3) ne suffit pas à atteindre la performance d'un système totalement entraîné sur le domaine cible (ligne 3). Ce constat justifie la nécessité de développer une méthode d'adaptation de domaine pour le réseau *x-vector*.

La régularisation du réseau *x-vector* avec la *Maximum Mean Discrepancy* est un succès. Les *x-vectors* produits par ce réseau ont acquis une robustesse au changement de canal puisqu'un classifieur final entraîné sur le domaine source avec ces *x-vectors* (ligne 4 du Tableau 3) obtient une bonne performance sur le domaine cible. En fait, pour tous les canaux à l'exception du canal D, l'adaptation de domaine du réseau *x-vector* avec un classifieur final entraîné sur le domaine source est plus performante que l'entraînement de tout le système sur le domaine cible (ligne 3). Ce résultat confirme nos expériences préliminaires : non seulement les valeurs de sortie mais aussi les activations de la couche *x-vector* du réseau acquièrent une invariance au domaine grâce à la méthode de régularisation.

Finalement, l'entraînement d'un classifieur final sur le domaine cible avec les *x-vectors* obtenus par régularisation (ligne 5 du Tableau 3) conduit à des EER moyens significativement inférieurs à un système totalement entraîné sur le domaine cible (ligne 3). C'est donc que la régularisation a un intérêt pour améliorer la qualité des *x-vectors*, même lorsqu'on dispose d'étiquettes de langues sur le domaine cible. D'autre part, pour ces *x-vectors* régularisés avec la *MMD*, un entraînement du classifieur final sur le domaine cible (ligne 5) améliore la performance de classification par rapport à un entraînement sur le domaine source (ligne 4). Les *x-vectors* ne sont donc pas totalement invariants entre les domaines et notre approche pourrait être combinée avec une adaptation du classifieur final.

5 Conclusion

Nous avons introduit une méthode d'adaptation de domaine non supervisée d'un réseau de neurones pour un système d'identification de la langue. Cette méthode consiste en une modification de la fonction de coût utilisée lors de l'entraînement du réseau de neurones par l'ajout d'un terme de régularisation. Par des expériences préliminaires avec un réseau de neurones convolutionnel, nous avons sélectionné une fonction de coût basée sur la *Maximum Mean Discrepancy*. Dans un second temps, nous avons appliqué cette approche à un système récent d'identification de la langue, constitué d'un extracteur de *features*, d'un extracteur de *x-vectors* et d'un classifieur final.

Les résultats démontrent l'efficacité de la méthode proposée pour prendre en compte la distorsion due au canal de transmission du signal. Lorsque la régularisation est appliquée au réseau x-vector, elle produit des vecteurs ayant acquis une robustesse au changement de domaine et permettant donc le transfert d'un apprentissage entre le domaine source et le domaine cible. De plus, la régularisation proposée améliore notablement la capacité de discrimination des *x-vectors* ainsi produits par rapport à un apprentissage supervisé. Elle est donc pertinente même dans le cas où on disposerait de données annotées sur le domaine cible.

Références

BEN-DAVID S., BLITZER J., CRAMMER K., KULESZA A., PEREIRA F. & VAUGHAN J. W. (2010). A theory of learning from different domains. In *Machine learning*, volume 79, p. 151–175 : Springer.

BHATTACHARYA G., ALAM J. & KENNY P. (2019). Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training. In *Proc. ICASSP*, p. 6041–6045.

BOUSQUET P.-M. & ROUVIER M. (2019). On robustness of unsupervised domain adaptation for speaker recognition. In *Proc. INTERSPEECH*, p. 2958–2962.

FER R., MATĚJKA P., GRÉZL F., PLCHOT O., VESELÝ K. & ČERNOCKÝ J. H. (2017). Multilingually trained bottleneck features in spoken language recognition. In *Computer Speech & Language*, volume 46, p. 252–267 : Elsevier.

FEYDY J., SÉJOURNÉ T., VIALARD F.-X., AMARI S.-I., TROUVÉ A. & PEYRÉ G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proc. The Twenty-second International Conference on Artificial Intelligence and Statistics*, p. 2681–2690.

GANIN Y., USTINOVA E., AJAKAN H., GERMAIN P., LAROCHELLE H., LAVIOLETTE F., MARCHAND M. & LEMPITSKY V. (2016). Domain-adversarial training of neural networks. In *The Journal of Machine Learning Research*, volume 17, p. 2096–2030.

HAN K. J., GANAPATHY S., LI M., OMAR M. K. & NARAYANAN S. (2013). TRAP language identification system for RATS phase II evaluation. In *Proc. INTERSPEECH*, p. 1502–1506.

LEI Y., FERRER L., LAWSON A., MCLAREN M. & SCHEFFER N. (2014). Application of convolutional neural networks to language identification in noisy conditions. In *Proc. Odyssey*, volume 41, p. 1–8.

LIN W.-W., MAK M.-W., LI L. & CHIEN J.-T. (2018). Reducing domain mismatch by maximum mean discrepancy based autoencoders. In *Proc. Odyssey*, p. 162–167.

LONG M., CAO Y., WANG J. & JORDAN M. I. (2015). Learning transferable features with deep adaptation networks. In *Proc. ICML 2015*, p. 97–105.

LOZANO-DIEZ A., ZAZO CANDIL R., GONZÁLEZ DOMÍNGUEZ J., TOLEDANO D. & GONZÁLEZ-RODRÍGUEZ J. (2015). An end-to-end approach to language identification in short utterances using convolutional neural networks. In *Proc. INTERSPEECH*, p. 403–407.

MATĚJKA P., ZHANG L., NG T., MALLIDI S. H., GLEMBEK O., MA J. & ZHANG B. (2014). Neural network bottleneck features for language identification. In *Proc. Odyssey*, p. 299–304.

NIST (2016). Evaluation plan for the NIST open evaluation of speech activity detection (OpenSAD15). In *www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation*.

PEYRÉ G. & CUTURI M. (2019). Computational optimal transport. In *Foundations and Trends® in Machine Learning*, volume 11, p. 355–607 : Now Publishers, Inc.

PLCHOT O., MATĚJKA P., NOVOTNÝ O., CUMANI S., LOZANO-DIEZ A., SLAVICEK J., DIEZ M., GRÉZL F., GLEMBEK O., MOUNIKA K. V., SILNOVA A., BURGET L., ONDEL L., KESIRAJU S. & ROHDIN J. (2018). Analysis of BUT-PT submission for NIST LRE 2017. In *Proc. Odyssey*, p. 47–53.

ROHDIN J., STAFYLAKIS T., SILNOVA A., ZEINALI H., BURGET L. & PLCHOT O. (2019). Speaker verification using end-to-end adversarial language adaptation. In *Proc. of ICASSP*, p. 6006–6010.

SNYDER D., GARCIA-ROMERO D., MCCREE A., SELL G., POVEY D. & KHUDANPUR S. (2018). Spoken language recognition using x-vectors. In *Proc. Odyssey*, p. 105–111.

SUN B. & SAENKO K. (2016). Deep CORAL : Correlation alignment for deep domain adaptation. In *Proc. ECCV 2016*, p. 443–450 : Springer.

WALKER K. & STRASSEL S. (2012). The RATS radio traffic collection system. In *Proc. Odyssey*, p. 291–297.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Modifications des flux aérodynamiques de la parole après chirurgie naso-sinusienne

Amélie Elmerich, Angélique Amelot, Lise Crevier-Buchman
Laboratoire de Phonétique et Phonologie
19 rue des Bernardins 75005 Paris
amelie.elmerich@sorbonne-nouvelle.fr, angelique.amelot@sorbonne-nouvelle.fr, lise.buchman1@gmail.com

RÉSUMÉ

Cette étude a pour but de déterminer dans quelle mesure la polypose naso-sinusienne impacte l'aérodynamique des flux oral et nasal. Ainsi, nous avons enregistré des patients atteints de cette pathologie avant et après chirurgie. Plusieurs éléments ont pu être mis en lumière : une modification du passage de l'air dans la cavité nasale et une meilleure coordination des flux d'air oral et nasal après chirurgie.

ABSTRACT

Speech aerodynamic airflow modifications after sinonasal surgery

The purpose of this study is to determine how much nasal polyposis impact the aerodynamics of the nasal and oral airflow. Thus, we recorded patients with this pathology before and after the surgery. Several elements have been highlighted: a modification of the airflow passage in the nasal cavity and a coordination of the nasal and oral airflow after surgery.

MOTS-CLÉS : Sinus, Nasalité, Polypose Naso-Sinusienne, Aérodynamique, Chirurgie.
KEYWORDS: Sinus, Nasality, Nasal Polyposis, Aerodynamics, Surgery

1 Introduction

1.1 Cavités nasales et sinusiennes

La cavité nasale fait partie des résonateurs de la parole. Elle permet la réalisation des phonèmes nasals par l'ouverture du port vélo-pharyngé. Acoustiquement, ce phénomène de nasalité s'observe par la présence d'anti-formants et aérodynamiquement par un débit d'air nasal positif. Autour de la cavité nasale, se trouvent des cavités remplies d'air appelées sinus. Quant à eux, leur rôle dans la parole est controversé. Certains auteurs (Lindqvist-Gauffin et Sunberg, 1976 ; Proctor, 1980 ; Maeda, 1982) les considèrent comme des résonateurs de la parole leur donnant la fonction de résonateur d'Helmholtz (Masuda, 1992) et pour d'autres (Bunch, 1992), les sinus n'ont pas de rôle dans la parole ou un rôle extra-linguistique (allègement de la masse crânienne, réchauffement de l'air inspiré etc.).

Volume 1 : Journées d'Études sur la Parole, pages 199–207.
HAL : hal-02798537.

1.2 La polypose naso-sinusienne

Il peut arriver que certaines pathologies comme la polypose naso-sinusienne (PNS), qui correspond à une inflammation chronique de la muqueuse tapissant les cavités nasales et sinusiennes et aboutissant à la formation de petites tumeurs bénignes (Bonfils *et al*., 2017), viennent perturber la résonance en obstruant les voies nasales. Il existe plusieurs stades de gravité selon le degré d'obstruction des cavités allant du stade I au stade IV (sévère). Cela va interférer sur la configuration et la circulation du flux d'air au sein des cavités en affectant la résonance mais aussi la qualité des sons nasals (Hong *et al.*, 1997). Il s'agit de la forme la plus sévère de sinusite chronique. Selon le degré de sévérité et en réponse à un traitement médicamenteux non efficace, une chirurgie des sinus peut s'avérer nécessaire. Celle-ci consiste à enlever les polypes pour permettre un passage conséquent de l'air mais aussi à modifier la structure anatomique des cavités naso-sinusiennes. En effet, la chirurgie augmente la communication entre cavités nasales et sinusiennes ce qui impacte la résonance nasale et la production des phonèmes nasals. Borel (2005) a montré qu'elle permettait d'augmenter les capacités de production des consonnes nasales. Giron et Mas (2016) ont montré que la différence majeure entre préopératoire et postopératoire se situait au niveau aérodynamique. Nous avons donc voulu approfondir ce niveau.

L'objectif de ce travail est de déterminer l'impact de la PNS mais aussi de la chirurgie sur le plan aérodynamique de la parole. Notre première hypothèse serait qu'il y a une modification du passage de l'air dans la cavité nasale dans le cadre d'une polypose naso-sinusienne c'est-à-dire un débit d'air nasal beaucoup moins important en préopératoire en raison de l'encombrement des cavités par les polypes. Et par compensation, un débit d'air oral plus important en préopératoire (Giron et Mas, 2016). L'air ayant des difficultés à traverser la cavité nasale en préopératoire, il se répartirait donc plus vers la cavité orale. Notre seconde hypothèse serait qu'il y a une meilleure coordination entre le flux d'air oral et nasal durant la production de la consonne nasale après une chirurgie des sinus. La coordination des flux peut se définir par une simultanéité de l'extinction du débit d'air oral et l'initialisation du débit d'air nasal en début de la réalisation de la consonne nasale.

2 Méthode

Ce travail cible 4 patients âgés de 48 à 63 ans (2 hommes et 2 femmes) atteints d'une polypose naso-sinusienne (Table 1). Ils ont été pris en charge au sein de l'Hôpital Européen Georges Pompidou à Paris en vue d'une chirurgie des sinus. Les patients recrutés devaient avoir un âge minimum de 20 ans, être francophone, présenter une obstruction de la cavité naso-sinusienne par polypes et opter pour l'intervention chirurgicale. Ils ont été enregistrés dans un environnement silencieux à l'hôpital la veille de leur chirurgie puis, 3 mois plus tard. Aux patients ont été appariés selon leur sexe 4 témoins. Nous avons fait un appariement par groupe féminin/masculin en raison de la grande variabilité anatomique des cavités naso-sinusiennes entre individus. La station d'Evaluation Vocale Assistée (EVA2TM, SQLab) nous a permis de recueillir les données aérodynamiques (débit d'air oral et nasal en litre/seconde (l/s)). Une calibration de l'appareil était réalisée avant chaque enregistrement, pour faire un ré-étalonnage. Une fois le patient installé, nous lui demandions de

respirer par le nez et la bouche afin de vérifier que l'appareil détectait les flux et que les embouts nasals étaient correctement positionnés. En effet, ils doivent être positionnés à l'entrée de chaque narine de manière verticale afin de suivre l'écoulement naturel du flux d'air nasal. Le débit d'air oral était recueilli à l'aide d'un masque en silicone souple. La segmentation et l'étiquetage des données en phonème et phrase ont été faits manuellement grâce au logiciel Praat à partir des enregistrements acoustiques obtenus à l'aide d'un microphone AKG C240 relié à la pièce à main de la station EVA2TM (Teston et Galindo, 1995). Notre corpus était composé de phrases et d'un texte contenant un nombre important de consonnes nasales ([m] et [n]). Nos 4 phrases provenaient du corpus AUPELF-UREF (Vaissière *et al.*, 1998) et le texte est une composition personnelle de Mme Hélène Villet. Les phrases alliaient consonnes nasales et les voyelles orales ([a] et [i]), par exemple : « Nana a nagé naguère à nadi », « Nini ne nie les nids ni les anis ». Le texte était composé de 6 phrases, par exemple : « La panique se lut dans les yeux de la jeune Catherine quand elle remarqua la magnifique tarentule qui traînait sur le meuble de cuisine. ». Ce corpus a l'avantage de présenter les consonnes nasales [m] et [n] dans des contextes variés : Voyelle-Consonne Nasale-Voyelle (VNV), Voyelle-Consonne Nasale-Consonne (VNC), Consonne-Consonne Nasale-Voyelle (CNV), Consonne-Consonne Nasale-Semi-Voyelle (CNSV), Consonne Nasale-Voyelle (NV). En somme, notre corpus était composé de 29 [m] et 40 [n] et a été produit une seule fois par chaque locuteur.

Patient	Sexe	Age	Diagnostic
M1	M	62	PNS Stade II
F1	F	48	PNS Stade II
M2	M	51	PNS Stade III Gauche PNS Stade II Droite
F2	F	63	PNS Stade IV

TABLE 1 : Tableau récapitulatif de nos 4 patients

3 Résultats

Les statistiques ont été réalisées grâce au logiciel R. Nous avons réalisé des analyses de variance (ANOVA). Le seuil de significativité a été considéré comme suit : $p<0,05$.

Les abréviations utilisées ci-dessous (NAF, OAF) correspondent respectivement à Nasal Air Flow (débit d'air nasal) et Oral Air Flow (débit d'air oral).

Notre première hypothèse s'est trouvée vérifiée. En effet, nous avons observé une augmentation significative du débit d'air nasal en postopératoire sur nos deux consonnes nasales [m] (p =3.72e-14) et [n] (p<2e-16) ainsi qu'une baisse significative du débit d'air oral sur la consonne [n] (p=0,0426).

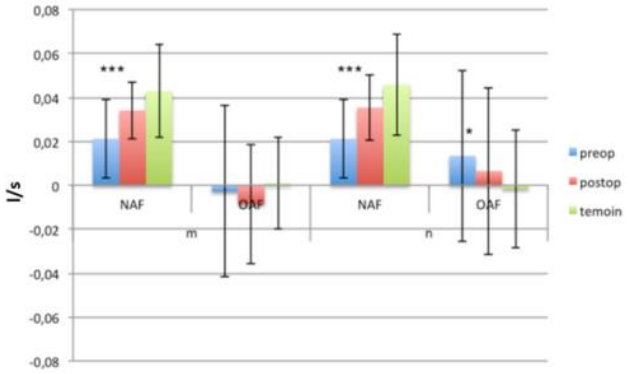


FIGURE 1 : Graphique présentant les moyennes en l/s de NAF et OAF sur [m] et [n] en préopératoire (bleu), postopératoire (rouge) pour les 4 patients et pour les témoins (vert)

Nous avons à présent souhaité observer ce résultat au cas par cas. En effet, nous avons des stades de gravité différents dans notre cohorte, il est intéressant de voir les degrés d'altérations selon le stade de gravité et le sexe.

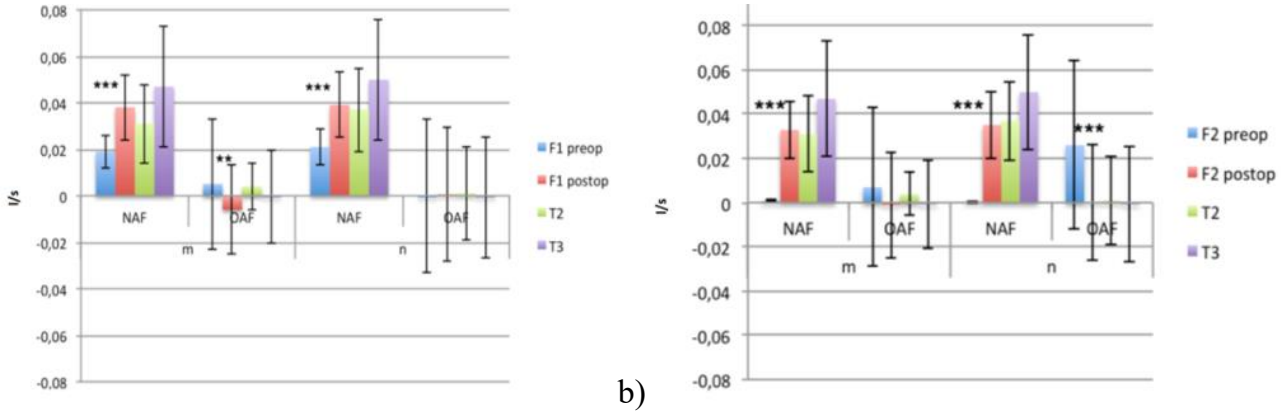


FIGURE 2 : a) Graphique des moyennes en l/s de NAF et OAF sur [m] et [n] en préopératoire (bleu), postopératoire (rouge) pour la patiente F1 et pour les témoins T2 et T3 (vert et violet), b) Graphique des moyennes en l/s de NAF et OAF sur [m] et [n] en préopératoire (bleu), postopératoire (rouge) pour la patiente F2 et pour les témoins T2 et T3 (vert et violet)

Pour la patiente F1 (Figure 2), l'augmentation du NAF est significative pour [m] (p=8.15e-14) et [n] (p=7.05e-16). Or, la baisse d'OAF n'est significative que pour [m] (p=0.00845). La patiente F2 (Figure 2) connaît une amélioration de son débit d'air nasal des plus importantes : passant de 0,0017 l/s pour [m] en préopératoire à 0,033 l/s en postopératoire. Il en va de même pour [n] : 0,0009 l/s à

0,035 l/s. Ces augmentations sont significatives (p<2e-16). Nous pouvons corréler ce constat au fait que c'était la patiente qui avait le stade de gravité le plus élevé (stade IV). De ce fait, les cavités naso-sinusiennes étaient complètement encombrées par les polypes. C'est aussi la patiente qui a le plus compensé oralement en préopératoire pour [m] : une moyenne de 0,007 l/s contre 0,005 l/s pour F1, -0,023 et -0,007 l/s pour M1 et M2. Mais aussi pour [n] : 0,026 l/s en préopératoire (environ 0,01 l/s pour M1 et M2, -0,0003 l/s pour F1). La baisse de l'OAF est significative en postopératoire sur [n], p= 4.12e-06. Pour les patients masculins M1 et M2 (Figure 3), les différences préopératoires et postopératoires relevées chez les patientes féminines F1 et F2, n'apparaissent pas de manière aussi marquée :

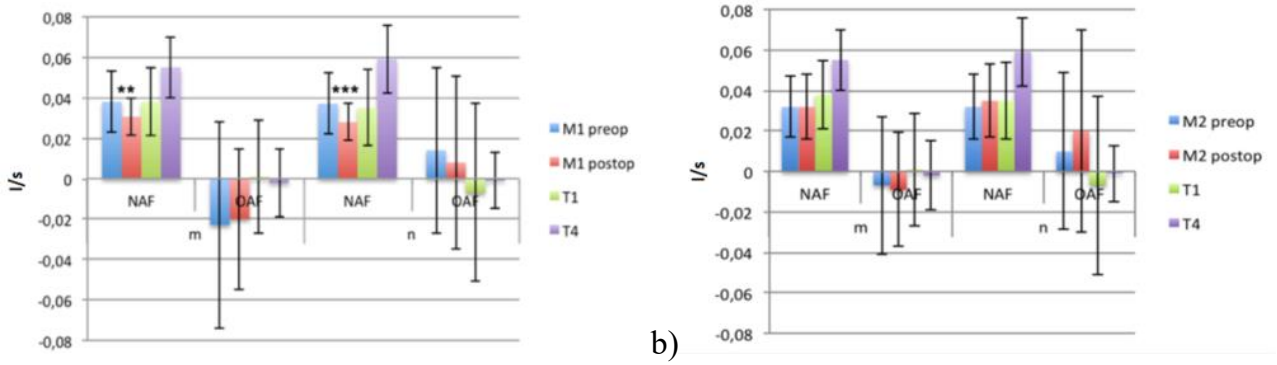


FIGURE 3 : a) Graphique des moyennes en l/s de NAF et OAF sur [m] et [n] en préopératoire (bleu), postopératoire (rouge) pour le patient M1 et pour les témoins T1 et T4 (vert et violet), b) Graphique des moyennes en l/s de NAF et OAF sur [m] et [n] en préopératoire (bleu), postopératoire (rouge) pour le patient M2 et pour les témoins T1 et T4 (vert et violet)

Concernant M1, son débit d'air nasal baisse en postopératoire sur les deux consonnes nasales : 0,030 l/s pour [m] et 0,028 l/s pour [n] contre respectivement 0,038 l/s et 0,037 l/s en préopératoire. Ces baisses sont significatives : [n] : p=0.000301 et [m] : p=0.00334. Concernant le débit d'air oral, il est quasiment identique pour [m] avant et après chirurgie et pour [n] il baisse légèrement (baisse de 0,006 l/s). Pour M2, le débit d'air nasal est identique sur [m] avant et après chirurgie et augmente légèrement pour [n] (+0,003 l/s). Au niveau du débit d'air oral, il diminue pour [m] (-0,007 l/s en préopératoire et -0,009 l/s en postopératoire) mais augmente pour [n] (+0,01 l/s).

Ce sont aussi nos deux locuteurs masculins qui ont le plus de débit d'air nasal en préopératoire : une moyenne de 0,038 l/s pour M1 et 0,032 l/s pour M2 contre 0,02 l/s pour F1 et une absence de débit d'air nasal pour F2. Nos locuteurs masculins vont à l'encontre de notre hypothèse c'est-à-dire qu'ils n'augmentent pas leur débit d'air nasal en postopératoire. Le facteur "sexe" pourrait être pris en compte (variables anatomiques notamment dues au volume des cavités nasales, par exemple) mais il est impossible de l'affirmer avec seulement 2 patients masculins de stade de gravité différente, et un groupe féminin hétérogène au niveau du stade de gravité. Il faudrait pouvoir observer cela sur une cohorte plus nombreuse de patients. La caractéristique articulatoire de la bilabiale [m] et de l'apico-alvéodentale [n] ne semblent pas avoir d'effet sur l'aérodynamique : il y a une certaine homogénéité

pour ces deux consonnes dans l'augmentation du débit d'air nasal (hormis pour M1 et M2) et la baisse du débit d'air oral.

Notre seconde hypothèse reposait dans le fait qu'une chirurgie des sinus entraînerait une meilleure coordination des flux oral et nasal. Nous envisageons ici la coordination dans une approche spatiale en tant que répartition des flux entre la sortie nasale et orale. Ce n'est pas le résultat d'un contrôle volontaire de la part du locuteur mais plutôt le résultat de la résistance ou non du passage de l'air dans une cavité nasale plus ou moins encombrée. Pour illustrer cette coordination, nous pouvons nous reporter à la figure 4 et plus particulièrement aux courbes NAF (4) et OAF (5) sur le segment [+Nasal]. Ce cas idéal de coordination des flux serait attendu chez les témoins et chez les patients en postopératoire.

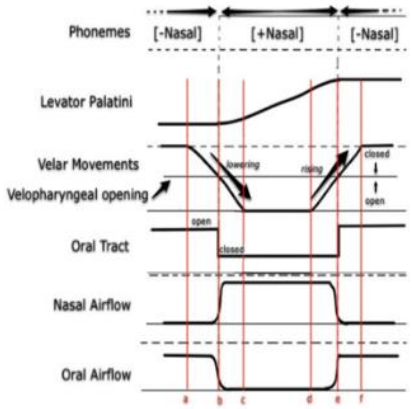


FIGURE 4 : Représentation schématique de la coordination aérodynamique et articulatoire de la réalisation d'une nasale (Clements *et al. (*2014))

Nous nous reporterons aux figures 1 à 3 présentes dans la première hypothèse. On peut constater grâce à la figure 1 que la coordination des flux se trouve améliorée : elle est meilleure pour [m] où le flux d'air oral est plus négatif (-0,0008 l/s en postopératoire contre -0,0002 l/s en préopératoire) mais ce dernier reste positif pour [n] mais tout de même à un volume moins important qu'en préopératoire (baisse de 0,007 l/s). Il serait attendu d'avoir un flux d'air oral nul au moment de la consonne nasale, la présence de flux d'air négatif peut refléter un léger flux d'air oral ingressif lié à l'abaissement du voile. Pour [m], la coordination était déjà visible en préopératoire, elle s'avère plus difficile pour [n]. Quant au flux d'air nasal, il est positif sur les consonnes nasales en préopératoire et postopératoire. Enfin, nous avons voulu observer si la coordination était plus ou moins aggravée selon le stade de gravité. En effet, un stade de gravité élevé (par exemple, stade IV) pourrait impacter de manière plus importante la coordination qu'un stade moyen (stade II à III). En préopératoire, la patiente F2 (Figure 2) a du mal à coordonner les flux oral et nasal : au niveau du flux d'air nasal, on trouve très peu d'air au moment de la consonne nasale (une moyenne générale de 0,0009 l/s pour les consonnes nasales) dû à l'encombrement des cavités naso-sinusiennes. Au niveau du flux d'air oral, la moyenne est très élevée sur les consonnes nasales (0,007 l/s pour [m] et 0,026 l/s pour [n]). En postopératoire, on observe une nette amélioration avec un flux nasal de 0,033 l/s

pour [m] et 0,035 l/s pour [n]. Par ailleurs, le flux d'air oral est fortement réduit. Le schéma idéal de la coordination se retrouve en postopératoire. Pour les patients M1 et M2 (Figure 3), la coordination se réalise en préopératoire notamment de manière idéale pour [m], [n] présentant un flux d'air oral positif. En postopératoire, la coordination du flux d'air oral reste toujours problématique pour [n] : la moyenne augmente même pour M2 de 0,01 l/s. Quant à F1 (Figure 2), c'est le phonème [m] au niveau du flux d'air oral qui s'est le plus amélioré en postopératoire (0,005 l/s en préopératoire et -0,006 l/s en postopératoire) puisqu'il devient négatif au moment de la consonne nasale.

4 Discussion

Notre première hypothèse se trouve en partie vérifiée. En effet, l'augmentation du débit d'air nasal est significative sur les consonnes nasales après chirurgie mais la baisse du flux d'air oral ne l'est que pour [n]. Cependant pour dresser des généralités sur la différence aérodynamique entre le préopératoire et le postopératoire, il nous faudrait plus de données. L'augmentation du NAF, en plus d'être liée à l'opération peut aussi traduire un volume d'air expiré plus important dans le cas d'un accent d'insistance par exemple. Ladefoged (1963) et Van Hattum (1952, cité par Van Hattum, 1967) ont montré qu'il pouvait y avoir des variations importantes de débit d'air expiré dans le cas où l'intensité, la hauteur de la voix et la durée n'étaient pas contrôlées, ce qui est le cas dans cette étude. En regardant au cas par cas, des différences ont été relevées : les patients F1, F2 et M1 ont baissé leur débit d'air oral sauf M2. F1, F2 et M2 ont augmenté leur débit d'air nasal sauf M1. Nos résultats ont convergé avec l'étude de Giron et Mas (2016). En effet, nous avons pu observer un débit d'air nasal beaucoup plus important dû à l'ablation des polypes. Cette augmentation s'est beaucoup plus remarquée chez la patiente F2, qui avait le stade de gravité le plus élevé. De ce fait l'augmentation de son débit d'air nasal est beaucoup plus saillante car c'est elle qui a le moins de débit d'air nasal en préopératoire et donc progresse le mieux en postopératoire. On peut mettre en relation cette augmentation du flux d'air nasal avec la nasalance. Elle est liée à l'obstruction du flux d'air nasal. Des auteurs comme Hong *et al.* (1997), Soneghet *et al.* (2002) ont montré une augmentation du score de nasalance en postopératoire. Nous n'avons pas utilisé cette mesure dans notre étude mais ces résultats rejoignent les nôtres. Le faible débit d'air nasal chez les témoins pourrait s'expliquer par la méthode d'enregistrement non systématisée. Les contraintes techniques devraient trouver des améliorations à l'avenir afin d'obtenir des données consistantes. En effet, le protocole d'enregistrement (placement de l'embout dans la narine) doit être réalisé de la même manière à chaque enregistrement. Ces données étant acquises par une tierce personne il est difficile de pouvoir s'assurer que toutes les précautions pour les enregistrements ont été effectuées. Nous avons pu remarquer que c'est plutôt dans un premier temps, le stade de gravité que le sexe qui joue un rôle. Constat qui rejoint celui de Thompson et Hixon (1979), les différences anatomiques, structurelles et fonctionnelles, entre hommes et femmes sont avérées, mais les conséquences sur la variabilité du débit d’air nasal sont peu significatives sur des tâches de parole.

Grâce à notre seconde hypothèse, nous cherchions à connaître l'impact de la modification du passage de l'air durant la production de phonèmes nasals. L'amélioration générale de la coordination des flux met en lumière une augmentation des capacités de production des consonnes nasales, ce qui

rejoint la conclusion émise par Borel (2005). Cependant, il convient de rester prudent face à certaines mesures de flux qui sont à un niveau très faible, il peut être délicat d'interpréter des flux de ce niveau. La question de la coordination pourrait être orientée aussi vers la coarticulation nasale c'est-à-dire l'anticipation ou la propagation de la nasalisation de la consonne nasale sur les voyelles adjacentes. En comparant les voyelles orales entourant une consonne nasale et celles entourant une consonne orale, nous pourrions voir la différence entre les deux du point de vue de la moyenne de flux d'air nasal et du pourcentage nasalisé de la voyelle, et ainsi voir la potentielle amélioration en postopératoire.

5 Conclusion

Cette étude avait pour ambition de se demander quel serait l'impact d'une pathologie des sinus et de sa chirurgie sur le niveau aérodynamique de la parole. Nous pouvons à présent répondre qu'elle a permis, pour certains patients, d'augmenter le flux d'air nasal, de réduire le phénomène de compensation orale mais aussi de faciliter la coordination des consonnes nasales.

6 Perspectives

Ce travail pourrait être enrichi en augmentant notre cohorte de patients en ayant des groupes conséquents avec des stades de gravité similaires. Mais aussi en adoptant une approche multiparamétrique de la problématique. Aborder le côté articulatoire en utilisant l'imagerie médicale serait une piste innovante : cela nous permettrait de rendre compte de la variabilité interindividuelle du volume des sinus et fosses nasales en lien avec le phénomène de résonance nasale. Au niveau de la perception, il serait intéressant de voir comment est caractérisée la voix des patients en préopératoire et postopératoire. Il est donc primordial d'approfondir les problématiques liées à cette pathologie : la littérature scientifique actuelle ne permet pas de répondre aux questionnements relatifs à la voix des patients. Très peu d'études se rapportant à la parole se sont souciées jusqu'à présent de l'impact d'une PNS sur la parole. En outre, un tel approfondissement permettrait aussi d'en apprendre plus sur le phénomène de nasalité qui reste encore méconnu de nos jours et sur la contribution ou non des sinus dans la parole. Nous n'avons pas de modèles qui permettent d'envisager les répercussions possibles d'une telle chirurgie sur la voix des patients.

Remerciements

Ce travail est soutenu par le Labex EFL (ANR-10-LABX-0083). Nous remercions les patients qui ont accepté de participer aux enregistrements avant et après leur chirurgie. Ainsi que les relecteurs anonymes pour leurs précieux commentaires.

Références

BONFILS, P., HALIMI, P., GAULTIER, A.-L. ET LISAN,Q. (2017). Polypose nasosinusienne. Rhinosinusite chronique avec polypes. *EMC Oto-Rhino-Laryngologie*, 12(2):1-21. DOI : 10.1016/S0246-0351(16)76892-9.

BOREL, S. (2005). *Analyse perceptive et acoustique des consonnes nasales dans la polypose nasosinusienne avant et après chirurgie des sinus*. Mémoire de l'Université de la Sorbonne nouvelle-Paris III.

BUNCH, MA. (1992) *Dynamics of the Singing Voice.* Vienna: Springer-Verlag.

CLEMENTS, N. (2014). The feature nasal. Dans Rialland et al. (dir), *Features in Phonology and Phonetics : Posthumous Writings by Nick Clements and Coauthors* (p.195-217). Berlin, Allemagne : De Gruyter Mouton

GIRON, M. ET MAS, B. (2016). *Evaluation de la qualité vocale avant et après chirurgie naso-sinusienne*. Mémoire d'orthophonie de l'Université Paris VI Pierre et Marie Curie.

HONG, K. H., KWON, S. H. ET JUNG, S. S. (1997). The Assessment of Nasality with a Nasometer and Sound Spectrography in Patients with Nasal Polyposis. *Otolaryngology-Head and Neck Surgery*,117(4), 343-348. DOI : 10.1016/S0194-5998(97)70124-4.

LADEFOGED, P. (1963). Some physiological parameters in speech. *Language and speech*, *6*(3), 109-119. DOI : 10.1177/002383096300600301

LINDQVIST-GAUFFIN, J. ET SUNBERG, J. (1976). Acoustic properties of the nasal tract. *Phonetica*, 33 (3):161-8.DOI:10.1159/000259720.

MAEDA, S. (1982). The role of the sinus cavities in the production of nasal vowels. In ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing, 7, 911-914. DOI : 10.1109/ICASSP.1982.1171561

MASUDA, S. (1992). Role of the maxillary sinus as a resonant cavity. *Nihon Jibiinkoka Gakkai Kaiho*, 95(1), 71-80.

PROCTOR DF. (1980) *Breathing, Speech and Song.* Vienna: Springer-Verlag.

SONEGHET, R., SANTOS, R. P., BEHLAU, M., HABERMANN, W., FRIEDRICH, G. ET STAMMBERGER, H. (2002). Nasalance changes after functional endoscopic sinus surgery. *Journal of Voice*, 16(3), 392-397. DOI : 10.1016/S0892-1997(02)00110-8

TESTON, B., et GALINDO, B. (1995). A diagnostic and rehabilitation aid workstation for speech and voice pathologies. In *Fourth European Conference on Speech Communication and Technology*.

THOMPSON, A. E. ET HIXON, T. J. (1979). Nasal air flow during normal speech production. *Cleft Palate Journal*, 16, 412-420. PMID : 290432

VAISSIERE, J., BASSET, P., SU, T., AMELOT, A., CORBIN, O. ET MICHAUD, A., (1998). Corpus AUPELF-UREF. Récupéré sur la plateforme COCOON, <http://purl.org/poi/crdo.vjf.cnrs.fr/cocoon-e61f60b2-f20c-3b62-81ba-b286f1fb8de4>.

VAN HATTUM, R. J., & WORTH, J. H. (1967). Air flow rates in normal speakers. *The Cleft palate journal*, *4*(2), 137-147.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole

Solène Evain [1] Adrien Contesse [2] Antoine Pinchaud [3] Didier Schwab [1]
Benjamin Lecouteux [1] Nathalie Henrich Bernardoni [4]

(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) http://www.vocalgrammatics.fr/
(3) ÉSAD Amiens, De-sign-e Lab, 80080 Amiens, France
(4) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

solene.evain@univ-grenoble-alpes.fr, AdrienContesse@gmail.com,
APinchaud@gmail.com,Didier.Schwab@imag.fr, Benjamin.Lecouteux@imag.fr,
nathalie.henrich@gipsa-lab.fr

RÉSUMÉ

Le *human-beatbox* est un art vocal utilisant les organes de la parole pour produire des sons percussifs et imiter les instruments de musique. La classification des sons du beatbox représente actuellement un défi. Nous proposons un système de reconnaissance des sons de beatbox s'inspirant de la reconnaissance automatique de la parole. Nous nous appuyons sur la boîte à outils Kaldi, qui est très utilisée dans le cadre de la reconnaissance automatique de la parole (RAP). Notre corpus est composé de sons isolés produits par deux beatboxers et se compose de 80 sons différents. Nous nous sommes concentrés sur le décodage avec des modèles acoustiques monophones, à base de HMM-GMM. La transcription utilisée s'appuie sur un système d'écriture spécifique aux beatboxers, appelé Vocal Grammatics (VG). Ce système d'écriture s'appuie sur les concepts de la phonétique articulatoire.

ABSTRACT

BEATBOX SOUNDS RECOGNITION USING A SPEECH-DEDICATED HMM-GMM BASED SYSTEM [1]

Human beatboxing is a vocal art making use of speech organs to produce percussive sounds and imitate musical instruments. Beatbox sounds classification is a current challenge. We propose a beatbox sounds recognition system with an adaptation of the Kaldi toolbox, widely used for automatic speech recognition (ASR). Our corpus is composed of isolated sounds produced by two beatboxers and is composed of 80 different sounds. We focused on decoding with monophones acoustic models, trained with a HMM-GMM model. One type of transcription was used : a beatbox specific writing system named Vocal Grammatics (VG) which uses concepts of articulatory phonetics.

MOTS-CLÉS : human-beatbox, reconnaissance automatique de la parole, Kaldi, reconnaissance de sons isolés.

KEYWORDS: Human beatbox, automatic speech recognition, Kaldi, isolated sounds recognition.

1. Cet article a été publié en anglais dans le Workshop international MAVEBA http://maveba.dinfo.unifi.it/

Volume 1 : Journées d'Études sur la Parole, pages 208–216.
HAL : hal-02798538.

1 Introduction

Le *human-beatbox* est apparu durant les années 80, dans le Bronx, un quartier de New York, et est associé à la culture hip-hop. Il consiste à produire des percussions vocales ainsi que des imitations d'instruments de musique, comme la trompette ou la guitare. La classification des sons de Beatbox peut être utilisée pour la recherche d'information musicale, comme une requête, pour la recherche de différents types de musique (Kapur *et al.*, 2004) ou pour des applications à commande-vocale avec un nombre de classes défini par l'utilisateur (Hipke *et al.*, 2014) afin de composer des morceaux de beatbox. Le beatbox est aussi utilisé dans le cadre de rééduction orthophonique : un système de reconnaissance peut permettre de travailler les exercices. Dans la littérature, des taux de classification corrects ont été obtenus sur un éventail limité de classes, c'est-à-dire cinq principaux sons de beatbox : *bass drum*, *open hi-hat*, *closed hi-hat*, *k-snare* et *p-snare* (Sinyor *et al.*, 2005). À notre connaissance, la reconnaissance automatique des sons de beatbox à l'aide d'un système de reconnaissance vocale n'a été explorée que par (Picart *et al.*, 2015). Leur base de données se compose de 5 sons percussifs de beatbox : *cymbal*, *hi-hat*, *kick*, *rimshot* , *snare* et 8 imitations d'instruments. Les performances étaient faibles pour les imitations d'instruments (taux d'erreur de reconnaissance de 41 %), mais plutôt correctes pour les classes se limitant aux sons percussifs (taux d'erreur de reconnaissance de 9 %).

En orientant nos efforts vers le développement d'un système de reconnaissance automatique des sons de beatbox efficace et fiable, nous visons à étendre le nombre de classes de sons et à permettre la reconnaissance de variantes subtiles dans la production de sons de beatbox. Nous considérons le *human-beatbox* comme un langage musical composé d'unités sonores que nous appellerons *boxèmes* en référence aux phonèmes de la parole. Par ailleurs, ce travail a été réalisé dans le but de créer un dispositif artistique interactif qui fournit des retours visuels lors de la production de sons de beatbox.

Le document est structuré comme suit : La section II présente la base de données. Le système de reconnaissance est présenté dans la section III. Différentes expériences sont décrites dans la section IV et leurs résultats sont donnés dans la section V. Les sections VI et VII présentent une discussion puis une conclusion, ainsi que les perspectives de nos travaux. Par ailleurs, ce travail a été présenté lors du workshop international MAVEBA (Evain *et al.*, 2019).

2 Corpus et matériel utilisé

Notre corpus de sons de beatbox appelé beatbox-VG2019 a été enregistré par deux beatboxeurs masculins : un beatboxeur professionnel (troisième auteur, nom de scène *Andro*) et un amateur (deuxième auteur). Il est composé de 80 boxèmes et peut être considéré comme un vaste corpus par rapport aux corpus précédemment présentés dans la littérature (et utilisés pour de la classification). Seuls les sons isolés sont considérés dans nos travaux, les séquences rythmiques étant écartées dans un premier temps.

Un système d'écriture pictographique basé sur l'articulation, développé par le deuxième auteur et s'appelant *Vocal Grammatics* (Contesse & Pinchaud, 2019) a été utilisé pour l'annotation. Dans cette écriture, les glyphes sont composés de deux informations : l'une sur les organes de la parole

Microphone	Distance de la bouche	Spécifications
Brauner VM1 (braun)	10 cm	condensateur + filtre pop
DPA 4006 (ambia)	50 cm	condensateur, micro d'ambiance
DPA 4060 (tie)	10 cm	condensateur
Shure SM58 (sm58p)	10 cm	dynamique
Shure SM58 (sm58l)	15 cm	dynamique
Shure beta 58 (beta)	1 cm	dynamique + encapsulé

TABLE 1: Récapitulatif des différents microphones

utilisés, l'autre sur la manière dont les sons sont produits (plosives, fricatives...). La Figure 1 illustre ce système d'écriture dans le cas d'un son plosif bilabial avec un glyphe morphologique représentant deux lèvres et un glyphe symbolique en forme de croix représentant la plosion.

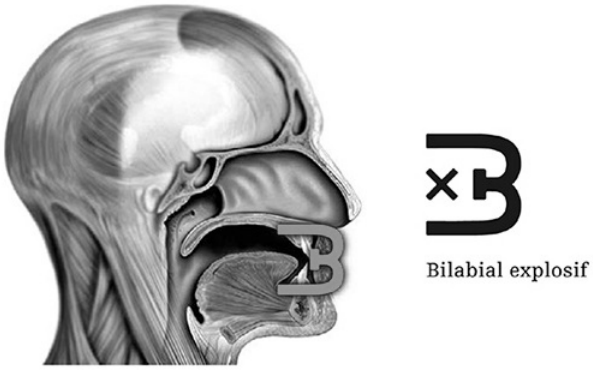


FIGURE 1: Représentation d'un son plosif bilabial avec le système d'écriture *Vocal Grammatics*.

Notre corpus de *boxèmes* a été enregistré avec six microphones. Cinq d'entre eux enregistraient simultanément et le dernier était encapsulé (une ou deux mains recouvrent la capsule du microphone). Les microphones différaient en termes de spécificités (par exemple, à condensateur ou dynamique) et de placement. Le tableau 1 donne les détails des microphones alors que le tableau 2 récapitule la composition du corpus.

L'apprentissage des modèles acoustiques a été réalisé avec la boîte à outils Kaldi (Povey *et al.*, 2011). En ce qui concerne les données de test, il s'agit de répétitions de différents sons de beatbox (pas toujours les même à la suite), dont la production est relativement lente (on peut percevoir une légère pause entre chaque son). Dans ce cadre de test, l'utilisation d'un système de reconnaissance automatique de la parole continue révèle son intérêt.

3 Reconnaissance du beatbox

Notre approche part de l'hypothèse que le *human-beatbox* est structuré comme un langage musical, utilisant les organes de la parole pour produire des unités sonores qui peuvent être distinguées les unes

Beatboxers	Adrien (amateur), Andro (professionnel)
Nom du corpus	beatbox-VG2019
Nombre de boxèmes (= taille du vocabulaire)	80
Nombre de boxèmes par beatboxeur	Adrien : 56/80 Andro : 80/80
Transcription	Vocal Grammatics
Microphone	5 simultanés + 1 encapsulé
Fréquence d'échantillonnage et précision	44100 Hz, 16 bits, mono
Durée totale d'enregistrement	∼206 min
Apprentissage	
Durée d'enregistrement	∼92 min
Nombre de répétitions des boxèmes	6 ou 2
Test	
Durée d'enregistrement	∼114 min
Nombre de répétitions des boxèmes	7 en moyenne

TABLE 2: Caractéristiques du corpus beatbox-VG2019

des autres et qui ont chacune une signification musicale spécifique pour le beatboxer. Dans ce contexte, un système de reconnaissance initialement dédié à la parole pourrait permettre de reconnaître les productions du beatbox. Généralement, dans le cadre de la reconnaissance automatique de la parole, les mots sont décomposés en unités (phonèmes, syllabes etc.) qui permettent de définir un lexique associant chaque mot à sa représentation sous forme d'unités atomiques. Les modèles acoustiques sont alors entraînés pour reconnaître ces unités.

Dans ce travail préliminaire, nous avons considéré chaque son de beatbox comme étant atomique : nous partons sur une approche de reconnaissance de mots isolés. La co-articulation ou la frontière extra-boxèmes ont été écartées, tout en conservant les contraintes de traitement du bruit, de variabilité intra et inter-locuteur. À terme nous souhaitons travailler sur des sous-unités au niveau des sons, mais nous manquons encore de données pour généraliser ces sous-unités.

Les paramètres utilisés sont de type MFCC. Ces paramètres se basent sur le système auditif périphérique humain (Tiwari, 2010) et sont largement utilisées dans les systèmes de reconnaissance automatique de la parole. Chaque son de beatbox a été associé à un modèle de Markov caché (HMM). Nous nous sommes limités à une approche HMM-GMM car les quantités de données d'apprentissage sont très faible (quelques dizaines de minute) et la parole beatboxée est tellement spécifique qu'il nous semblait difficile d'exploiter des méthodes neuronales (par apprentissage direct ou même par transfert d'apprentissage).

Notre objectif est avant tout d'appliquer les principes d'un système de reconnaissance de la parole continue avec la constitution d'unités acoustiques (nos boxèmes), d'un lexique (nos 80 sons pour l'instant) et d'un modèle de langage (qui n'est pas traité ici, mais qui représenterait la rythmique des séquences beatboxées).

Dans nos expériences, l'apprentissage a été réalisé avec une sélection du nombre de Gaussiennes automatique en fonction de la quantité de données (cependant, nous avons essayé différentes quantités de Gaussiennes, sans observer le moindre impact). Par ailleurs, au niveau du modèle de langage,

la probabilité d'émission d'un boxème est identique pour chacun d'eux étant donné que dans ces expériences préliminaires nous ne prenons pas en compte les séquences. Le système est donc capable de reconnaître de manière continue les boxèmes produits sans avoir de connaissance *a priori* sur la rythmique.

4 Méthode

Plusieurs systèmes ont été conçus dans le but de tester divers paramètres. L'influence dans l'apprentissage de chaque microphone avec différents placements et sensibilités a été étudiée afin de savoir si tous les enregistrements pouvaient être utilisés ensemble pour former un système plus robuste. Pour chaque microphone, nous avons découpé les enregistrements en une base d'apprentissage (6 répétitions de boxèmes) et une base de test (1 à 12 répétitions de boxèmes). Les résultats nous permettent de classer les microphones du plus efficace au moins efficace.

Pour cette première étude, nous avons souhaité nous concentrer sur les boxèmes produits de façon non-encapsulée. Le fait d'encapsuler le microphone (le recouvrir de la main) modifie le résultat sonore pour un boxème donné. Comme cinq des six microphones étudiés ont été utilisés de façon non-encapsulée par les beatboxeurs, le nombre d'enregistrements disponibles est plus conséquent que celui des boxèmes produits de façon encapsulés.

L'impact de différents paramètres sur la reconnaissance a été testé : le nombre d'états des HMM, les paramètres MFCC, la probabilité d'apparition d'un silence, ainsi que l'ajout ou non d'un phonème de pause dans le lexique . Certains choix ont été basés sur l'article de (Picart *et al.*, 2015). Nous présentons ici les résultats pour quatre configurations du système de reconnaissance :

- configuration A : les quatre paramètres cités ci-dessus sont par défaut, à savoir 3 états HMM, 13 paramètres MFCC, la probabilité d'apparition d'un silence à 0.5 et l'absence de phonème pause dans le lexique ;
- configuration B : une pause a été ajoutée dans le lexique et la probabilité d'apparition d'un silence est fixée à 0.8 ;
- configuration C : même base que la configuration B. Le nombre de coefficients MFCC passe à 22 ;
- configuration D : même base que la configuration B. Le nombre d'états HMM passe à 5.

Le lexique d'un système de reconnaissance de la parole est de la forme 'mot : transcription phonétique'. Ici, le mot est un boxème. La pause indiquée dans les systèmes ci-dessus a été ajoutée dans le lexique de la façon suivante : 'boxème : pause transcription_phonétique pause'. Cette pause n'est pas présente dans la transcription manuelle des corpus de test et n'est pas présente dans l'hypothèse de décodage. Les systèmes B, C et D sont donc comparables au système A puisque la valeur de dénominateur du BER est la même.

La mesure d'évaluation 'BER' -*Boxeme Error Rate*- est utilisée pour évaluer le système. Elle est directement inspirée du taux d'erreur sur les mots (WER) puisqu'il est calculé en additionnant le nombre de substitutions, d'insertions et de suppressions divisé par le nombre de boxèmes dans la référence. Plus la reconnaissance est bonne, plus la valeur du BER est faible. Le CBR (taux de boxèmes correct) est également utilisé en table 3 afin d'avoir une deuxième mesure de l'efficacité du système. Il indique le pourcentage de boxèmes correctement reconnus.

5 Résultats

Les figures 2 et 3 donnent le BER pour différents décodages. La ligne "but" sur l'axe horizontal représente notre objectif : obtenir un BER de 10 % ou moins, *a priori* fixé pour garantir une utilisation intéressante de notre système par le public.

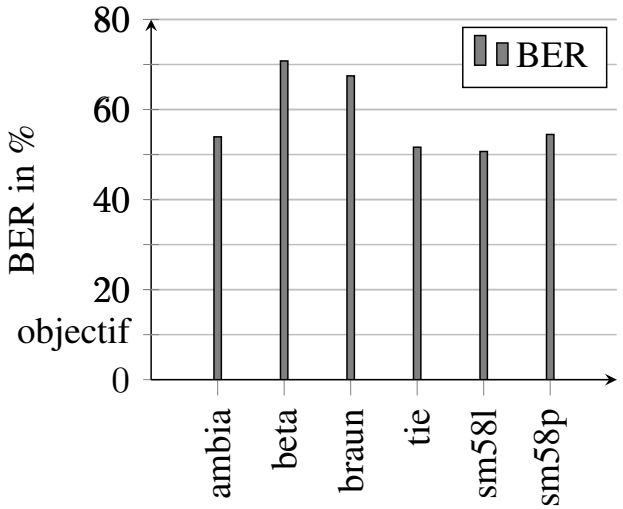


FIGURE 2: BER obtenu avec des modèles acoustiques monophones pour les six microphones

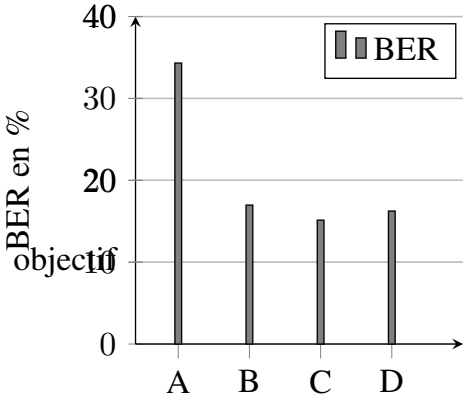


FIGURE 3: Évolution du BER pour les configurations A, B, C et D du système
A : par défaut / **B** : prob. silence=0.8 + pause / **C** : B + MFCC=22 / **D** : B + états HMM=5

En figure 2 nous pouvons observer l'efficacité de chacun des six microphones. Nous avons constaté que les microphones à condensateur DPA et les microphones dynamiques Shure SM58, placés à proximité ou loin de la bouche du beatboxer, offrent des performances similaires. Des taux de reconnaissance plus faibles sont constatés pour les enregistrements avec le microphone dynamique Shure beta 58 encapsulé et le microphone à condensateur Brauner VM1.

Nous avons ensuite voulu tester l'incidence de certains paramètres sur la reconnaissance. Nous avons sélectionné un microphone pour ces tests : le Shure SM58 - 'sm58p' - utilisé proche de la bouche. Notre choix s'est porté sur celui-ci particulier car c'est un des microphones le plus utilisé par les beatboxeurs à l'échelle mondiale et que l'utilisation du microphone dans cette discipline est le plus souvent effectuée proche de la bouche. Un premier test a été de faire varier la probabilité de silence de 0,5 (par défaut) à 0,9 avec un pas fixé à 0,1. Notre meilleur modèle a été obtenu avec une probabilité de silence de 0,9, ce qui a donné un BER de 26,94 %. En spécifiant en plus un phonème de "pause" avant et après chaque boxème dans le lexique, nos meilleurs résultats sont de 16,97% de BER. Ceux-ci sont obtenus avec une probabilité de silence de 0,8. Cela s'explique par la configuration de nos expériences : nous avons demandé aux beatboxers de répéter un même son ; des pauses ont donc artificiellement été ajoutées en début ou fin de chaque son.

La figure 3 montre l'évolution des résultats avec différents paramètres : probabilité de silence plus élevée, ajout d'une pause dans le lexique, 22 paramètres MFCC au lieu de 13 par défaut et 5 états par HMM au lieu de 3 par défaut. Pour rappel, la configuration A est représentative d'un système avec les quatre paramètres laissés 'par défaut'. L'ensemble d'apprentissage a été réalisé avec des enregistrements de microphones non encapsulés.

Notre meilleur modèle est obtenu avec la configuration C et donne un BER de 15,13 %. Les configurations B et D sont très proches avec des BER de 16,97% et 16,24% respectivement (voir tableau 3 pour les détails concernant les substitutions, les insertions, les suppressions et les taux de boxèmes corrects).

Dans la figure 3 et le tableau 3, nous observons que chaque changement de paramétrage est bénéfique pour les substitutions, les insertions, les suppressions et les taux de boxèmes corrects. Le bénéfice le plus évident est pour le taux d'insertion qui passe à zéro. Le taux de boxème correct atteint 85% avec la configuration C.

	A	**B**	**C**	**D**
Substitutions	19.19%	12.73%	10.70%	12.36%
Insertions	9.41%	0.18%	0.18%	0%
Deletion	5.72%	4.06%	4.24%	3.87%
CBR	75.09%	83.21%	85.06%	83.76%

TABLE 3: Insertions, substitutions, suppressions et taux de boxème correct (CBR) pour les configurations A B C D
A : par défaut / **B** : probabilité de silence 0.8 + pause / **C** : B + 22 MFCC / **D** : B + HMM à 5 états

6 Discussion

Comme nous l'avons vu précédemment, l'impact des différents microphones est assez faible, à l'exception des microphones Shure beta 58 et Brauner VM1 qui sont moins performants. Nous supposons que c'est à cause de la façon dont nous les avons utilisés (proximité par rapport au locuteur). En effet, le microphone Shure beta 58 est encapsulé et cette utilisation affecte les performances du microphone. Quant au microphone à condensateur Brauner VM1, nous pouvons observer qu'il fonctionne moins bien que l'autre microphone à condensateur de notre test (DPA 4060) et supposons qu'il a été placé trop près de la bouche du beatboxer. Enfin, ni le nombre de paramètres MFCC ni le nombre d'états dans le HMM n'apportent une nette amélioration. Nous supposions qu'augmenter le nombre d'états HMM était intéressant pour les sons complexes qui sont composés de deux ou plusieurs boxèmes. Ces aspects seront analysés dans des études ultérieures.

7 Conclusion et perspectives

Notre approche démontre qu'utiliser un système de reconnaissance vocale pour reconnaître les sons de beatbox isolés est pertinent. Cela ouvre des perspectives pour la reconnaissance vocale de phrases beatboxées.

Jusqu'à présent, notre meilleur modèle a été obtenu avec une augmentation de la probabilité de silence (0,8 au lieu de 0,5), l'insertion d'un phonème de silence "pause" étant ajouté dans les contextes droits et gauche du vocabulaire et 22 paramètres pour les MFCC. Le meilleur BER obtenu est alors de 15,13%.

Nous avons pu observer que le type de microphone utilisé pour l'enregistrement ne semble pas avoir

d'influence sur le système. Il dépend plutôt de leur utilisation (encapsulé ou non). Mettre de côté le microphone encapsulé pour l'apprentissage donne de meilleurs résultats.

Quant aux différents types de production, lorsqu'ils sont mélangés, ils semblent dégrader fortement les performances. Pour l'instant, en ce qui concerne les substitutions, nous ne pouvons rien conclure car le système semble mélanger des sons qui sont assez semblables à l'oreille ou qui ont une articulation assez similaire, et des sons qui sont très différents. Nous supposons que la division du corpus en fonction de la longueur du son et l'adaptation du nombre d'états HMM pourraient améliorer le système.

Diviser chaque son en plus petits morceaux, comme on le fait pour les langues comportant des phonèmes ou des syllabes, est une perspective. En effet, à mesure que le vocabulaire du corpus augmentera, nous serons confrontés à un manque d'exemples pour l'apprentissage. Le fait de disposer d'un modèle basé sur des boxèmes réduirait le nombre de modèles nécessaires au système et permettrait le traitement de la coarticulation. De plus, il reste à explorer les séquences rythmiques (que nous pourions apparenter à un modèle de langage) et la reconnaissance des sons encapsulés. Enfin, il serait intéressant de voir si la reconnaissance des voix des femmes ou des enfants pose des problèmes dans le cadre des sons de beatbox.

Des perspectives plus techniques visent à résoudre le fait que les données annotées de beatbox sont pour l'instant très précieuses et rares. En effet, dans les expériences décrites nos ensembles d'apprentissage ne représentent tout au plus que quelques dizaines de minutes. Pour cette raison nous sommes restés concentrés sur des modèles de type HMM-GMM qui sont moins gourmands en données que des modèles à base de réseaux de neurones profonds : nous envisageons d'exploiter des techniques d'augmentation de données, de synthèse de données et l'utilisation d'outils non supervisés tels que wav2vec.

Références

CONTESSE A. & PINCHAUD A. (2019). *vocal grammatics*. Web page, www.vocalgrammatics.fr, Last consulted : 2019-08-29.

EVAIN S., CONTESSE A., PINCHAUD A., SCHWAB D., LECOUTEUX B. & HENRICH BERNARDONI N. (2019). Beatbox sounds recognition using a speech-dedicated hmm-gmm based system.

HIPKE K., TOOMIM M., FIEBRINK R. & FOGARTY J. (2014). BeatBox : End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations. p. 121–124, Como, Italy : ACM.

KAPUR A., TZANETAKIS G. & BENNING M. (2004). Query-by-Beat-Boxing : Music Retrieval For The DJ. Barcelona, Spain.

PICART B., BROGNAUX S. & DUPONT S. (2015). Analysis and automatic recognition of Human BeatBox sounds : A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4255–4259, Brisbane, QLD, Australia. DOI : 10.1109/ICASSP.2015.7178773.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The Kaldi Speech Recognition Toolkit. p.4̃, Hilton Waikoloa, Big Island, Hawaii, US.

SINYOR E., MCKAY C., FIEBRINK R., MCENNIS D. & FUJINAGA I. (2005). Beatbox classification using ACE. p.$\tilde{4}$, London, UK.

TIWARI V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, p. 19–22.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Perception et production du trait de nasalité vocalique chez l'enfant porteur d'implants cochléaires.

Sophie Fagniart[1], Brigitte Charlier[2,3], Véronique Delvaux[1,4], Anne Huberlant[3], Kathy Huet[1], Myriam Piccaluga[1], Isabelle Watterman[2] et Bernard Harmegnies[1]

(1) Institut de Recherche en Sciences et Technologie du Langage UMONS, Belgique
(2) Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050 Bruxelles, Belgique
(3) Centre « Comprendre et Parler », Rue de la Rive 101, 1200 Woluwe-Saint-Lambert, Belgique
(4) Fond National de la Recherche Scientifique, Belgique
sophie.fagniart@umons.ac.be

RÉSUMÉ

L'implant cochléaire, malgré une amélioration considérable de la perception auditive, ne fournit qu'une information acoustique partielle, pouvant donner lieu à des difficultés de perception de certains contrastes phonétiques. L'étude présentée vise à déterminer les compétences de perception et de production des voyelles nasales et orales d'enfants porteurs d'implants cochléaires en comparaison aux compétences d'enfants normo-entendants. Malgré des résultats très satisfaisants dans les deux groupes, on observe des patterns d'erreurs spécifiques au groupe d'enfants implantés dans les tâches perceptives, ainsi que certaines particularités dans la réalisation phonétique des voyelles nasales, portant notamment sur les valeurs de bande passante.

ABSTRACT

Perception and production of the nasal vowels in cochlear implanted children.

Cochlear implant, despite a considerable improvement of the auditory perception, provides only partial acoustic information, which can lead to difficulties in perceiving some phonetic contrasts. The present study aimed to determine the skills in perception and production of nasal and oral vowels between children with cochlear implants and normal hearing children. Despite very satisfactory results in both groups, there are patterns of errors specific to the group of children implanted in perceptual tasks, as well as some characteristics in the phonetic realization of nasal vowels, in particular with the bandwidth values.

MOTS-CLÉS : Implant cochléaire, nasalité vocalique, phonétique, perception, production
KEYWORDS: Cochlear implant, vocalic nasality, phonetics, perception, production

1 Introduction

Les voyelles nasales, du fait de leur réalisation articulatoire impliquant le couplage de trois systèmes de résonance, donnent lieu à un signal acoustique riche et complexe. Dans ce cadre, il y a lieu de s'interroger sur la perception de cette classe de phonèmes auprès des personnes atteintes de surdité, dont l'accès à l'information acoustique est limité, la lecture labiale ne permettant pas de distinguer la nasalité vocalique (Borel, 2015). En ce qui concerne plus particulièrement la population atteinte de

Volume 1 : Journées d'Études sur la Parole, pages 217–225.
HAL : hal-02798539.

surdité et porteuse d'un implant cochléaire, il y a lieu, qui plus est, de se questionner sur la représentativité du signal émis par le dispositif – celui-ci étant composé de 12 à 22 électrodes de transmission contre des milliers de cellules ciliées dans l'oreille saine. Hawks, Fourakis, Skinner Holden & Holden (1997) ont notamment mis en évidence des difficultés à discriminer des voyelles orales aux bandes passantes plus larges auprès d'une population implantée. Borel (2015) a centré différentes recherches sur la perception de la nasalité auprès d'une population adulte implantée (76 implantations unilatérales, 6 implantations bilatérales), au sein de différents paradigmes expérimentaux. Les 82 sujets testés ont présenté, dans une tâche d'identification phonémique, des performances significativement inférieures à celles de leurs pairs entendants en ce qui concerne la perception des voyelles nasales, qu'ils perçoivent comme des voyelles orales, et ce même après un délai post-implantation d'un an. Dans son étude de 2018, Crouzet a investigué la perception des consonnes et voyelles nasales auprès de 19 participants normo-entendants, en traitant les phonèmes utilisés via un vocodeur permettant de simuler la déformation sonore liée à l'implant en faisant varier différents paramètres permettant de moduler la résolution spectrale du son. Les résultats sont très marqués : tandis que les sujets ne présentent pas de difficulté accrue à traiter la nasalité au sein des consonnes – avec une tendance à l'amélioration des performances avec l'augmentation de la résolution spectrale - les voyelles nasales sont significativement moins bien perçues, quel que soit le paramétrage sélectionné pour les synthétiser. Ces différentes études corroborent l'hypothèse de difficultés accrues dans la perception des voyelles nasales chez les individus porteurs d'implant(s) cochléaire(s). Toutefois, nous notons l'absence d'études portant sur les compétences de production de cette distinction entre voyelles orales et nasales chez les individus porteurs d'implant(s) cochléaire(s) : est-ce que les déficits perceptifs se manifestent par une distinction moins marquée des voyelles orales/nasales en production ? De plus, la perception de la distinction des voyelles orales/nasales n'a été évaluée que chez l'adulte implanté. Qu'en est-il de l'enfant atteint de surdité pré-linguale porteur d'implants cochléaires ?

Par ailleurs, les études de la perception du son au travers d'un vocodeur par des sujets normo-entendants, outre le fait qu'elles peuvent ne pas représenter précisément l'input auditif perçu au travers de l'implant, ne permettent pas de rendre compte des stratégies perceptuelles mises en place par la population implantée. L'étude décrite ci-après vise à répondre à ces interrogations, en investiguant les capacités de perception et de production de voyelles orales et nasales auprès d'enfants atteints de surdité pré-linguales et porteurs d'implants cochléaires.

2 Méthodologie

2.1 Participants

L'étude a été menée auprès de deux groupes d'enfants âgés entre 5 et 12 ans : un groupe d'enfants atteints de surdité et porteurs d'implants cochléaires (groupe IC) et un groupe témoin d'enfants normo-entendants (groupe NE). Le groupe IC était constitué de 13 enfants (7 filles et 6 garçons), âgés entre 5;8 ans et 11;6 ans, présentant une surdité de perception pré-linguale profonde bilatérale, tous porteurs d'implants cochléaires bilatéraux de la marque « Cochlear » (implantation entre 9 et 30 mois), recrutés grâce aux logopèdes du centre « Comprendre et Parler » de Bruxelles (n=9) et au « Centre Médical d'Audiophonologie (CMAP) » de Montegnée (n=4). Ils ont tous reçu une rééducation auditive de type « oraliste », aussi bien dans le contexte de leur centre de réhabilitation que dans leur contexte familial. Le « Langage parlé complété », soutien à la discrimination phonologique et à la lecture labiale, est couramment utilisé auprès de ces enfants avec les orthophonistes et aidants de leur centre de réhabilitation, et certains le pratiquent également en contexte familial (n=7). Leur courbe d'audiométrie vocale en répétition de mots/pseudo-mots se situe de 88 à 100% à 55/60dB. Le groupe

NE était constitué de 25 enfants (11 filles et 14 garçons) âgés de 5 à 12 ans (moyenne : 8;6 ans – écart-type : 2;4 ans), scolarisés dans l'enseignement général, ayant pour langue maternelle le français (familles monolingues), notons que certains sont toutefois en enseignement néerlandophone (n= 7). Les sujets ayant reçu ou suivant une prise en charge orthophonique ont été exclus lors du recrutement.

2.2 Tâches

Trois tâches ont été administrées : une tâche d'identification de mots-cible (pseudo-mots C1V1C2V2 avec C1=C2=/t/ et V1=V2 parmi /ɑ̃, ɔ̃, ɛ̃/ ou /a, o, ɛ, u/ : /tɑ̃tɑ̃/, /tɔ̃tɔ̃/, /tɛ̃tɛ̃/, /tata/, /toto/, /tɛtɛ/, /tutu/) en contexte phrastique, une tâche de discrimination et une tâche de répétition de ces pseudo-mots. La tâche d'identification consistait à présenter une phrase au sein de laquelle se trouvait un mot-cible contenant une voyelle orale ou nasale. Nous avons choisi d'utiliser des pseudo-mots afin d'éviter tout effet de fréquence lexicale, et afin de pouvoir utiliser une structure commune pour tous les items. Les pseudo-mots cibles ont ainsi été construits de telle façon qu'il soit nécessaire de traiter l'information spectrale des voyelles orales et nasales, en limitant des effets de coarticulation, de longueur de segments ou encore d'indices prosodiques. Les voyelles orales sélectionnées présentent soit une proximité phonologique, soit une proximité phonétique avec une des voyelles nasales du français. Par exemple, l'opposition phonologique de la voyelle nasale /ɑ̃/ est constituée par la voyelle orale /a/, tandis que la réalisation phonétique de cette même voyelle nasale sans nasalité est davantage proche phonétiquement de la voyelle orale /ɔ/ / (français parisien : Borel, 2015 ; français de Belgique : Delvaux, 2012). La consonne /t/ a été choisie afin de limiter des effets de variabilité liés à la coarticulation. Les pseudo-mots ont été construits sur base de stimuli naturels, produits par un locuteur masculin. La durée de V1 et V2 a été contrôlée et fixée à 100ms afin d'éviter des biais de réponse liés à la longueur des voyelles – celle-ci étant un indice perceptif important dans la discrimination des voyelles orales et nasales (Delattre & Monnot, 1968). Huit phrases (produites de façon naturelle par un locuteur masculin) ont été construites de sorte que le pseudo-mot soit placé en position intermédiaire (« J'ai vu tantan près du bus ») ou en position finale (« Près du bus, j'ai vu tantan »), pour un total de 56 items (8 phrases*7 mots-cible).

Lors de la tâche d'identification, chaque pseudo-mot était associé à un personnage représenté sur une carte disposée sur la table. Durant une phase d'apprentissage, l'expérimentateur apprenait à l'enfant le nom des personnages par l'association d'un geste et d'une phrase mnémotechnique afin de faciliter leur rétention. Ensuite, il était demandé à l'enfant de prendre la carte correspondante au personnage cité au sein d'une phrase qu'il entendait, pour la placer sur une image correspondant à la phrase produite (pour exemple : « J'ai vu tonton près du bus », l'enfant prend la carte « Tonton » et la place à côté de l'image du bus). La tâche de discrimination (même/différent) a consisté en la présentation de 63 paires de pseudo-mots (5*9 paires différentes et 1*18 paires identiques, écart inter stimuli = 100ms) constituées de façon à évaluer si les voyelles orales étaient moins aisément discriminées si elles étaient présentées avec une voyelle orale proche phonétiquement ou phonologiquement (tableau 1).

nasales/orales – ***correspondant phonologique***	Tantan – Tata (5x) Tantan – Tantan Tata – Tata	Tintin – Tètè (5x) Tintin – Tintin Tètè – Tètè	Tonton – Toutou (5x) Tonton – Tonton Toutou – Toutou
nasales/orales – ***correspondant phonétique***	Tantan – Toto (5x) Tantan – Tantan Toto – Toto	Tintin – Tata (5x) Tintin - Tintin Tata – Tata	Tonton – Toto (5x) Tonton – Tonton Toto – Toto
Orales/orales	Tata – Toto (5x) Tata – Tata Toto - Toto	Tètè – Tata (5x) Tètè – Tètè Tata - Tata	Toutou – Toto (5x) Toutou – toutou Toto - Toto

TABLEAU 1 : Tâche de discrimination : paires de stimuli présentés.

Les réponses des enfants étaient récoltées via une application informatique développée par le laboratoire de phonétique de l'UMONS sur tablette tactile (Microsoft Surface Pro3). Afin de faciliter la rétention des consignes, des pictogrammes étaient placés sur les zones de réponses. Pour les deux tâches (identification et discrimination), les stimuli sonores étaient présentés aux enfants en champ libre via haut-parleurs (Bose Soundlink II) dont le volume sonore moyen était contrôlé grâce à un sonomètre et ajusté à 60dB, placés à 1 mètre du sujet.

Lors de la tâche de répétition, il était demandé aux enfants de produire une partie des stimuli-phrases de la tâche d'identification : les 4 phrases avec le pseudo-mot placé en position finale, soit 28 items (7 mots-cibles*4 phrases). Lors de cette tâche, afin de laisser au sujet la possibilité de capter les mouvements faciaux, l'expérimentateur produisait les phrases, tout en plaçant la carte du pseudo-mot cible sur la scène correspondante afin d'illustrer la phrase-cible produite. L'enfant était alors invité à produire oralement la phrase (production enregistrée via un magnétophone portable Zoom H5 posé à 25cm du sujet).

2.3 Traitement et analyse des données

Pour la tâche d'identification, nous avons analysé les pourcentages de bonnes réponses de chaque sujet pour chacun des phonèmes-cibles testés. Pour la tâche de discrimination, nous avons analysé les scores ***d'***, calculés selon la méthode de MacMillan & Creelman (1991). Pour la tâche de répétition de phrases, nous avons calculé les pourcentages des différents mots-cibles correctement produits au sein des phrases. Un mot-cible était jugé comme correct lorsqu'il était identifié de façon univoque par l'expérimentateur lors de la réécoute attentive des enregistrements.

Par ailleurs, une analyse acoustique a été menée sur les productions de la tâche de répétition : segmentation manuelle des voyelles et, extraction de leur durée, mesures des valeurs des trois premiers formants et des bandes passantes grâce à une procédure automatique via le logiciel PRAAT (paramétrage : 5 formants, fréquence maximale = 5500Hz et fenêtre temporelle = 0,025s).

Les analyses acoustiques, plus spécifiquement de certains paramètres acoustiques liés à la distinction voyelles orales/nasales, doivent nous permettre de qualifier objectivement les patterns de réalisations acoustiques de nos sujets. Ainsi, nous avons mesuré la durée des différentes voyelles produites, afin de voir si l'allongement caractéristique des voyelles nasales (Delattre, 1968) se retrouvait davantage dans un de nos groupes. Nous avons également mesuré les valeurs des trois premiers formants, afin d'observer si les réalisations acoustiques des deux groupes d'enfants étaient similaires sur les voyelles orales et nasales. D'autre part, nous avons utilisé ces valeurs afin de situer chaque production d'enfants dans le plan F1/F2 et de procéder à différentes comparaisons entre les paires de voyelles orales et nasales nous intéressant. Enfin, nous avons aussi extrait les valeurs des bandes passantes des trois formants de chacune des productions. En effet, certains auteurs ont rapporté que les bandes passantes des deux premiers formants pourraient être impliquées dans la perception de la nasalité (Hawks, Fourakis, Skinner & Holden, 1997 ; House & Stevens, 1956 ; Delvaux, 2013).

3 Résultats

Pour la tâche d'identification, on observe tout d'abord significativement ($U(1) = 462084$; $p < .001$) de moins bonnes performances chez les enfants IC (88,5%) que chez les enfants NE (97,8%). Une analyse des comparaisons spécifiques indique que les différences de performance sont associées aux phonèmes /ɑ̃/ avec une performance de 90% pour le groupe NE contre 75% pour le groupe IC ($U(1) = 79,5$; $p = .009$) ; aux phonèmes /a/ avec une performance de 98% pour le groupe NE contre 87,5%

pour le groupe IC (U(1) = 88,5 ; p=.021) et aux phonèmes /u/ avec une performance de 100% pour le groupe NE contre 76,92% pour le groupe IC (U(1) = 37,5 ; p<.001).

Stimulus	Réponse						
	ɑ̃	**ɔ̃**	**ɛ̃**	**a**	**o**	**u**	**ɛ**
ɑ̃	NE : 90% **IC : 76%**	NE : 10% **IC : 22,1%**	NE : / **IC : 0,9%**	NE : / IC : /	NE : / **IC : 1%**	NE : / IC: /	NE : / IC: /
ɔ̃	NE : / **IC : 5,8%**	NE : 97,5% **IC : 92,3%**	NE : / **IC : 1%**	NE : / IC : /	NE : 2% IC : 1%	NE : 0,5% IC : /	NE : / IC: /
ɛ̃	NE : / IC : /	NE : / IC : /	NE : 98,5% **IC : 90,4%**	NE : 1,5% **IC : 9,6%**	NE : / IC : /	NE : / IC : /	NE : / IC : /
a	NE : 0,5% IC : 1,9%	NE : / IC : /	NE : 1% **IC : 9,6%**	NE : 98,5% **IC : 87,5%**	NE : / IC : /	NE : / IC : /	NE : / **IC : 1%**
o	NE : / IC : /	NE : / IC : 1,9%	NE : / IC : /	NE : / IC : /	NE : 100% IC : 98,1%	NE : / IC : /	NE : / IC : /
u	NE : / IC : /	NE : / **IC : 1%**	NE : / IC : /	NE : / IC : /	NE : / **IC : 15,4%**	NE : 100% **IC : 76,9%**	NE : / **IC : 6,7%**
ɛ	NE : / IC : /	NE : / IC : /	NE : / IC : /	NE : / IC : /	NE : / **IC : 1,9%**	NE : / IC : /	NE : 100% IC : 98,1%

TABLEAU 1 : Matrice de confusions des enfants normo-entendants (NE) et implantés (IC) pour chacun des phonèmes-cibles.

Lorsque l'on examine la matrice de confusion (tableau 2), on voit que l'erreur la plus fréquente aussi bien chez les enfants IC que les enfants NE sont les substitutions des phonèmes /ɑ̃/ par /ɔ̃/. Chez les enfants IC, les erreurs les plus fréquentes et peu/pas observées chez les enfants NE sont des substitutions des phonèmes /u/ par /o/, et de /ɛ̃/ par /a/, et de /a/ par /ɛ̃/, de /u/ par /ɛ/ et de /ɔ̃/ par /ɑ̃/.

Pour la tâche de discrimination, la moyenne des scores ***d'*** des enfants NE (4,41 – ce qui correspond à 98,45%) est significativement plus élevée que celle des enfants IC (4,12 – ce qui correspond à 94,87%), témoignant de moins bonnes capacités de discrimination chez ces derniers (U(1) = 11421 ; p=.04). En analysant plus spécifiquement les scores d', nous voyons des différences significatives entre les deux groupes pour /ɛ̃/-/a/-/ɛ/ (IC = 4.21 ; NE=4.47 ; U(1)=1233.5 ; p=.027), et /ɔ̃/-/u/-/o/ (IC = 3.86 ; NE=4.39 ; U(1)=1153 ; p=.013), les enfants IC commettant davantage d'erreurs que les enfants NE. On observe des performances équivalentes des deux groupes pour /ɑ̃/-/a/-/o/ (U(1) = 1434 ; p=.808).

Pour la tâche de répétition, on observe que les deux groupes ont des performances élevées, le taux d'erreurs étant très bas pour ce qui est de l'adéquation des deux voyelles produites au sein de chaque pseudo-mot, les enfants IC étant néanmoins globalement significativement meilleurs (U(1) =469694 ; p=.01). En ce qui concerne les mesures de durées des phonèmes en fonction du type de voyelle (figure 1), on constate : (1) que les voyelles nasales sont significativement plus longues lors de la production (orales= 0,097s – nasales = 0,135s - (U)=808056 ; p<.001), et (2) que les enfants du groupe IC produisent des segments significativement plus longs que ceux du groupe NE, et ce aussi bien pour les voyelles orales (NE = 0,093s - IC = 0,105s - (U)=193408 ; p<.001) que nasales (NE = 0,129s – IC = 0,147s - (U)=11262,5 ; p<.001).

L'analyse des valeurs de F1, F2 et F3 au sein des deux groupes, indique : (1) que les deux groupes présentent des valeurs formantiques similaires pour les voyelles nasales sauf pour le phonème /ɛ̃/ dont la valeur de F2 est significativement plus élevée au sein du groupe IC, (2) qu'en ce qui concerne les voyelles orales, il existe des différences significatives pour les valeurs de F1 de /a/ et /o/ qui sont plus élevées chez les enfants du groupe IC, et pour les valeurs du F2 du /o/ et /u/ qui sont cette fois moins élevées chez le groupe IC. Nous avons également comparé, entre les deux groupes, les distances euclidiennes dans le plan F1/F2 entre les paires de phonèmes associés selon leur proximité phonétique

ou phonologique. On constate que les groupes diffèrent significativement pour les valeurs de distances euclidiennes de /ɑ̃/ et /o/, la distance relative entre ces deux phonèmes étant plus faible chez les IC par rapport aux NE.

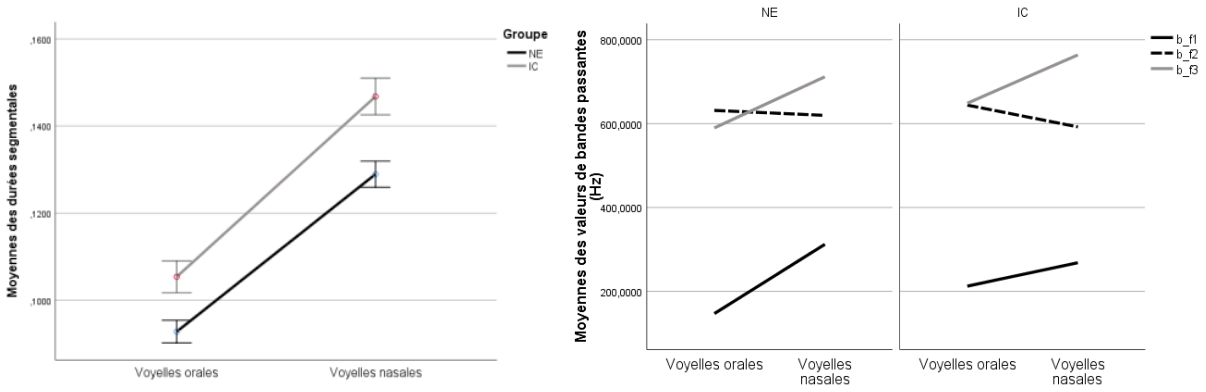


FIGURE 1 (gauche) : Moyennes des durées segmentales, en fonction du groupe et du type de voyelle - FIGURE 2 (droite) : Moyennes des valeurs moyennes de bandes passantes des trois premiers formants, en fonction du groupe et du type de voyelle.

Enfin, nous avons comparé, entre les deux groupes, les moyennes des bandes passantes des trois premiers formants (Figure 2), pour chaque type de voyelle (orales/nasales). On observe une différence significative entre les valeurs de bande passante de F1 : (1) pour les voyelles orales ((U)=80056 ; p=.030), les enfants IC ayant des valeurs supérieures (IC= 212,44 Hz; NE= 147,00 Hz), et (2) pour les nasales ((U)=170372,5 ; p=.006), les enfants IC ayant des valeurs plus basses (IC= 268,24 Hz; NE =312,14 Hz), mais également des différences significatives sur les valeurs de bande passante de F3 uniquement pour les nasales ((U)=176371 ; p<.001), les enfants IC ayant alors des valeurs plus élevées (IC = 763,9 Hz; NE = 711,77 Hz).

Par ailleurs, un effet d'interaction significatif entre la variable groupe et la variable nasalité sur les valeurs de bande passante de F1 (F(1, 2064)= 14,919 : p<.001) a été mis en évidence : les enfants NE marquent de façon nette la nasalité par une augmentation de 212%, correspondant à une augmentation de 168 Hz, l'augmentation de bande passante des nasales pour le groupe IC n'est que de 126%, correspondant à une augmentation de 56 Hz.

4 Discussion

Tout d'abord, en ce qui concerne la première tâche perceptive (identification de pseudo-mots au sein de phrases), notons que les enfants IC, malgré des performances moins élevées que le groupe NE, présentent des performances assez satisfaisantes de presque 90% de réponses correctes. Lors de la tâche de discrimination, les enfants IC ont à nouveau présenté des scores inférieurs aux NE. En outre, l'erreur la plus fréquente lors de la tâche d'identification au sein du groupe IC porte sur le phonème nasal /ɑ̃/ confondu avec une autre nasale /ɔ̃/, et davantage d'erreurs au sein du triplet /ɑ̃/-/a/-/o/ lors de la tâche de discrimination, erreurs qui sont également les plus fréquentes dans le groupe NE, ne semblant pas témoigner de difficultés spécifiques au groupe IC. Par ailleurs, on note certaines erreurs chez les IC qui sont très peu présentes voire totalement absentes chez les enfants NE, à savoir : (1) des erreurs de confusion entre les phonèmes /ɛ̃/ et /a/, ou sur le phonème /u/, confondu avec /o/ et /ɛ/ lors de la tâche d'identification, et (2) davantage d'erreurs pour /ɛ̃/-/a/-/u/ et /ɔ̃/-/o/-/u/ en discrimination. Des confusions entre voyelles nasales /ɛ̃/, /ɔ̃/ avec les orales /a/, /o/, /u/ proches phonétiquement avaient déjà été rapportées comme une erreur fréquente dans la population implantée

adulte (Borel, 2015), ce qui corrobore l'hypothèse d'une difficulté à percevoir la nasalité de certains phonèmes, alors confondus avec une voyelle orale proche phonétiquement. Pour ce qui est du phonème /u/, nous n'avions pas connaissance d'études ayant rapporté de difficultés de perception de ce phonème chez l'enfant implanté. Nous pouvons évoquer comme piste explicative la compacité de ce phonème, ayant des valeurs formantiques plus basses et très rapprochées par rapport aux autres voyelles orales du français. Du fait du manque relatif de finesse des informations spectrales par l'implant, notamment en basses fréquences les informations spectrales liées à ce phonème pourraient être moins bien perçues et donner lieu à des confusions. Le phonème /o/, ayant des valeurs spectrales proches de ce phonème, est un parfait candidat pour le substituer. De plus, /u/ et /o/ ont une configuration articulatoire proche, du moins sur la dimension la plus visible, à savoir l'arrondissement des lèvres.

Au niveau de la production, on a remarqué des performances similaires entre les deux groupes d'enfants, les enfants IC n'ayant donc pas de difficultés à répéter adéquatement des pseudo-mots contenant des voyelles orales ou nasales. Toutefois, nous remarquons des différences se marquant au niveau acoustique. Tout d'abord, les enfants IC présentent des durées phonémiques plus élevées que les NE, cet allongement étant davantage marqué sur les voyelles nasales. Sachant que l'allongement de la durée des segments aide à la perception de la nasalité (Delattre, 1968), il est possible que cet allongement témoigne de l'utilisation d'indices temporels dans la perception et la production de segments vocaliques. Pour ce qui est des valeurs formantiques, on remarque que les valeurs sont pratiquement similaires au sein des trois premiers formants pour la production des voyelles orales et nasales, traduisant un geste articulatoire très proche entre les deux groupes. Les quelques différences portent sur les voyelles orales /a/ et /o/ ayant des valeurs de F1 plus élevées, pouvant traduire une aperture plus importante, et /o/, /u/ dont les valeurs de F2 sont plus faibles, peut-être donc davantage antériorisées. La voyelle nasale /ɛ̃/ présente quant à elle des valeurs de F2 plus élevées, témoignant d'une antériorisation de cette voyelle. Par ailleurs, nous avons mesuré les distances euclidiennes entre chaque voyelle nasale et son correspondant phonétique et phonologique au sein des deux groupes. On a remarqué une tendance générale à produire de façon plus proche des paires dites « phonétiques » (/ɑ̃/ et /o/, /ɛ̃/ et /a/, /ɔ̃/ et /o/) montrant à nouveau que ces paires de phonèmes semblent plus proches perceptuellement et qu'il y a donc lieu de les voir comme les principaux correspondants de l'opposition « orale/nasale ». Cette observation corrobore celle de différents auteurs ayant rapporté davantage de proximité perceptuelle et acoustique entre ces différentes paires (Delvaux, 2013 ; Carignan, 2014). Par ailleurs, nous avons noté que les distances euclidiennes de l'opposition /ɑ̃/ et /o/ étaient plus faibles au sein des productions des enfants IC, signifiant encore davantage de proximité dans la réalisation acoustique de ces deux phonèmes.

Enfin, les analyses réalisées sur les bandes passantes des formants nous ont permis de montrer, d'une part, une augmentation significative généralisée de la bande passante de F1 et F3 pour les voyelles nasales au sein des deux groupes, et, d'autre part, une augmentation moindre pour les enfants IC par rapport aux enfants NE sur les valeurs de bande passante de F1. L'augmentation de la bande passante pour les voyelles nasales avait déjà été rapportée dans la littérature. Les auteurs justifient cette augmentation par un amortissement du pic fréquentiel liée à la mise en résonance des cavités nasales, dont la surface plus étendue et surmontée d'une muqueuse absorbante va amortir le son, augmentant sa bande passante et diminuant son intensité (Bernthal & Beuckelman, 1977 ; Delvaux, 2012). Nos résultats confirment que la nasalité est bien associée à des valeurs de bande passante augmentées, du moins pour F1 et F3. Par ailleurs, nous avons vu que cette augmentation liée à la nasalité se réalisait de façon moins importante pour les valeurs de F1 dans notre groupe IC, d'une part car les valeurs de bande passante étaient plus élevées pour les voyelles orales, et d'autre part car ces valeurs étaient beaucoup plus faibles pour les voyelles nasales, donnant une augmentation très faible liée à la production de la nasalité (d'environ 120% contre 210% pour les NE). Ces différences suggèrent une

production atypique des voyelles, avec une faible différenciation des voyelles nasales et orales concernant les valeurs de bande passante. Or, si cette augmentation de bande passante traduit la mise en résonance de la cavité nasale, il est ainsi possible que les mouvements du port vélo-pharyngé soient imprécis, incomplets ou encore absents lors de la transition entre résonance orale et nasale. Cette proposition serait congruente avec les résultats de l'étude de Baudonck, Van Lierde, D'haeseleer et Dhooge (2015), observant en nasalance davantage de nasalité dans le timbre des enfants IC. Si les enfants IC ont des difficultés à gérer le contrôle entre nasalité et oralité via les mouvements de leur port vélo-pharyngé, il est normal que ceux-ci présentent une résonance nasale plus proche lors de sons oraux et nasaux. Notons toutefois que ces différences n'ont pas été perçues lors des productions qui ont été identifiées comme adéquates à plus de 95%, pouvant laisser penser à l'emploi de stratégies alternatives permettant la production de variantes phonétiques adéquates pour permettre de distinguer les voyelles orales des voyelles nasales. L'allongement des segments oraux et nasaux observé chez les enfants IC, combiné à une configuration articulatoire orale très proche de la réalisation phonétique des voyelles nasales, pourrait ainsi constituer une stratégie efficace de marquage de la nasalité vocalique.

Les erreurs observées lors des tâches de perception et les particularités relevées lors des productions de segments oraux et nasaux peuvent être mises en lien. En effet, si l'implant cochléaire transmet incomplètement et de manière atypique les informations spectrales nécessaires à la perception de certains indices acoustiques, comme la nasalité vocalique, l'enfant présentera, dès lors, des difficultés à traiter ces phonèmes mais aussi à les produire, du fait du fonctionnement imprécis de la boucle audio-phonatoire. Le fonctionnement de cette boucle permet un feedback auditif à l'enfant de ses propres productions, permettant lors de son développement de les ajuster afin de maîtriser les sons de sa langue. Or, si le système auditif est défaillant, la boucle audio-phonatoire fonctionnera de manière imprécise, donnant un feedback biaisé de l'enfant sur ses propres productions. On peut donc aisément imaginer que si certains traits phonologiques ne sont pas perçus adéquatement, ils seront également sujets à des imprécisions au niveau articulatoire. Ceci pourrait expliquer les particularités acoustiques relevées dans notre groupe d'enfants implantés concernant la production des voyelles nasales.

5 Conclusion

Ces résultats semblent intéressants à prendre en compte pour le diagnostic et la prise en charge des enfants porteurs d'implants cochléaires. En effet, malgré des performances en apparence très satisfaisantes, les patterns d'erreurs en perception et les particularités en production laissent à penser que le signal acoustique n'est pas traité de la même manière chez l'enfant implanté et chez l'enfant normo-entendant. Par ailleurs, bien que les résultats obtenus par les enfants implantés de notre étude soient en apparence très proches de ceux obtenus par les enfants normo-entendants, il faut garder à l'esprit que les tâches administrées ne reflètent peut-être pas le fonctionnement perceptif et productif de l'enfant dans sa vie de tous les jours, où il sera confronté à du bruit, différents locuteurs, des états de fatigue et de santé variables, ou tout autre facteur pouvant limiter la mise en place de stratégies de compensation pour percevoir adéquatement un input auditif dégradé. De plus, notons que le trait phonétique de nasalité a été ici évalué dans un contexte lexical. Bien que les mots-cible aient été des pseudo-mots, l'association préalable avec des personnages les ont lexicalisés, et, en outre, ces mots-cible étaient présentés aussi bien en perception qu'en production comme des éléments saillants au sein de phrases. De plus, les possibilités de réponse étaient restreintes aux choix proposés à l'enfant, on peut se demander, en contexte d'identification libre des phonèmes, comment les enfants implantés classifient les sons de parole perçus. Si l'on remarque certaines particularités perceptives et productives dans des contextes privilégiés comme ceux-ci, qu'en est-il du traitement de ce type de contraste phonétique au sein de segments moins saillants et plus abstraits du langage, comme les morphèmes grammaticaux ? Les morphèmes grammaticaux sont davantage vulnérables à des

difficultés de traitement cognitif et/ou perceptif. L'investigation du lien entre le développement phonético-phonologique et les composantes langagières dites de niveaux supérieurs, comme la morphosyntaxe, nous semble être du plus grand intérêt au sein de la population d'enfants implantés, et fera l'objet de nos travaux de thèse.

Références

BAUDONCK, N., VAN LIERDE, K., D'HAESELEER, E., & DHOOGE, I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. International journal of pediatric otorhinolaryngology, 79(4), 541-545. doi.org/10.1016/j.ijporl.2015.01.025

BERNTHAL, J. E., & BEUKELMAN, D. (1977). The effect of changes in velopharyngeal orifice area on vowel intensity. Cleft Palate Journal, 14(1), 63-77.

BOERSMA, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

BOERSMA, P. & WEENINK, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.56, retrieved 13 March 2019 from http://www.praat.org/

BOREL, S. (2015). Perception auditive, visuelle et audiovisuelle des voyelles nasales par les adultes devenus sourds. Lecture labiale, implant cochléaire, implant du tronc cérébral (Doctoral dissertation). Université de Sorbonne Nouvelle.

CARIGNAN, C. (2014). An acoustic and articulatory examination of the "oral" in "nasal" : The oral articulations of French nasal vowels are not arbitrary. Journal of phonetics, 46, 23-33. doi.org/10.1016/j.wocn.2014.05.001

CORRETGE, R. (2019). Praat Vocal Toolkit. http://www.praatvocaltoolkit.com

CROUZET, O. (2018). Perception des consonnes et voyelles nasales en parole vocodée: Analyse de la contribution des niveaux de résolution spectrale et temporelle. Actes des XXXIIèmes Journées d'Études sur la Parole–JEP2018, Aix-en-Provence, France, 4-8.

DELATTRE, P., & MONNOT, M. (1968). The role of duration in the identification of French nasal vowels. IRAL-International Review of Applied Linguistics in Language Teaching, 6(1-4), 267-288.

DELVAUX, V., METENS, T., & SOQUET, A. (2002). Propriétés acoustiques et articulatoires des voyelles nasales du français. XXIVèmes Journées d'étude sur la parole, Nancy, 1, 348-352.

DELVAUX, V., DEMOLIN, D., SOQUET, A., & KINGSTON, J. (2004). La perception des voyelles nasales du français. Actes des XXVèmes JEP, 157-160.

DELVAUX, V. (2012). Les voyelles nasales du français. Aérodynamique, articulation, acoustique et perception. Presses Interuniversitaires de Bruxelles, Belgique : GRAMM-R.

HAWKS, J. W., FOURAKIS, M. S., SKINNER, M. W., HOLDEN, T. A., & HOLDEN, L. K. (1997). Effects of formant bandwidth on the identification of synthetic vowels by cochlear implant recipients. Ear and hearing, 18(6), 479-487.

MACMILLAN, N. A., & CREELMAN, C. D. (1991). Detection theory: A user's guideCambridge University Press. New York.

MAEDA, S. (1993). Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. *In* Huffman, M.K. & Krakow, R.A. (Eds.) *Phonetics and phonology : Nasals, nasalization, and the velum* (pp. 147-167). New York : Academic Press. doi.org/10.1016/B978-0-12-360380-7.50010-7

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Une nouvelle mesure de la réverbération pour prédire les performances a priori de la transcription de la parole

Sébastien Ferreira[1,2], Jérôme Farinas[1], Julien Pinquier[1], Julie Mauclair[1] et Stéphane Rabant[2]

(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
(2) Authôt, 52 Avenue Pierre Semard, 94200, Ivry-sur-Seine, France
prenom.nom@irit.fr[1],sferreira@authot.com,srabant@authot.com

RÉSUMÉ

Dans cette étude, nous explorons la prédiction *a priori* de la qualité de la transcription automatique de la parole dans le cas de la parole réverbérée enregistrée avec un seul microphone. Cette prédiction est faite avant le décodage pour informer les utilisateurs de la qualité de la transcription attendue. Dans cette étude, nous nous concentrons uniquement sur les pertes de performance liées à la réverbération. Une nouvelle mesure de réverbération appelée « Excitation Behavior » est introduite. Cette mesure exploite le résidus de la prédiction linéaire sur les fenêtres voisées du signal de parole. L'expérience a été menée sur le corpus Wall Street Journal, réverbéré par des réponses impulsionnelles provenant du REVERB Challenge. Par rapport aux autres mesures de réverbération testées, notre mesure obtient une amélioration relative de 20% de la prédiction du taux d'erreur (aussi bien au niveau des phonèmes que des mots).

ABSTRACT

A new reverberation measure to predict *a priori* ASR performance

In this study, we explore the *a priori* prediction of the quality of automatic speech transcription in the case of reverberant speech recorded with a single microphone. This prediction is made before decoding to inform users of the expected transcription quality. We studied only the performance losses related to reverberation. A new reverberation measure called "Excitation Behavior" is introduced. This measure exploits the residuals of the linear prediction on the voiced windows of the speech signal. The experiment was conducted on the Wall Street Journal corpus, reverberated with impulse responses from the REVERB Challenge. Compared to the other reverberation measurements tested, our measurement obtains a relative prediction improvement of more than 20% (both at phone and word level).

MOTS-CLÉS : prédiction de performance, reconnaissance automatique de la parole, réverbération.

KEYWORDS: performance prediction, automatic speech recognition, reverberation.

1 Introduction

Au cours de la dernière décennie, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont atteint de bonnes performances sur de la parole « propre ». L'amélioration de la robustesse des systèmes de RAP par rapport aux différents environnements acoustiques est un problème de recherche toujours d'actualité. Même si actuellement les systèmes de RAP sont de plus en plus

Volume 1 : Journées d'Études sur la Parole, pages 226–234.
HAL : hal-02798541.

robustes, la qualité de la transcription reste fortement dépendante de la qualité du signal de parole. Dans les systèmes commerciaux de RAP, les enregistrements soumis à une tâche de transcription ont généralement (voir même quasi-exclusivement) un seul canal audio. Ce constat limite les méthodes pouvant être employées afin d'améliorer la robustesse des systèmes car les algorithmes exploitant plusieurs microphones, comme les traitements par faisceaux (« beamforming » en anglais), ne peuvent pas être utilisés (Kinoshita *et al.*, 2016).

Comme nous l'avions fait précédemment lors d'une étude sur la parole bruitée (Ferreira *et al.*, 2018), l'objectif de cet article est de prédire *a priori* la qualité de la transcription des systèmes de RAP pour la parole réverbérée. L'analyse est *a priori* car elle est entièrement effectuée avant le décodage de la parole. L'avantage de cette prédiction est d'informer au plus tôt un utilisateur de la qualité de la transcription attendue. Nous avons choisi de traiter uniquement le cas de la réverbération afin d'isoler les sources de dégradation de la qualité de la parole. Ces travaux ont vocation à être utilisés par des systèmes commerciaux de transcription automatique de la parole (comme par exemple dans le modèle économique de la société Authôt), donc certaines contraintes ont été fixées :

- la mesure doit être non-intrusive,
- la réponse impulsionnelle de la pièce d'enregistrement est inconnue,
- la prédiction doit être réalisée avant le décodage : aucun score de confiance ou hypothèse de transcription ne peuvent être utilisés,
- le signal est mono-canal,
- le système de RAP doit être considéré comme une boîte noire.

Afin de créer cette prédiction, nous cherchons des mesures qui quantifient l'impact de la réverbération sur la qualité de la transcription automatique de la parole. Nous cherchons un score d'« intelligibilité de la parole » pour les systèmes de RAP lorsque le signal de parole est réverbéré.

Les méthodes d'estimation de la réverbération qui satisfont l'ensemble de nos contraintes se répartissent en deux catégories : l'estimation aveugle du T60 (temps de réverbération) et les scores d'intelligibilité de la parole non-intrusive. Le temps de réverbération ou T60 est défini comme le temps nécessaire pour que le niveau de pression acoustique diminue de 60 dB après extinction de la source d'excitation sonore (ISO 3382, 1997). Dans de nombreuses situations, le T60 est estimé (et non calculé) car la réponse impulsionnelle de la pièce (RIR pour Room Impulse Response) est inconnue. Il existe deux méthodes pour estimer le T60 : la distribution de décroissance spectrale (SDD pour Spectral Decay Distribution) qui estiment la distribution de la décroissance de l'enveloppe de puissance du signal dans le temps afin d'estimer le temps de réverbération (Dumortier & Vincent, 2014) et l'analyse des résidus de la Prédiction Linéaire (PL) (Keshavarz *et al.*, 2012). Les méthodes orientées vers l'intelligibilité de la parole reposent sur la quantification des déformations du signal de parole. Par exemple, dans cet article (Falk *et al.*, 2010), ce sont les caractéristiques spectrales de modulation qui sont exploitées. Les méthodes qui estiment le T60 n'ont pas été conçu pour informer de la distance entre le locuteur et le microphone : il s'agit pourtant d'une variable importante de la qualité de la parole. Les méthodes orientées vers l'intelligibilité de la parole ont été conçu pour prédire l'intelligibilité humaine. Les tests que nous avons effectués sur la prédiction *a priori* de la qualité de la transcription avec ces mesures n'ont pas été satisfaisants. Pour ces raisons, nous avons décidé de créer une nouvelle mesure, que nous appelons Excitation Behaviour (EB), qui quantifie l'impact de la réverbération sur les performances des systèmes de RAP.

Le calcul de cette mesure est détaillée en section 2. Dans la section 3, le protocole d'expérimentation est décrit afin d'évaluer l'EB : le taux d'erreur mots (WER pour Word Error Rate) et le taux d'erreur phonèmes (PER pour Phone Error Rate) sont prédits sur différentes conditions de réverbération et discutés en section 4.

2 La mesure Excitation Behaviour

2.1 Architecture

L'Excitation Behaviour (EB) est un paramètre qui appartient aux méthodes d'analyse des résidus de la Prédiction Linéaire (PL). Les résidus de la PL ou le signal d'erreur de prédiction $e(n)$ est la différence entre la parole d'entrée et la parole estimée (Makhoul, 1975).

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

où a_k correspond aux coefficients de la prédiction linéaire (LPC pour Linear Prediction Coefficient), p l'ordre du filtre et $s(n)$ l'échantillon de parole.

Sur les fenêtres voisées de la parole, le résidu de la PL contient des informations sur l'instant de fermeture glottale et sur la source d'excitation (Ananthapadmanabha & Yegnanarayana, 1979). Lorsque la parole est réverbérée, la différence entre les impulsions glottales et la source d'excitation est plus faible (voir figure 1). C'est cette distorsion des résidus de la PL qui sera exploitée par l'EB. L'architecture de l'extraction de l'EB se fait en trois étapes : la sélection automatique des fenêtres de parole voisées, l'extraction du résidu de la PL et le calcul d'une valeur statistique basée sur des ratio entre percentiles de la distribution de ce résidu.

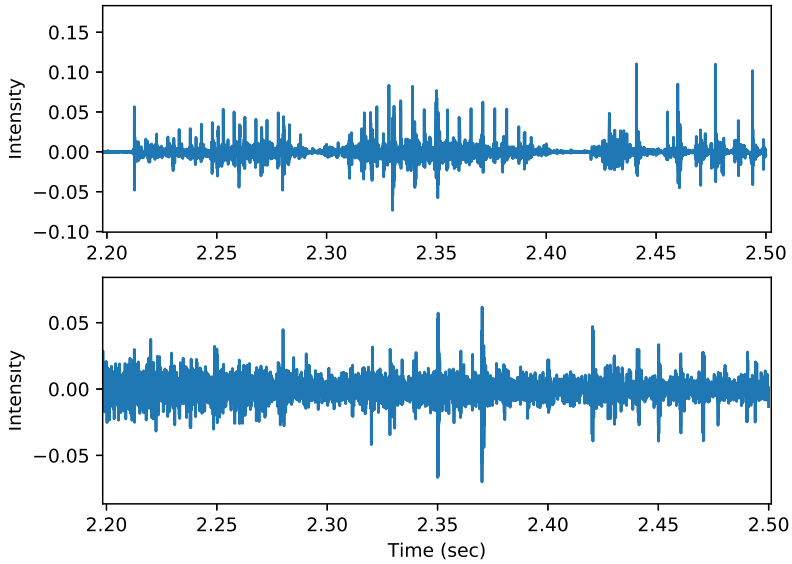


FIGURE 1 – Résidus de la PL d'un signal de parole. En haut signal propre. En bas version réverbérée

2.2 Sélection des fenêtres voisées

Pour sélectionner les fenêtres voisées de la parole, nous analysons la différence entre la racine quadratique moyenne (RMS pour Root Mean Square) et le taux de passage par zéro (ZCR pour Zero Crossing Rate) du signal. Une fenêtre de 16 ms est sélectionnée comme parole prononcée lorsque :

$$\frac{RMS_{trame}}{\max RMS_{global}} - \frac{ZCR}{2} > 0 \tag{2}$$

Un exemple de sélection de fenêtres est présenté dans la figure 2. Pour filtrer d'avantage cette sélection, seules les fenêtres voisées suivies de deux autres fenêtres voisées sont considérées comme voisées.

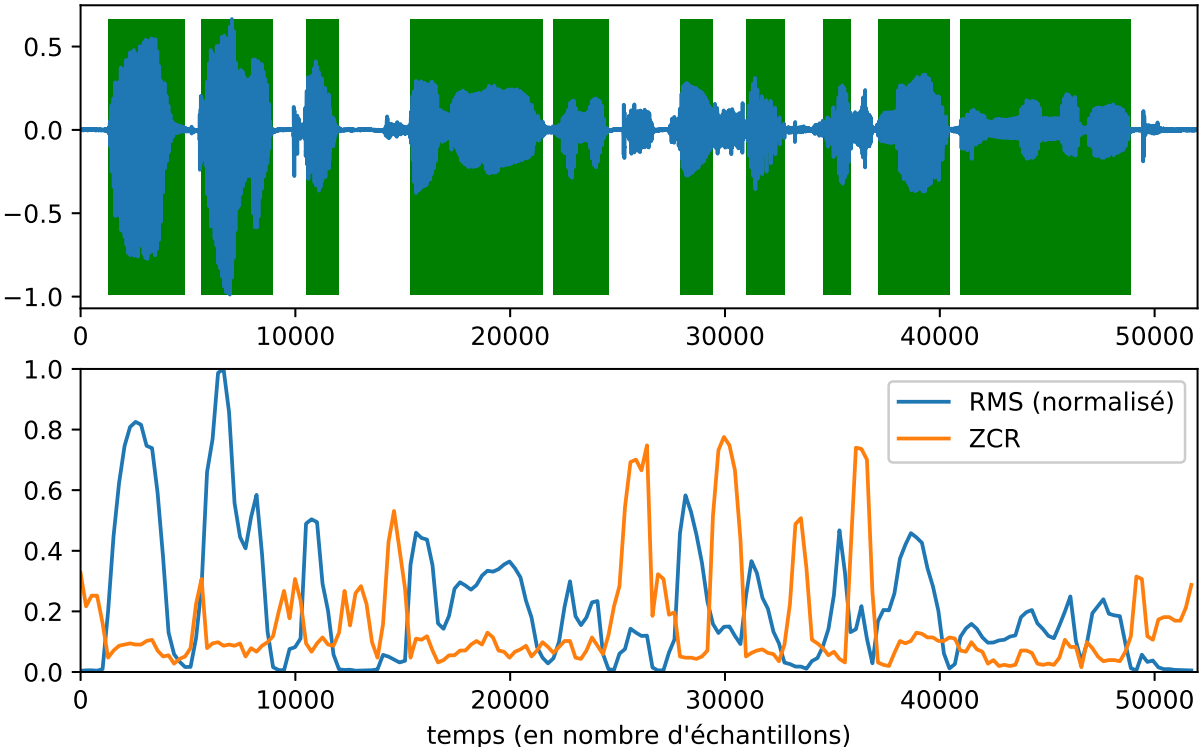


FIGURE 2 – En haut, le signal avec en vert les zones voisées sélectionnées automatiquement (équation 2). En bas, les valeurs du RMS normalisés et du ZCR correspondantes.

2.3 Calcul de la mesure

Après avoir sélectionné les fenêtres voisées, le signal est préaccentué. Pour calculer les LPC du signal, un modèle autorégressif est estimé avec l'algorithme de Levinson-Durbin pour minimiser l'erreur de prédiction (résidus de la PL). Les résidus de la PL sont calculés avec une fenêtre de 10ms et un ordre de 21. Pour chaque succession de 4 trames voisées, la transformée de Hilbert est utilisée sur les résidus de la PL et est normalisée par la valeur maximale : le maximum correspond à l'impulsion glottal la plus grande. La distribution de la transformée de Hilbert normalisée des résidus de la PL est affectée par la réverbération. Si nous observons les valeurs p_{90}, p_{50} ou p_{10} seules, la réverbération augmente ces valeurs. Cependant, ces valeurs de percentile sont très variables, selon le locuteur. C'est pourquoi le rapport défini dans l'équation 4 est calculé pour définir la mesure de l'EB, tel que :

$$p_{90}p_{50} = p_{90} - p_{50} \quad \text{et} \quad p_{50}p_{10} = p_{50} - p_{10} \tag{3}$$

où p_x définit les percentiles x^{th}. Avec la liste des $p_{90}p_{50}$ et des $p_{50}p_{10}$ obtenues pour la phrase, nous calculons le score EB avec :

$$EB = \frac{\overline{p_{90}p_{50}}}{\overline{p_{50}p_{10}}} \tag{4}$$

L'EB est proche de 1,8 lorsque la réverbération provoque beaucoup d'erreur par les système de RAP, et supérieur à 2,8 lorsque la réverbération est négligeable par les systèmes de RAP.

3 Expériences

3.1 Corpus

Le corpus Wall Street Journal est largement utilisé en RAP : WSJ0 (Garofalo *et al.*, 1993) et WSJ1 (Garofolo *et al.*, 1994). Ce corpus a été choisi pour limiter les erreurs induites par les modèles de langage, les accents et les disfluences. Les données *train_si284* (sans ajout de réverbération) sont utilisées pour entraîner le système de RAP. Pour travailler avec la parole réverbérée, les sous-ensembles *dev93* et *eval92* ont été convolués avec des réponses impulsionnelles des pièces (RIR pour Room Impulse Responses) mesurées dans différentes conditions. Les RIR proviennent du REVERB challenge (Kinoshita *et al.*, 2013). Les RIR enregistrées permettent de simuler 12 conditions de réverbération différentes : 6 pièces de tailles différentes (2 petites, 2 moyennes et 2 grandes) avec 2 types de distances entre un haut-parleur et un réseau de microphones (proche = 50 cm et loin = 200 cm). Les temps de réverbération des petites, moyennes et grandes pièces sont respectivement d'environ 0,25, 0,5 et 0,7s. Les sous-ensembles *Dev93* et *eval92* ont donc 13 conditions de réverbération différentes (le +1 vient du cas sans réverbération). Ces deux sous-ensembles sont respectivement utilisés pour entraîner et tester le modèle de régression.

3.2 Architecture du système de prédiction

Les systèmes de prédiction des performances des systèmes de RAP sont souvent appris avec une méthode de régression supervisée (voir Figure 3). L'objectif de cette régression est de modéliser le lien entre les caractéristiques extraites et la performances du système RAP : les systèmes de RAP sont plus ou moins robustes à certaines distorsions. Cette régression nous permet d'évaluer à quel point les différentes mesures de réverbération testées permettent de prédire l'impact de la réverbération sur les performances du système de RAP.

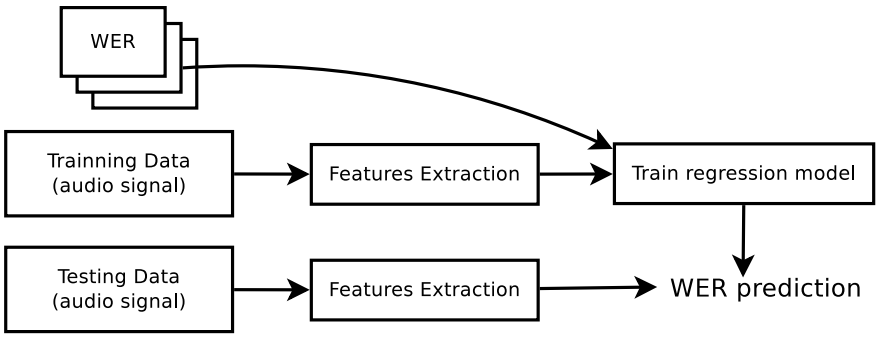


FIGURE 3 – Architecture du systèmes de prédiction de la qualité de la transcription.

3.3 Vérité terrain

Afin d'établir une vérité terrain pour notre système de prédiction, un système de RAP a été entraîné. Nous avons utilisé la recette de Karel Vesely avec Kaldi (Veselỳ *et al.*, 2013). Le système est hybride avec un réseau de neurones profond et des modèles de Markov cachés (DNN-HMM) entraîné par entropie croisée. Dans des conditions propres (train et test), le système de RAP entraîné avec *train_si284* obtient 5,84% de WER sur *dev93* et 3,42% sur *eval92*.

Pour construire le décodeur acoustico-phonétique utilisé pour prédire le PER, nous phonétisons au préalable les corpus d'entraînement et de test. Le dictionnaire de prononciation est modifié pour être composé uniquement des phonèmes possibles. Le modèle de langage est remplacé par un 1-gram appris sur un corpus de texte phonétisé. Ces changements permettent d'obtenir au décodage, une séquence de phonèmes sans aucun impact du modèle de langage. Une fois ces modifications apportées, le décodeur acoustico-phonétique est entraîné de la même manière que celle décrite au paragraphe précédent pour le système de RAP.

Pour calculer les modèles de régression, nous utilisons un perceptron multi-couche (MLP pour Multi-Layer Perceptron) que nous avons entraîné avec Scikit-learn.La MLP est assez simple et utilise une seul couche composée 3 neurones. Le but de la regression MLP est seulement d'obtenir une régression non linéaire.

3.4 Mesures de réverbération testées

Pour comparer la mesure de l'EB décrite dans la section 2, nous avons utilisé plusieurs mesures de l'état de l'art (voir le tableau 1) :

- SRMR+ : SRMR et SRMR normalisés (Falk *et al.*, 2010),
- Slope : ici c'est la valeur de "floored ratio of spectral substraction" (Tachioka *et al.*, 2013),
- Neg-side : une méthode de SDD qui utilise la variance négative et le skewness (Dumortier & Vincent, 2014),
- LP-kurto : moyenne des kurtosis des résidus de la PL d'ordre 10 (trame de 32ms) (Gillespie *et al.*, 2001).

4 Résultats & discussion

Afin d'évaluer la performance de la prédiction du WER, nous avons calculé l'erreur de prédiction absolue moyenne (MAE pour Mean Average Error) et l'écart-type (SD pour Standard Deviation). Dans (Willmott & Matsuura, 2005), les auteurs indiquent que le MAE est une mesure plus naturelle que le RMSE, et qu'elle est non ambiguë. Le MAE et le SD sont indiqués pour toutes les conditions de réverbération testées. Les résultats de la prédiction sont présentés dans le tableau 1.

En ce qui concerne la prédiction *a priori* du WER, l'EB obtient une meilleure précision de prédiction que les autres mesures de réverbération testées. Une amélioration relative de 23% de la prédiction de l'erreur moyenne de WER est obtenue avec EB par rapport à SRMR+. De plus, l'erreur de prédiction est moins dispersée.

Sachant que notre méthode de prédiction n'utilise pas d'informations linguistiques, nous avons voulu

TABLE 1 – Resultats de prédiction (WER et PER) avec une régression MLP.

	SRMR+	Slope	Neg-side+	LP-kurto	EB
WER (%)					
MAE	17,75	18,44	17,45	18,33	**13,66**
SD	14,26	14,95	13,75	15,69	**12,63**
PER (%)					
MAE	10,76	12,59	10,82	11,43	**7,86**
SD	8,14	9,04	7,75	9,15	**6,25**

observer les performances de prédiction du PER d'un système de décodage acoustico-phonétique : cela permet de retirer l'influence du modèle de langage. Le but est de savoir si la prédiction caractérise bien les distorsions acoustiques de la parole réverbérée. Nous pouvons voir, comme prévu, que la prédiction du PER est plus précise que la prédiction du WER. La précision de la prédiction de PER par notre méthode est meilleure que celle des autres mesures de réverbération testées. Une amélioration relative de 27% de la prédiction de l'erreur moyenne de PER est obtenue avec EB par rapport à SRMR+.

La mesure EB permet d'obtenir une prédiction des performance des système de RAP plus précise. Cependant, la précision reste insatisfaisante à l'échelle des phrases. La durée d'énonciation des phrases du corpus utilisé varie de 3 à 16 secondes, avec une moyenne de 7 secondes. Pour une tâche aussi difficile que la prédiction *a priori* des performances d'un système de RAP, une fenêtre temporelle plus longue est nécessaire. Dans la grande majorité des cas, les enregistrements soumis à une tâche de transcription automatique par les utilisateurs des services de RAP commerciaux dépassent 3 minutes. Nous avons donc analysé l'évolution de la prédiction en cumulant le nombre de phrases dans les même condition pour le même locuteur.

Dans la figure 4, la précision de la prédiction du WER devient plus précise, en fonction du nombre de tours de parole utilisés. Ainsi, la prédiction du WER atteint 7,13 d'erreur absolue moyenne avec 20 énoncés utilisés (durée d'environ 140 s). La prédiction de PER atteint 5,04 d'erreur absolue moyenne avec 20 énoncés utilisés (voir figure 5).

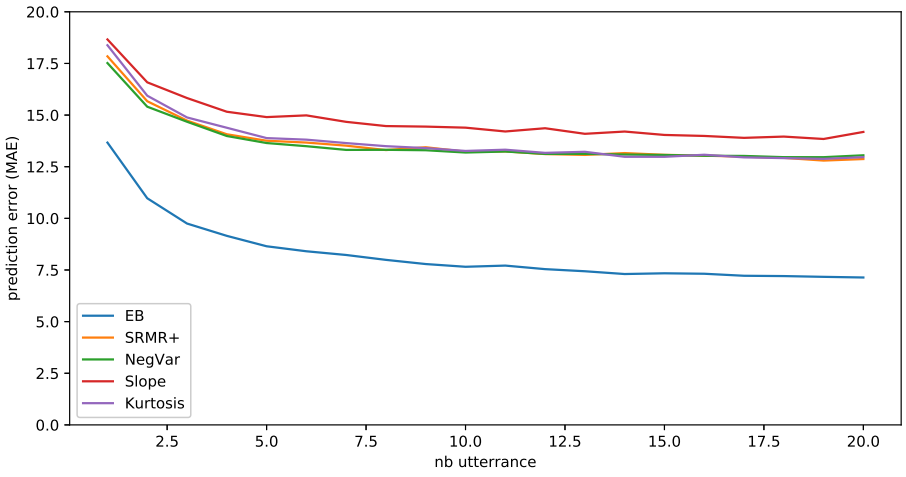


FIGURE 4 – Moyenne des erreurs de prédiction du WER en fonction du nombre de phrases utilisées.

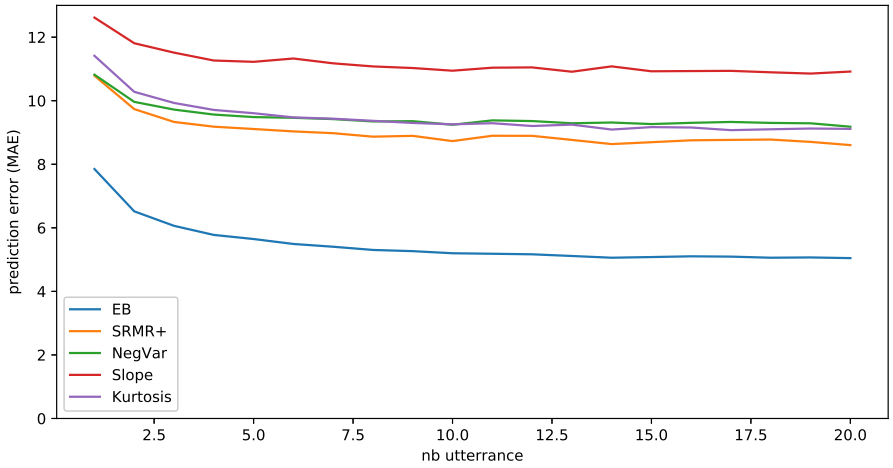


FIGURE 5 – Moyenne des erreurs de prédiction du PER en fonction du nombre de phrases utilisées.

5 Conclusions

Lorsqu'un signal de parole réverbéré est transcrit par un système de RAP, un grand nombre d'erreurs de transcription est possible. Dans les systèmes de RAP commerciaux, les fichiers audio sont principalement enregistrés avec un seul microphone : les méthodes exploitant plusieurs microphones ne peuvent pas être utilisées pour améliorer la robustesse des systèmes. Il est difficile de savoir dans quelle mesure la réverbération affecte les systèmes de RAP car le comportement de ces systèmes est très différent de celui des humains (lorsque la parole est réverbérée). Le principal objectif de ce papier est de prédire la qualité de la transcription *a priori* de la parole réverbérée, tout en respectant les contraintes des systèmes commerciaux de transcription automatique de la parole.

Nous avons contribué à l'élaboration de la mesure que nous appelons EB, pour quantifier l'impact de la réverbération sur les systèmes de RAP. L'EB extrait une mesure statistique sur le résidu de la PL, qui est bien corrélé à la performance des systèmes de RAP. Pour évaluer l'EB, nous avons prédit le WER et le PER obtenus avec les différentes mesures de réverbération testées. Au niveau d'une phrase, la mesure EB fournit une amélioration relative de la précision de prédiction du WER et du PER qui est de 20% supérieure aux autres mesures testées. La mesure EB obtient une erreur moyenne de prédiction de WER de 7,13 lorsque 140 secondes sont analysées, et une erreur moyenne de prédiction de PER de seulement 5,04. Dans tout les cas, la mesure de l'EB fournit une prédiction plus précise que les autres mesures testées.

Pour prédire *a priori* plus précisément la qualité de transcription des systèmes de RAP, d'autres analyses acoustiques du signal de parole pourraient être réalisées. Il serait intéressant d'observer l'influence du bruit et de la musique superposée. D'autre part, certaines variabilités au niveau des locuteurs, comme le débit de parole, le sexe ou l'âge, pourraient également améliorer la précision de la prédiction *a priori* de la qualité de la transcription automatique de la parole.

Références

ANANTHAPADMANABHA T. & YEGNANARAYANA B. (1979). Epoch extraction from linear

prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**(4), 309–319.

DUMORTIER B. & VINCENT E. (2014). Blind rt60 estimation robust across room sizes and source distances. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5187–5191 : IEEE.

FALK T. H., ZHENG C. & CHAN W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(7), 1766–1774.

FERREIRA S., FARINAS J., PINQUIER J. & RABANT S. (2018). Prédiction a priori de la qualité de la transcription automatique de la parole bruitée. *JEP*.

GAROFALO J. *et al.* (1993). CSR-I (WSJ0) complete LDC93S6A. Linguistic Data Consortium, Philadelphia, USA.

GAROFOLO J., GRAFF D., PAUL D. & PALLETT D. (1994). CSR-II (WSJ1) Complete. Linguistic Data Consortium, Philadelphia, USA.

GILLESPIE B. W., MALVAR H. S. & FLORÊNCIO D. A. (2001). Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 6, p. 3701–3704 : IEEE.

ISO 3382 (1997). *Acoustics : measurement of the reverberation time of rooms with reference to other acoustical parameters*. Standard, International Organization for Standardization, Genève, Suisse.

KESHAVARZ A., MOSAYYEBPOUR S., BIGUESH M., GULLIVER T. A. & ESMAEILI M. (2012). Speech-model based accurate blind reverberation time estimation using an lpc filter. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(6), 1884–1893.

KINOSHITA K., DELCROIX M., GANNOT S., P. HABETS E. A., HAEB-UMBACH R., KELLERMANN W., LEUTNANT V., MAAS R., NAKATANI T., RAJ B., SEHR A. & YOSHIOKA T. (2016). A summary of the REVERB challenge : state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, **2016**(1). DOI : 10.1186/s13634-016-0306-6.

KINOSHITA K., DELCROIX M., YOSHIOKA T., NAKATANI T., SEHR A., KELLERMANN W. & MAAS R. (2013). The reverb challenge : A common evaluation framework for dereverberation and recognition of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, p. 1–4 : IEEE.

MAKHOUL J. (1975). Linear prediction : A tutorial review. *Proceedings of the IEEE*, **63**(4), 561–580.

TACHIOKA Y., HANAZAWA T. & IWASAKI T. (2013). Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction. *Acoustical Science and Technology*, **34**(3), 212–215.

VESELỲ K., GHOSHAL A., BURGET L. & POVEY D. (2013). Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, p. 2345–2349.

WILLMOTT C. J. & MATSUURA K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, **30**(1), 79–82.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse de l'effet de la réverbération sur la reconnaissance automatique de la parole

Sébastien Ferreira[1,2], Jérôme Farinas[1], Julien Pinquier[1], Julie Mauclair[1] et Stéphane Rabant[2]

(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
(2) Authôt, 52 Avenue Pierre Semard, 94200, Ivry-sur-Seine, France
prenom.nom@irit.fr[1],sferreira@authot.com,srabant@authot.com

RÉSUMÉ

La Reconnaissance Automatique de la Parole (RAP) est moins performante lorsque le signal de parole est de mauvaise qualité. Dans cette étude, nous analysons les erreurs commises par les systèmes de RAP lorsque la parole transcrite est réverbérée afin de mieux comprendre les raisons de ces erreurs. Notre analyse permet de mettre en valeur les erreurs dues notamment à un mauvais alignement phonétique. Nous avons pu constater que les phonèmes de courte durée sont majoritairement supprimés lors du décodage phonétique. De plus, les phonèmes détectés, qu'ils soient corrects ou pas, ont tendance à avoir la même durée, ce qui est anormal pour certaines classes phonétiques comme les voyelles courtes ou les plosives. Nous avons aussi analysé les principales confusions entre les différentes classes phonétiques. Finalement, nous avons pu montrer que les erreurs lors de l'alignement phonétique des systèmes de transcription automatique entraînent beaucoup d'erreurs de détection.

ABSTRACT

Analyzing how reverberation affects Automatic Speech Recognition

Automatic Speech Recognition (ASR) is less effective when the speech signal is of poor quality. In this study, we analyze the errors made by ASR systems when the transcribed speech is reverberated in order to better understand the reasons for these errors. Our analysis allows us to highlight errors due to phonetic misalignment. We have found that short duration phonemes are mostly suppressed during phonetic decoding. Moreover, the detected phonemes, whether they are correct or not, tend to have the same duration. This is abnormal for certain phonetic classes such as short vowels or plosives. We also analyzed the main confusions between the different phonetic classes. We were able to show that errors in the phonetic alignment of automatic transcription systems lead to many detection errors.

MOTS-CLÉS : reconnaissance automatique de la parole, réverbération, analyse d'erreur.

KEYWORDS: automatic speech recognition, reverberation, error analysis.

1 Introduction

Au cours de la dernière décennie, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont atteint de bonnes performances. Néanmoins, la robustesse de ces systèmes reste insatisfaisante par rapport aux humains (Kinoshita *et al.*, 2016), notamment lorsque le signal de parole est réverbéré (surtout pour les fichiers enregistrés avec un seul microphone). Pour la parole réverbérée, le comportement des systèmes de RAP semble être différent de celui des humains (Lippmann, 1996).

Volume 1 : Journées d'Études sur la Parole, pages 235–243.
HAL : hal-02798542.

Dans (Sehr *et al.*, 2010), nous pouvons apprendre que les 50 premières millisecondes de la réponse impulsionelle de la salle d'enregistrement (RIR pour Room Impulse Response) affectent peu les performances des systèmes de RAP, contrairement aux millisecondes suivantes (réverbération tardive). Dans (Junqua, 1997), les auteurs présentent une tentative de caractériser la sensibilité d'un dispositif de reconnaissance de phonèmes en fonction de la source de distorsion du canal. On peut voir que les grandes classes phonétiques ne sont pas affectées de la même manière par la réverbération. Dans (Parada *et al.*, 2014), les auteurs montrent la robustesse relative à la réverbération de chaque phonème, et proposent un modèle pour estimer la confusion de chaque phonème. La méthode utilise l'indice de clarté C50, qui est bien corrélé avec les performances de la RAP (Parada *et al.*, 2016).

Nous proposons d'analyser les erreurs des systèmes de RAP dues à la réverbération. Les résultats détaillés de la substitution des phonèmes et de la durée des phonèmes détectés (ou supprimés) permettront d'observer les raisons des mauvaises performances de la RAP pour la parole réverbérée. Plutôt que TIMIT (Fisher, 1986), nous avons choisi de travailler sur un autre corpus, le "Wall Street Journal" (WSJ) (Paul & Baker, 2003), avec une recette plus récente que les précédentes études sur le sujet pour créer notre système de RAP : utilisation de DNN (Deep Neural Network) et adaptation fMLLR (feature space Maximum Likelihood Linear Regression).

Dans la section 2, nous présentons le contexte de cette étude et nous rappelons le mécanisme de réverbération et ses différentes formes. Dans la section 3, nous exposons notre plan d'expérimentation, le matériel et les systèmes utilisés dans cette étude. Dans la section 4, nous présentons les résultats que nous discuterons dans la section 5.

2 Contexte

Supposons qu'un signal s qui est traité dans un environnement acoustique réaliste soit modélisé par :

$$y(t) = s(t) \circledast h(t) + n(t)$$

où t représente le temps, h la RIR et n le bruit de fond. Le RIR correspond à l'enregistrement d'un bruit impulsif (un clap) afin d'enregistrer les résonances. Dans ce modèle, nous pouvons remarquer que le bruit de fond est indépendant de la parole (distorsion additive), et la réverbération est fortement corrélée à la parole (distorsion convolutionnelle). La RIR décrit de manière précise la propriété de réverbération d'une pièce. La figure 1 présente un exemple schématique tiré de l'article (Valimaki *et al.*, 2012).

Le RIR est composé de trois parties :

- **Trajet direct** (« Direct Path ») : l'onde sonore est directement capturée par le microphone.
- **Réverbération précoce** (« Early Reverberation ») : les ondes sonores sont réfléchies une fois. La coloration spectrale du signal de parole est due à cette réverbération précoce. Cela n'affecte que très peu les performances de la RAP.
- **Réverbération tardive** (« Late Reverberation ») : les ondes sonores sont réfléchies plusieurs fois. Le flou temporel du signal de parole est dû à la réverbération tardive. Ceci affecte grandement les performances de la RAP.

La distorsion principalement gênante, provoquée par la réverbération, est le flou temporel. Le flou temporel provoque un chevauchement du phonème précédent sur le phonème actuel. Sur la figure 2, tirée de l'article (Petrick *et al.*, 2008), nous pouvons voir l'énergie des phonèmes qui se chevauchent. Dans cette illustration, **v** correspond au phonèmes voisés, **u** au phonèmes non-voisés et **r** a la

réverbération due au flou temporel. Maintenant la question est de savoir comment se comporte les systèmes de RAP losque les phonèmes se chevauchent ?

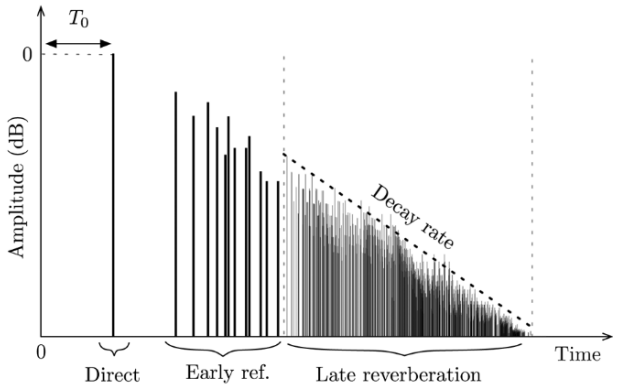


FIGURE 1 – Schéma d'une RIR générique, extrait de (Valimaki *et al.*, 2012).

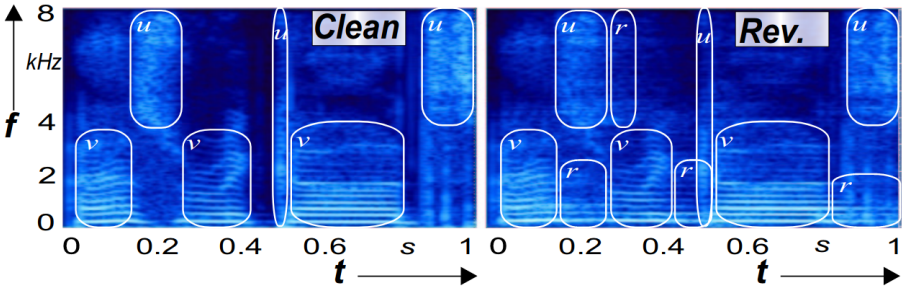


FIGURE 2 – Illustration des perturbations causées par la réverbération, extrait de (Petrick *et al.*, 2008).

3 Corpus et vérité terrain

Le corpus de parole utilisé vient du WSJ (Paul & Baker, 2003), et plus précisément le WSJ0 (Garofalo *et al.*, 1993) et le WSJ1 (Garofolo *et al.*, 1994). Les RIR provenant du REVERB Challenge (Kinoshita *et al.*, 2013) sont utilisées afin de réverbérer la parole artificiellement. Elles permettent de simuler 12 conditions différentes de réverbération : 6 pièces de tailles différentes (2 petites, 2 moyennes et 2 grandes) avec 2 types de distances au microphone (proche : 50 cm et lointaine : 200 cm). Les temps de réverbération T60 des petites, moyennes et grandes pièces sont respectivement d'environ 250, 500 et 700 millisecondes. Les salles de taille similaire ont été combinées, pour au final, obtenir 6 conditions réverbérées et une condition sans ajout de réverbération.

Pour créer le système de RAP, nous avons utilisé la recette Kaldi de Karel Vesely (Veselỳ *et al.*, 2013). Le système est hybride avec un réseau de neurones profond et des modèles de Markov caché (DNN-HMM) entraînés avec la cross-entropie sur le sous ensemble *train_si284* du WSJ (sans réverbération ajouté). Les données utilisées pour notre analyse sont composées des sous-ensembles *dev93* et *eval92* du WSJ qui ont été convolués par les différentes RIR pour obtenir 7 conditions de réverbération.

Afin d'analyser plus finement les erreurs commises par les systèmes de RAP, nous avons voulu nous détacher de l'influence du modèle de langage. Pour cela, nous utilisons un décodeur acoustico-phonétique pour prédire une suite de phonèmes (plutôt qu'une suite de mots). Pour concevoir ce

système, nous phonétisons au préalable les corpus d'entraînement et de test. Le dictionnaire de prononciation est modifié pour être composé uniquement des phonèmes possibles. Le modèle de langage est remplacé par un 1-gram appris sur un corpus de texte phonétisé.

4 Résultats

4.1 Décodage par le système de RAP

Nous avons décodé les sous-ensembles *dev93* et *eval92* pour les 7 conditions décrites dans la section 3. Les WER (Word Error Rate) et les PER (Phone Error Rate), liés à la taille de la pièce et à la distance au microphone, sont indiqués dans le tableau 1. Le terme « propre » correspond au fichier originel (sans convolution avec une réponse impulsionnelle).

TABLE 1 – Résultats de WER et PER en fonction des différentes conditions de réverbération : moyenne, écart-type et pourcentages de substitution, insertion, délétion de phonèmes.

Taille salle	Propre	Petite		Moyenne		Grande	
Distance		proche	loin	proche	loin	proche	loin
WER en %							
Moyenne	4,9	6,9	12,9	16,5	52,1	20,1	78,4
Écart-type	8,1	9,2	13,7	15,8	24,4	16,8	17,5
PER en %							
Moyenne	9,8	14,1	24,2	28,5	50,6	31,7	63,8
Écart-type	5,6	6,7	8,9	9,2	8,9	9,5	7,4
Ratio des erreurs des phonèmes en %							
Substitution	61,2	59,8	60,8	60,7	56,2	61,1	53,8
Insertion	15,3	15,5	11,4	9,4	4,0	9,2	2,1
Délétion	23,5	24,9	27,8	29,9	39,8	29,7	44,1
C50							
Moyenne		42,20	20,29	16,43	6,84	13,34	5,91

Dans les même conditions, nous pouvons constater que l'écart-type du WER est plus importante que l'écart-type du PER. Nous remarquons aussi, que passé un certain seuil de PER (environ 50%), le modèle de langage a plus de difficultés à retrouver les mots correct : une fois les 50% de PER atteint le WER est supérieur au PER.

Nous voyons que deux facteurs impactent fortement les performances des systèmes de transcriptions :

— la **taille de la pièce**. Plus le T60 est important et plus l'énergie provenant de la réverbération est importante.

— la **distance au microphone**. Plus la distance au microphone augmente et plus l'énergie de parole provenant du trajet direct est atténué.

Ces deux facteurs ont une influence directe sur la mesure du C50, qui est fortement corrélée avec les performances des systèmes de RAP (Parada *et al.*, 2016).

Comme la réverbération est une distorsion acoustique, nous allons dans la suite de cet article observer uniquement les erreurs du décodage phonétique car il est difficile de comprendre les erreurs commises

par les systèmes de RAP sans dissocier le modèle de langage.

Nous avons aussi observé, parmi les erreurs commises par le décodeur acoustico-phonétique, les ratios d'erreur provenant des substitutions, des insertions et des délétions des phonèmes dont les résultats sont visibles dans le tableau 1. Nous pouvons observer que la contribution des insertions est plus faible lorsque la réverbération augmente. Par contre, le nombre de délétions augmente. Les substitutions restent relativement stables, sauf dans les conditions les plus réverbérées où elles diminuent. Dans tous les cas testés, les substitutions restent la cause principale d'erreurs.

4.2 Durée des phonèmes

Comme la réverbération provoque un flou temporel du signal, nous avons décidé d'observer la durée des phonèmes transcrit par le décodeur acoustico-phonétique. Pour obtenir la durée de chaque phonème, nous avons utilisé les résultats issus de Kaldi. Cela implique que la durée des phonèmes est obtenue automatiquement (sans annotation manuelle) : les durées des phonèmes de référence seront les durées obtenues sur le décodage des tours de parole n'ayant pas été réverbérés artificiellement (condition propre) [1]. Dans un souci de lisibilité, nous avons choisi de regrouper les phonèmes par classe phonétique pour l'affichage des résultats.

Nous avons observé la durée des phonèmes corrects et incorrects (substitution et délétion) dans des conditions réverbérées que nous comparons avec la durée des phonèmes dans des conditions propres (non-réverbérées) (voir figure 3).

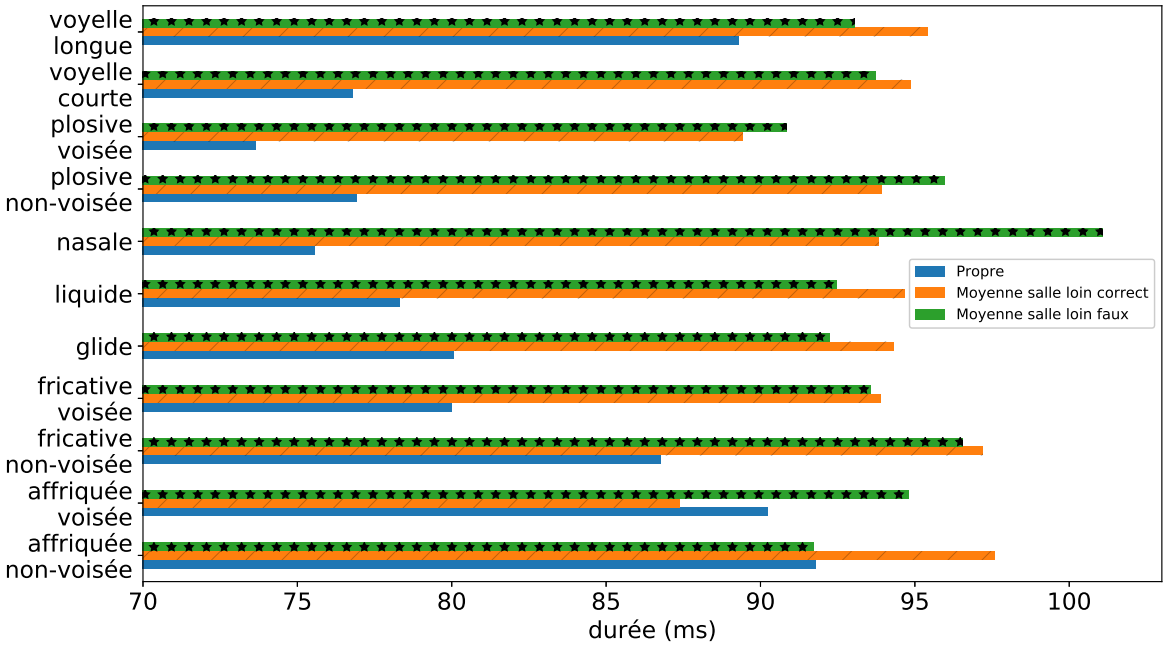


FIGURE 3 – Durée des phonèmes dans des conditions propre et réverbéré. Dans le cas réverbéré, les phonèmes correctement reconnus et les phonèmes erronés sont affichés séparément.

Sur la condition de réverbération salle moyenne et longue distance au microphone, la durée des

1. À noter que seuls les phonèmes correctement détectés dans des conditions propres sont pris en compte !

phonèmes réverbérés est globalement accrue pour attendre environ 95 ms en moyenne. Pour information, pour la condition *propre* la moyenne est de 82 ms, pour la condition *petite pièce lointaine* c'est 85 ms, et pour la condition *grande pièce lointaine* cela atteint 120 ms. Les phonèmes corrects ont globalement la même durée que les phonèmes substitués et insérés : hormis pour les consonnes affriquées et les nasales. En plus d'une augmentation de durée, nous avons aussi observé que les phonèmes tendent à avoir des durées similaires lorsque la réverbération s'accroît. Par exemple, la différence de durée entre voyelle courte et voyelle longue est clairement visible lorsque le fichier est non-réverbéré, mais lorsque le fichier est suffisamment réverbéré, les durées sont similaires.

Nous souhaitons maintenant analyser les phonèmes qui ont subit une délétion. Nous avons fait le lien entre les phonèmes supprimés dans des conditions réverbérées et leurs durées de référence (dans notre cas, cela correspond aux durées obtenues dans des conditions propres). Les résultats sont présentés dans le tableau 2.

TABLE 2 – Moyenne de la durée des phonèmes (condition propre) qui sont supprimés lorsque le fichier est réverbéré.

Taille salle	Petite		Moyenne		Grande	
Distance	proche	loin	proche	loin	proche	loin
Durée des délétion (ms)	58	61	66	68	67	71

Nous pouvons voir que les phonèmes qui sont supprimés ont en moyenne une durée moins importante. Ainsi, les phonèmes de faible durée ont plus de chance d'être supprimés lorsque la parole est réverbérée. Nous remarquons aussi que plus la réverbération augmente, et plus la durée moyenne des phonèmes supprimés augmente.

4.3 Résultats par classe phonétique

Commençons par observer le pourcentage de phonèmes correctement détectés en fonction de leur classe phonétique, dont les résultats sont affichés sur la figure 4a.

Nous avons choisi de montrer uniquement la condition de réverbération (moyenne salle, loin du microphone). Le comportement des résultats est similaire pour d'autres conditions de réverbération (plus la réverbération est importante et plus les résultats sont marqués). Sans réverbération, le système de transcription phonétique obtient des résultats similaires pour chaque phonème (environ 90% correct). Par contre, ce n'est pas le cas dans des conditions réverbérées. La classe phonétique des liquides est la moins impactée par la réverbération. Par contre, les consonnes affriquées voisées et les plosives sont les plus impactées. Les autres catégories sont moyennement impactées et obtiennent des résultats similaires. Nos résultats sont similaires à ceux trouvés dans cette étude (Junqua, 1997), ce qui montre que l'utilisation d'une recette de systèmes de RAP plus récente (DNN) ne modifie pas ces constats.

Afin d'identifier les erreurs, nous avons ensuite calculé une matrice de confusion entre classes phonétiques (figure 4b). Nous pouvons remarquer que les résultats sont très similaires [2].

Sur la figure 4a, l'impact des insertions et des substitutions entre même classe phonétique est pris

2. Hormis pour le cas des voyelles courtes (57% à 70%) qui s'explique par les insertions (25% des erreurs).

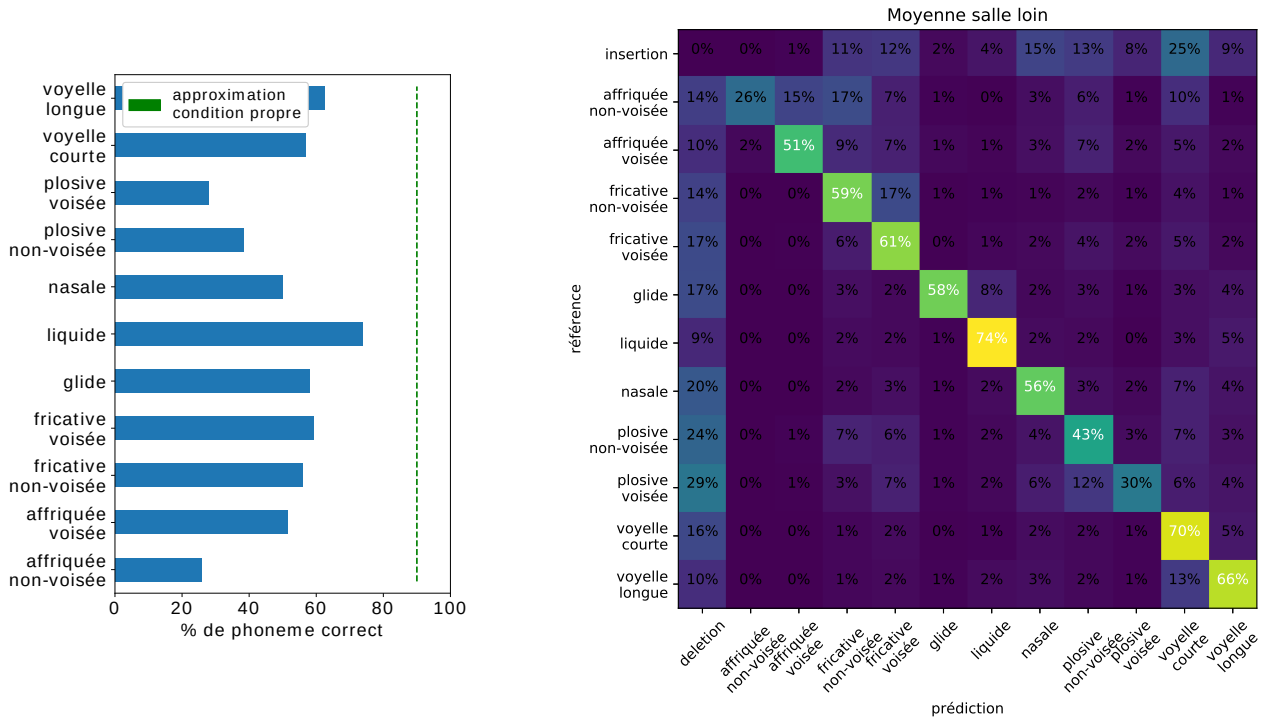


FIGURE 4 – (a) (partie à gauche) Pourcentage de phonèmes correctement détectés par classe phonétique (b) (partie de droite) et matrice de confusion dans une condition de réverbération forte (salle de taille moyenne, distance lointaine au micro).

en compte, ce qui n'est pas le cas lorsque nous observons la diagonale de la matrice de confusion (figure 4b). En effet, si nous observons les confusions entre phonèmes (et non au niveau des classes phonétiques), les substitutions sont majoritairement effectuées entre des classes phonétiques différentes (sauf pour les nasales).

Nous pouvons voir que les erreurs de détection proviennent principalement des suppressions, à l'exception des affriquées non-voisées, des fricatives non-voisées et des voyelles longues. Le nombre de délétions est toujours moins important que le nombre total de substitutions. Certaines classes phonétiques sont plus souvent supprimées que d'autres : les plosives et les nasales en particulier. Concernant les substitutions. Les fricatives non-voisées sont dans 17% des cas substituées par des fricatives voisées. La superposition avec le phonème précédent, provoquée par la réverbération, peut ajouter du voisement aux fricatives non-voisées. Le même effet se constate sur les affriquées, ou les non-voisées sont dans 17% des cas substituées par des voisées. Les affriquées non-voisées sont dans 17% des cas substituées par des fricatives non-voisées. Les affriquées partagent des caractéristiques communes aux plosives et aux fricatives. Le flou temporel provoqué par la réverbération rend plus similaire les affriquées aux fricatives car la composante plosive est fortement détériorée (étirée). Nous remarquons aussi que les plosives sont supprimées dans plus de 25% des cas. Les plosives se caractérisent par une phase de silence suivie par une impulsion. Lorsque que l'énergie provenant de la réverbération d'un précédent phonème se superpose aux plosives, il est plus bien plus difficile de détecté ce court silence. Concernant la substitution de 12% de plosives voisées en plosives non-voisées, l'effet inverse était plus attendu (de non-voisé à voisé).

Les autres substitutions remarquables, comme les plosives voisées qui sont substituées par des plosives non-voisées et les voyelles longues qui sont substituées par des voyelles courtes, sont plus difficilement explicables.

5 Discussion

La reconnaissance automatique de phonèmes dans la chaîne de traitement Kaldi est composée de deux étapes : l'alignement puis la classification. Le niveau de réverbération a un impact sur la précision de l'alignement des phonèmes. Nous pouvons le constater grâce à ces différentes observations :

— un allongement de la durée moyenne des phonèmes (82 ms pour la condition propre à 120 ms pour la condition la plus réverbérée),
— une uniformisation de la durées des phonèmes (voir figure 2),
— le ratio du nombre de délétions augmente (23,5% en condition propre et 44.1% dans la condition la plus réverbérée),
— la moyenne des durées des délétions de phonème augmente (de 58 ms pour une réverbération très faible à 71 ms pour la réverbération la plus forte),

Comme les phonèmes détectés sont de plus en plus long, de moins en moins de phonèmes peuvent être détectés. De plus, la réverbération cause de nombreuse erreurs de substitution, suite au chevauchement des phonèmes. Cela s'observe principalement par les nombreuses substitutions entre les phonèmes non-voisés et ceux voisés.

La distorsion du signal de parole provoquée par la réverbération est avant tout une superposition. L'effet d'un point de vue acoustique est similaire à de la conversation superposée. Les systèmes de transcription automatique de la parole sont conçus pour n'attribuer qu'un seul label phonétique par unités de temps. Or, dans des conditions réverbérées, le systèmes de RAP doit souvent choisir entre deux labels. Pour l'alignement, cela rend la détection de la transition entre deux phonèmes beaucoup plus difficile. Nous pensons que les systèmes automatiques favorisent l'allongement du phonème précédent plutôt que la transition vers le phonème suivant. Le début de certains phonèmes disparaît à cause de la superposition avec la réverbération du phonème précédent. Il est ainsi plus difficile de reconnaître ces phonèmes. Enfin, les phonèmes ayant une durée plus courte sont en général supprimés lorsqu'ils sont intégralement recouvert par la réverbération du phonème précédent : ce qui explique de nombreuses observations effectuées dans ce papier.

6 Conclusions

Dans cet article, nous avons analysé les erreurs de détection commises par les systèmes de RAP dues à la réverbération. Concernant la confusion entre phonèmes, nous avons obtenu des résultats similaires aux études précédentes, même avec une recette de RAP plus récente, utilisant des réseaux de neurones profond. De plus, nous avons montré les limites de l'alignement phonétique des systèmes de RAP actuels dans le cas de la parole réverbéré. Enfin, nous avons étudié les confusions en classes phonétiques, en expliquant lorsque cela a été possible les erreurs de substitution.

Parmi les perspectives de nos travaux, nous pensons que la détection d'anomalie de la durée des phonèmes (l'uniformisation de la durée par exemple) pourrait servir aux travaux de prédiction de la performance des systèmes de RAP, par exemple en utilisant les postériogrammes des phonèmes comme (Meyer *et al.*, 2017). Il serait également intéressant de comparer les erreurs phonétiques entre un système avec un alignement supervisé et un autre non supervisé. Il serait également pertinent de tester d'autres méthodes d'alignement comme les CTC (Connectionist Temporal Classification), afin de savoir si les erreurs d'alignement sont similaires à ce que nous avons observé. Enfin, nous pourrions observer la robustesse des systèmes de RAP appris sur des données réverbérées.

Références

FISHER W. M. (1986). The DARPA speech recognition research database : specifications and status. In *Proc. DARPA Workshop on Speech Recognition, Feb. 1986*, p. 93–99.

GAROFALO J. *et al.* (1993). CSR-I (WSJ0) complete LDC93S6A. Linguistic Data Consortium, Philadelphia, USA.

GAROFOLO J., GRAFF D., PAUL D. & PALLETT D. (1994). CSR-II (WSJ1) Complete. Linguistic Data Consortium, Philadelphia, USA.

JUNQUA J.-C. (1997). Impact of the unknown communication channel on automatic speech recognition : A review. In *Fifth European Conference on Speech Communication and Technology*.

KINOSHITA K., DELCROIX M., GANNOT S., P. HABETS E. A., HAEB-UMBACH R., KELLERMANN W., LEUTNANT V., MAAS R., NAKATANI T., RAJ B., SEHR A. & YOSHIOKA T. (2016). A summary of the REVERB challenge : state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, **2016**(1). DOI : 10.1186/s13634-016-0306-6.

KINOSHITA K., DELCROIX M., YOSHIOKA T., NAKATANI T., SEHR A., KELLERMANN W. & MAAS R. (2013). The REVERB challenge : A common evaluation framework for dereverberation and recognition of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, p. 1–4 : IEEE.

LIPPMANN R. (1996). Speech perception by humans and machines. In *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, p. 309–316.

MEYER B. T., MALLIDI S. H., KAYSER H. & HERMANSKY H. (2017). Predicting error rates for unknown data in automatic speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5330–5334 : IEEE.

PARADA P. P., SHARMA D., LAINEZ J., BARREDA D., VAN WATERSCHOOT T. & NAYLOR P. A. (2016). A single-channel non-intrusive C50 estimator correlated with speech recognition performance. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **24**(4), 719–732.

PARADA P. P., SHARMA D., NAYLOR P. A. & VAN WATERSCHOOT T. (2014). Reverberant speech recognition : A phoneme analysis. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, p. 567–571 : IEEE.

PAUL D. B. & BAKER J. M. (2003). The design for the Wall Street Journal-based CSR corpus. In *Proceedings Title*, volume II, p. 803–806 : IEEE.

PETRICK R., LOHDE K., LORENZ M. & HOFFMANN R. (2008). A new feature analysis method for robust asr in reverberant environments based on the harmonic structure of speech. In *Signal Processing Conference, 2008 16th European*, p. 1–5 : IEEE.

SEHR A., HABETS E. A., MAAS R. & KELLERMANN W. (2010). Towards a better understanding of the effect of reverberation on speech recognition performance. In *Proc. IWAENC*.

VALIMAKI V., PARKER J. D., SAVIOJA L., SMITH J. O. & ABEL J. S. (2012). Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(5), 1421–1448.

VESELỲ K., GHOSHAL A., BURGET L. & POVEY D. (2013). Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, p. 2345–2349.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Représentation du genre dans des données open source de parole

Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Avec l'essor de l'intelligence artificielle (IA) et l'utilisation croissante des architectures d'apprentissage profond, la question de l'éthique et de la transparence des systèmes d'IA est devenue une préoccupation centrale au sein de la communauté de recherche. Dans cet article, nous proposons une étude sur la représentation du genre dans les ressources de parole disponibles sur la plateforme *Open Speech and Language Resource*. Un tout premier résultat est la difficulté d'accès aux informations sur le genre des locuteurs. Ensuite, nous montrons que l'équilibre entre les catégories de genre dépend de diverses caractéristiques des corpus (discours élicité ou non, tâche adressée). En nous appuyant sur des travaux antérieurs, nous reprenons quelques principes concernant les métadonnées dans l'optique d'assurer une meilleure transparence des systèmes de parole construits à l'aide de ces corpus.

ABSTRACT

Gender representation in open source speech resources [1]

With the rise of artificial intelligence (AI) and the growing use of deep-learning architectures, the question of ethics and transparency in AI systems has become a central concern within the research community. We address transparency and fairness in spoken language systems by proposing a pilot study about gender representation in speech resources available through the *Open Speech and Language Resource* platform. We show that finding gender information in open source corpora is not straightforward and that gender balance depends on other corpus characteristics (elicited/non elicited speech, speech task targeted). The paper ends with recommendations about metadata and gender information for researchers in order to assure better transparency of the speech systems built using such corpora.

MOTS-CLÉS : traitement automatique de la parole, corpus, genre, locuteurs, données open source.

KEYWORDS: speech processing, corpora, gender, speakers, open source data.

1 Introduction

L'utilisation généralisée de l'apprentissage machine (*machine learning*) a fait des données un enjeu majeur de l'industrie et de la recherche. Les systèmes ont besoin d'une grande quantité de données étiquetées pour "apprendre" à modéliser correctement la tâche adressée. Les corpus de données, à l'instar de la puissance de calcul des machines, sont devenus de plus en plus grands, nous faisant entrer dans ce qu'on appelle aujourd'hui le *big data*. Mais outre la taille des corpus d'apprentissage, les chercheurs et chercheuses s'intéressent à une autre caractéristique de ces masses de données : leur

1. Publication originale en anglais, à LREC 2020.

Volume 1 : Journées d'Études sur la Parole, pages 244–252.
HAL : hal-02798544.

qualité. La notion de qualité des données peut trouver différentes définitions (Cai & Zhu, 2015), nous soutenons dans cet article que l'une des qualités principales de ces corpus est leur *transparence*.

Questionnant l'impact de tels outils sur nos sociétés, plusieurs études se sont intéressées aux biais existant dans ces systèmes : un cas bien connu dans le domaine du traitement automatique des langues (TAL) est l'exemple des plongements de mots (*word-embeddings*), avec les travaux de Bolukbasi *et al.* (2016) et de Caliskan *et al.* (2017) qui montrent le caractère socialement construit des données, encapsulant représentations et structures de pouvoir, incluant de fait les stéréotypes de genre. Des biais sexistes ont également été mis à jour pour des tâches de traduction automatique (Vanmassenhove *et al.*, 2018), ainsi que dans des systèmes de reconnaissance faciale (Buolamwini & Gebru, 2018). Dans une étude précédente, nous avons questionné l'impact du déséquilibre entre les catégories de genre dans les données d'entraînement sur les performances d'un système de reconnaissance automatique de la parole (RAP), montrant que la sous-représentation des femmes entraînait un biais de performance du système pour les locutrices (Garnerin *et al.*, 2019).

Dans la continuité de ce questionnement sur les liens entre représentations, données, et impacts sociétaux, nous étudions dans cet article la représentation du genre au sein d'une plateforme ouverte rassemblant des ressources langagières pour développer des outils de TAL. L'objectif de cette enquête est tout d'abord, d'observer les répartitions entre les catégories de genre au sein des corpus de parole. Cette répartition est envisagée en termes de nombre de locuteurs et de locutrices mais également en termes de temps de parole disponible pour chaque catégorie. Dans un second temps, nous proposons une réflexion sur les pratiques générales de mise à disposition de ressources, en nous appuyant sur quelques recommandations issues de travaux antérieurs.

2 OpenSLR

Open Speech Language Resources [2] (OpenSLR) est une plateforme créée par Daniel Povey, ayant pour objectif de centraliser des ressources langagières accessibles et téléchargeables gratuitement pour aider au développement de systèmes de parole. OpenSLR héberge actuellement 83 ressources. [3] Ces ressources sont constituées d'enregistrements audio transcrits, de logiciels, ainsi que de lexiques et de données textuelles nécessaires à la création de modèles de langue. D'autres plateformes proposent également ce type de ressources, la plupart du temps payantes. Nous nous concentrons donc sur les corpus de parole disponibles sur OpenSLR en raison de leur libre accès pour étudier à grande échelle la représentation du genre dans les corpus de parole.

Parmi les ressources disponibles sur la plateforme, nous avons analysé les 53 ressources de parole. Nous n'avons pris en compte ni les versions multiples d'une même ressource ni les sous-ensembles de ressources (e.g. LibriTTS, étant inclu dans LibriSpeech). Dans le cas de multi-versions, seule la dernière version a été conservée (e.g. TED LIUM). Plusieurs ressources contiennent de la parole dans différents dialectes ou langues et nous étudions chaque langue séparément. Nous avons ainsi un total de 66 corpus, dans 33 langues différentes avec 51 variantes dialectales/accentuelles. Les types de discours sont également variés (parole élicitée et lue, émissions radiophoniques, TEDTalks, enregistrements de réunions, appels téléphoniques, livres audio, etc.), ce qui n'est pas surprenant, étant donné le nombre d'acteurs ayant contribué sur la plateforme. Nous étudions cet échantillon

2. http://www.openslr.org.
3. Dernière consultation au 14 novembre 2019

Info. dispo.		#corpus
Non		24 (36.4%)
Oui	metadata	9 (13.6%)
	indexed	28 (42.4%)
	paper	5 (7.6%)
Total	-	**66**

TABLE 1 – Disponibilité des informations concernant le genre dans les corpus OpenSLR.

Info. dispo.	#corpus
Nombre d'individus	40
Nombre d'énoncés	32
Durée de parole	5
Nombre total de corpus	**42**

TABLE 2 – Type d'information disponible en fonction du genre dans les 42 corpus contenant des informations genrée.

pour aborder la question de la représentation du genre dans les corpus de parole. [4] OpenSLR ne fournissant pas de format défini et n'ayant pas d'exigences explicites concernant les structures de données, les ressources présentes sont également un bon reflet des pratiques des créateurs et créatrices de ressources concernant les méta-données.

3 Méthodologie

Afin d'étudier la représentation du genre dans les ressources disponible pour le traitement de la parole, commençons par définir ce que nous entendons par genre. Le genre est entendu ici comme variable binaire (homme/femme). Néanmoins, contrairement aux critiques faites par les sociologues J. Stacey et B. Thorne, qui dénoncent une cooptation du terme de genre (Stacey & Thorne, 1985), si le genre est envisagé ici comme une propriété des individus, il n'en reste pas moins l'expression d'un rapport social qui structure les relations et se retrouve dans les données de parole, de façon plus ou moins marquée en fonction des modalités de recueil des données langagières. Nous sommes également conscientes que les identités de genre sont plurielles et dépassent ces deux catégories, mais nous n'avons trouvé aucune mention de locuteurs ou locutrices non-binaires au sein des corpus étudiés dans notre étude.

Suite au travail de Doukhan & Carrive (2018), analysant la représentation du genre dans les flux télévisuels français, nous avons voulu explorer les corpus OpenSLR en réutilisant leurs notions de "taux de présence" (nombre d'individus) et de "taux d'expression" (temps de parole) pour rendre compte de la représentation du genre dans les données. Après téléchargement, ces informations ont donc été extraites manuellement des corpus.

3.1 Informations sur les locuteurs et locutrices et absence de méta-données

La première difficulté rencontrée est l'absence générale d'information (cf. Table 1). La prise en compte du genre dans la technologie étant un sujet de recherche relativement récent, les données démographiques sur le genre ne sont, la plupart du temps, pas mises à disposition par les créateurs et créatrices de ressources. Ainsi, en plus des caractéristiques générales du corpus mentionnées plus loin (voir 3.3), il nous semblait important de renseigner dans notre tableau final, si des informations sur le

4. Notre étude de cas ne prétend pas être exhaustive et il serait nécessaire d'inclure des ensembles de données fournies par des agences de ressources telles qu'ELRA ou LDC pour généraliser nos conclusions.

genre étaient fournies en premier lieu et le cas échéant de quelle manière. Les différentes modalités de cet attribut sont : *paper*, si un article a été explicitement cité dans la ressource, *metadata* si un fichier de métadonnées a été inclus, *indexed* si le genre a été explicitement indexé dans les données ou si les données ont été structurées en termes de genre.

3.2 Informations sur les durées et homogénéité des données

La deuxième difficulté concerne le fait que les informations sur le temps de parole ne sont pas standardisées, rendant impossible la comparaison de temps de parole entre individus ou entre catégories de genre (cf. Table 2). Lorsque des informations de durée sont fournies, la granularité utilisée varie selon les corpus. Certains auteurs indiquent les temps de parole en heures (e.g. (Panayotov *et al.*, 2015; Hernandez *et al.*, 2018)), d'autres le nombre d'énoncés ou de phrases (e.g. (Juan *et al.*, 2015; Google, 2019)), la définition de ces deux termes n'étant jamais explicite. Nous avons également constaté qu'il n'y avait pas de cohérence entre la durée de parole et le nombre d'énoncés, ce qui exclut la possibilité d'approximer l'une par l'autre.

3.3 Corpus

Le résultat final de notre analyse se traduit par un tableau [5] présentant toutes les caractéristiques des corpus. Les caractéristiques étudiées sont les suivantes : l'identifiant de la ressource (*id*) tel que défini sur OpenSLR ; la langue (*lang*) ; le dialecte ou l'accent s'il est spécifié (*dial*) ; le nombre total de locuteurs et locutrices ainsi que leur nombre dans chaque catégorie de genre (*#spk*, *#spk_m*, *#spk_f*) ; le nombre total d'énoncés ainsi que le nombre total d'énoncés par catégorie de genre (*#utt*, *#utt_m*, *#utt_f*) ; la durée totale, ou temps de parole, ainsi que la durée par catégorie de genre (*dur*, *dur_m*, *dur_f*) ; la taille de la ressource en gigaoctets (*sizeGB*) ainsi qu'un label qualitatif (*size*, prenant sa valeur entre "grand", "moyen", "petit") ; le taux d'échantillonnage (*sampling*) ; la tâche de discours ciblée pour la ressource (*task*) ; le caractère élicité ou non de la parole (*elicited*) [6] ; le statut de la langue (*lang_status*) : une langue est considérée comme ayant peu (low-resource) ou beaucoup (high-ressource) de ressources. Le statut de la langue est défini d'un point de vue technologique (c'est-à-dire : y a-t-il des ressources ou des systèmes de TAL disponibles pour cette langue ?) Il est fixé à la granularité de la langue (d'où le nom), quel que soit le dialecte ou l'accent (si renseigné) ; l'année de la publication (*year*) ; les auteurs et autrices de la ressource (*producer*).

4 Analyse

4.1 Disponibilité des informations sur le genre

Parmi nos 66 corpus, 36,4% ne fournissent aucune information sur le genre des locuteurs et locutrices. Plus de 20% des corpus ne fournissent aucune information sur les locuteurs et locutrices, quelle qu'elle soit. La Table 1 résume le nombre de corpus pour lesquels des informations de genre ont été

5. Le tableau final et le script utilisé pour l'analyse sont disponibles à l'adresse suivante https://github.com/mgarnerin/openslr_gender_survey

6. Nous définissons comme données de parole non-élicitées, des données qui auraient existé sans la création des ressources (par exemple : TedTalks, livres audio, etc.), les autres données de parole sont considérées comme élicitées

fournies et, le cas échéant, l'endroit où celles-ci ont été trouvées. La procédure de recherche était la suivante : nous avons d'abord examiné le fichier de métadonnées (si existant) et dans le cas contraire, nous avons cherché si le genre était indexé dans la structure des données. Si aucune information n'était trouvée, nous avons cherché s'il existait un article décrivant les données.

La Table 2 indique les types d'information renseignées dans le sous-ensemble des 42 corpus contenant des informations sur le genre des locuteurs et locutrices. La plupart du temps, seul le nombre d'individus dans chaque catégorie est indiqué ; cinq corpus fournissent également le temps de parole pour chaque catégorie. De ce fait, nous n'avons pas pu étudier le taux d'expression de chaque catégorie, comme dans le travail de Doukhan & Carrive (2018), mais nous avons analysé le nombre d'énoncés lorsque renseigné. Il convient toutefois de rappeler que la notion d'énoncé n'est jamais définie dans les ressources (le découpage est-il syntagmatique ? basé sur les groupes de souffles ou du aux limites techniques du système ?), il n'existe donc pas de cohérence entre nombre d'énoncés et temps de parole, et ces résultats sont à prendre avec prudence. En plus des 42 corpus pour lesquels nous avons réussi à trouver des informations sur le genre, nous avons recueilli manuellement ces informations pour 4 autres corpus, atteignant une taille d'échantillon finale de 46 corpus.

4.2 Genre et taux de présence

Parole élicitée vs non-élicitée. Lorsqu'on analyse le taux de présence de chaque catégorie de genre dans notre échantillon, la parité est atteinte avec 3 050 locutrices et 3 022 locuteurs. Cependant, certaines données sont pré-existantes à la création de ressources, notamment les données issues des médias, dans lesquels les femmes sont moins représentées (Macharia *et al.*, 2015). Le même résultat a d'ailleurs été mis en avant par l'étude de Doukhan & Carrive (2018). Nous avons donc croisé cette répartition avec le caractère élicité ou non de la parole, considérant comme non-élicitée toute parole qui aurait existé indépendamment de la création du corpus (e.g. TEDTalks, les interviews, les émissions de radio, etc.) Les résultats sont présentés dans la Table 3. Dans les deux cas (parole élicitée, respectivement non-élicitée), la différence entre les genres est relativement faible (5,6 points, respectivement 5,8 points), loin des 30 points de différence observés dans (Garnerin *et al.*, 2019). Une explication possible de cette observation est que les corpus, élicités ou non, restent le résultat d'un processus contrôlé, de sorte que la disparité homme/femme sera réduite autant que possible par les créateurs et créatrices des corpus. Cependant, on remarque qu'hormis Librispeech (Panayotov *et al.*, 2015), tous les corpus non élicités sont de petits corpus. En retirant Librispeech de l'analyse, nous observons un rapport femme/homme de 1/3-2/3, ce qui semble cohérent avec nos résultats précédents.

On peut donc conclure que la disparité de genre n'est observable que lorsque les données ne sont pas élicitées ou sciemment équilibrées. Cette représentation déséquilibrée n'est donc pas observée à l'échelle de l'ensemble de la plate-forme OpenSLR, la majorité des corpus étant élicités (89,1%). Ces résultats démontrent une volonté d'assurer la parité durant le processus de création des corpus.

"How can I help ?" : l'impact de la tâche. Lorsque les corpus de parole sont construits pour l'entraînement de systèmes ce sont la plupart du temps des systèmes reconnaissance de la parole (RAP) ou de synthèse vocale. En croisant la représentation du genre avec la tâche adressée, nous obtenons les résultats reportés dans la Table 4. Nous observons que si les taux de présence sont presque équilibrés au sein des corpus de RAP, les femmes sont mieux représentées dans les ensembles de données pour la synthèse. Cette observation fait écho au rapport de recommandation de l'ONU pour une éducation numérique égalitaire entre les sexes, qui indique qu'aujourd'hui la plupart des assistants vocaux ont une voix de femme en abordant les problèmes éducatifs et sociétaux que cela

Type de parole	#corpus	#F	#H
Élicitée	41	1782	1596
		52.8%	47.2%
Non-élicitée	5	1268	1426
		47.1%	52.9%
Non-elicitée	4	67	143
(sans Librispeech)		31.9%	68.1%

TABLE 3 – Taux de présence en fonction du type de parole

Tâche	#corpus	#F	#H
Reco.	12	2523	2615
		49.1%	50.9%
Synthèse	10	124	70
		63.9%	36.1%
NA	25	403	337
		54.5%	45.5%

TABLE 4 – Taux de présence en fonction de la tâche

	F	M
Nombre de loc.	591	551
	51.8%	48.2%
Nombre d'énoncés	72,280	143,342
	33.5%	66.5%

TABLE 5 – Nombre d'énoncés par catégorie de genre pour les 32 corpus fournissant ces informations. *N.B : deux corpus reportaient uniquement des nombres d'énoncés, le nombre de loc. est donc donné à titre indicatif*

soulève (West *et al.*, 2019). Cette conception genrée des assistants vocaux est parfois justifiée par des stéréotypes tels que "les voix féminines sont perçues comme plus serviables, plus sympathiques ou plus agréables". Les systèmes de synthèse vocale étant souvent utilisés pour créer des assistants vocaux, on peut supposer que l'utilisation de voix féminines est devenue pratique courante pour garantir l'adhésion du public au système. Cette affirmation peut toutefois être nuancée, notamment par les travaux de Nass & Brave (2005) qui ont montré que d'autres facteurs pouvaient justifier l'utilisation de voix féminines, tels que l'identification sociale et les stéréotypes culturels liés au genre.

4.3 Genre et taux d'expression

En raison d'un manque global d'informations sur le temps de parole, nous n'avons pas analysé le taux d'expression par catégorie. Cependant, le nombre d'énoncés est souvent renseigné, ou facilement retrouvable dans les corpus, et nous avons pu récupérer des fréquences par catégorie de genre pour 32 corpus. Si l'équilibre entre hommes/femmes est presque atteint d'un point de vue du taux de présence, les hommes sont plus représentés lorsqu'on s'intéresse au nombre d'énoncés (voir Table 5). Cependant, cette disparité n'est en réalité que l'effet de trois corpus contenant 51 463 et 26 567 (Korvas *et al.*, 2014) et 8376 (Hernandez-Mena, 2019) énoncés de locuteurs, alors que le nombre moyen d'énoncés par corpus est respectivement de 1942 pour les hommes et 1983 pour les femmes. Après avoir retiré ces trois valeurs extrêmes, la quantité de parole est équilibrée entre les catégories de genre. Le nombre élevé d'énoncés des trois valeurs extrêmes est cependant surprenant, ces trois corpus étant petits (2,1 Go, 2,8 Go) et moyens (5,2 Go). Cela met une fois de plus en évidence le problème de la notion d'énoncé (*sentence* ou *utterances*) qui n'est jamais explicitement définie. Une telle différence de granularité rend donc difficile la comparaison entre les corpus.

5 Recommendations

L'impact social du *big data* et les problèmes éthiques soulevés par les systèmes de TAL ont déjà été abordés dans des travaux antérieurs. Wilkinson *et al.* (2016) ont élaboré des principes pour la gestion des données scientifiques, les principes FAIR Data, basés sur quatre caractéristiques fondamentales des données qui sont la repérabilité (*foundability*), l'accessibilité (*accessibility*), l'interopérabilité (*interoperability*) et la réutilisabilité (*reusability*). Dans notre cas, la repérabilité et l'accessibilité sont prises en compte dès la conception, les ressources sur OpenSLR étant librement accessibles. L'interopérabilité et la réutilisabilité des données ne sont cependant pas encore atteintes. Une autre discussion sur la description des données au sein de la communauté du TAL a été initiée par Couillault *et al.* (2014), qui ont proposé une Charte sur l'éthique et les *big data* (*Ethics and Big Data Charter*), pour aider les créateurs et créatrices de ressources à décrire leurs données d'un point de vue juridique et éthique. Le travail de Hovy & Spruit (2016) a mis en évidence l'articulation complexe entre données, systèmes de TAL et leurs différentes implications sociales, avec, entre autres, les notions d'*exclusion*, de *surgénéralisation* et d'*exposition*. Plus récemment, les travaux de Bender & Friedman (2018) ont proposé la notion de *data statement* pour garantir la transparence des données. Nous espérons que la présente étude encouragera les chercheurs et chercheuses à décrire de manière exhaustive leurs ensembles de données, en suivant les lignes directrices proposées ci-dessus.

Sur l'importance des méta-données. La première conclusion de notre enquête est qu'il n'est pas facile d'obtenir une description exhaustive sur les locuteurs et locutrices dans les ressources de parole. Ce manque de méta-données est problématique d'un point de vue scientifique, car il empêche de garantir la généralisation des systèmes ou des résultats linguistiques basés sur ces corpus, comme le soulignent Bender & Friedman (2018), mais également éthique rendant impossible tout contrôle quant à l'existence d'une disparité de représentation pouvant conduire à des biais. Cette absence d'informations contextuelles sur la parole traduit aussi une conception du langage comme entité abstraite, plutôt que comme production située, qui mérite d'être questionnée (Hovy & Spruit, 2016).

Lorsque des informations sur la représentation du genre dans les données étaient fournies, celles-ci se portaient majoritairement sur le nombre de locuteurs et locutrices. Il serait intéressant d'avoir également accès à la durée des ensembles de données en heures ou minutes, globalement et par individu et/ou catégorie de genre. Taux de présence et taux d'expression n'étant pas égaux, l'un mesurant la représentation de chaque catégorie et l'autre la quantité de données disponibles. Des informations de durée standardisées pourraient permettre de vérifier rapidement l'équilibre entre les catégories de genre, sans s'appuyer sur une notion d'énoncé peu fiable. Lors de la collecte des données, nous avons remarqué que plus les ressources étaient récentes, plus il était facile de trouver des informations sur le genre, attestant de la visibilité croissante des thématiques de genre dans la technologie, mais si ce travail descriptif et important pour les futurs corpus, il doit également être effectué pour les ensembles de données déjà publiés, car ils sont susceptibles d'être utilisés à nouveau par la communauté.

Transparence dans l'évaluation. Le taux d'erreur-mots (WER pour *word error rate*) est généralement calculé comme la somme des erreurs commises sur l'ensemble des données de test divisée par le nombre total de mots dans la référence. Mais si une telle évaluation permet de comparer facilement les systèmes, elle ne tient pas compte de leurs variations de performance. Dans notre enquête, 13 des 66 corpus étaient accompagnés d'un article décrivant les ressources. Lorsque les performance des systèmes de RAP étaient reportées, aucune évaluation en terme de genre n'était faite, même si des informations sur la représentation du genre dans les données étaient renseignées. La communication

des résultats pour les différentes catégories est le moyen le plus simple de vérifier l'absence de biais dans les performances. Décrire ses données est un premier pas, mais pour une science ouverte et juste, l'étape suivante devrait être de prendre également en compte ces informations dans le processus d'évaluation. Un travail récent dans ce sens a été réalisé par (Mitchell *et al.*, 2019) qui a proposé de décrire les performances des modèles dans des "cartes modèles" (*model cards*), encourageant ainsi un rapport transparent des résultats.

6 Conclusion

Dans notre enquête sur le genre dans les corpus disponibles sur la plateforme OpenSLR, nous observons les tendances suivantes : la parité est globalement atteinte, mais les interactions avec d'autres caractéristiques des corpus révèlent que la disparité homme/femme nécessite plus qu'un simple nombre d'intervenants pour être identifiée. Dans les données non élicitées (c'est-à-dire toutes données de paroles qui auraient existé sans la création d'un corpus, comme les TEDTalks ou les émissions radiophoniques), nous avons constaté que, sauf dans le cas de Librispeech où l'équilibre entre les catégories de genre est contrôlé, les hommes sont plus représentés que les femmes. Il semble également que la plupart des corpus visant à développer les systèmes synthèse vocale contiennent principalement des voix féminines, peut-être en raison du stéréotype associant la voix féminine aux activités de *care*. Nous observons également que la description genrée des données a été prise en compte par la communauté, avec un nombre croissant de corpus fournis avec des métadonnées sur le genre au cours des deux dernières années. Notre échantillon ne contenant que 66 corpus, nous reconnaissons que nos résultats ne peuvent pas nécessairement être étendus à toutes les ressources linguistiques, mais cela nous permet relancer le débat sur les pratiques générales de description des corpus, soulignant le manque de méta-données, et d'actualiser le discours autour des implications sociales des systèmes de TAL.

Références

BENDER E. M. & FRIEDMAN B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, **6**, 587–604.

BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? Debiasing word embeddings. In *Actes de NeurIPS 2016 (Neural Information Processing Systems)*, p. 4349–4357.

BUOLAMWINI J. & GEBRU T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. In *Actes de FAT 2018 (Fairness, Accountability and Transparency)*, p. 77–91, New-York City, USA : ACM.

CAI L. & ZHU Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science Journal*, **14**.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

COUILLAULT A., FORT K., ADDA G. & MAZANCOURT H. (2014). Evaluating corpora documentation with regards to the ethics and big data charter. In *Actes de LREC 2014 (Language Resources and Evaluation)*, p. 4225–4229, Reykjavik, Islande : ELRA.

DOUKHAN D. & CARRIVE J. (2018). Description automatique du taux d'expression des femmes dans les flux télévisuels français. In *Actes de JEP 2018 (Journées d'Études sur la Parole)*, p. 496–504, Aix-en-Provence, France.

GARNERIN M., ROSSATO S. & BESACIER L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *Actes de AI4TV 2019 (Workshop on AI for Smart TV Content Production, Access and Delivery)*, p. 3–9, Nice, France : ACM.

GOOGLE (2019). Crowdsourced high-quality UK and Ireland English Dialect speech data set. Web download at http ://www.openslr.org/83/.

HERNANDEZ F., NGUYEN V., GHANNAY S., TOMASHENKO N. & ESTÈVE Y. (2018). TED-LIUM 3 : Twice as much data and corpus repartition for experiments on speaker adaptation. In *Actes de SPECOM 2018 (Speech and Computer)*, p. 198–208, Leipzig, Allemagne : Springer.

HERNANDEZ-MENA C. D. (2019). TEDx spanish corpus. audio and transcripts in spanish taken from the tedx talks ; shared under the CC BY-NC-ND 4.0 license. Web Download.

HOVY D. & SPRUIT S. L. (2016). The social impact of Natural Language Processing. In *Actes de ACL 2016 (Volume 2 : Short Papers)*, p. 591–598, Berlin, Allemagne : Association for Computational Linguistics.

JUAN S. S., BESACIER L., LECOUTEUX B. & DYAB M. (2015). Using resources from a closely-related language to develop ASR for a very under-resourced language : a case study for Iban. In *Actes de INTERSPEECH 2015 (International Speech Communication Association)*, p. 1270–1274, Dresde, Allemange : ISCA.

KORVAS M., PLÁTEK O., DUŠEK O., ŽILKA L. & JURČÍČEK F. (2014). Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Actes de LREC 2014 (Language Resources and Evaluation)*, p. 4423–4428, Reykjavik, Islande : ELRA.

MACHARIA S., NDANGAM L., SABOOR M., FRANKE E., PARR S. & OPOKU E. (2015). Who makes the news. Global Media Monitoring Project (GMMP).

MITCHELL M., WU S., ZALDIVAR A., BARNES P., VASSERMAN L., HUTCHINSON B., SPITZER E., RAJI I. D. & GEBRU T. (2019). Model cards for model reporting. In *Actes de FAT 2019 (Fairness, Accountability and Transparency)*, p. 220–229, Atlanta, GA, USA : ACM.

NASS C. & BRAVE S. (2005). *Wired for Speech : How Voice Activates and Advances the Human-computer Relationship*. MIT Press.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an ASR corpus based on public domain audio books. In *Actes de ICASSP 2015 (Acoustics, Speech and Signal Processing)*, p. 5206–5210, Brisbane, Australie : IEEE.

STACEY J. & THORNE B. (1985). The missing feminist revolution in sociology. *Social problems*, **32**(4), 301–316.

VANMASSENHOVE E., HARMEIER C. & WAY A. (2018). Getting gender right in neural machin translation. In *Actes de EMNLP 2018 (Empirical Methods in Natural Language Processing)*, p. 3003–3008, Bruxelles, Belgique.

WEST M., KRAUT R. & EI CHEW H. (2019). I'd blush if I could : closing gender divides in digital skills through education.

WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J., APPLETON G., AXTON M., BAAK A., BLOMBERG N., BOITEN J.-W., DA SILVA SANTOS L. B., BOURNE P. E. *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, **3**.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Reconnaissance de phones fondée sur du Transfer Learning pour des enfants apprenants lecteurs en environnement de classe

Lucile Gelin[1, 2] Morgane Daniel[2] Thomas Pellegrini[1] Julien Pinquier[1]
(1) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
(2) Lalilo, Paris, France
lucile.gelin@irit.fr, morgane@lalilo.com

RÉSUMÉ

A conditions égales, les performances actuelles de la reconnaissance vocale pour enfants sont inférieures à celles des systèmes pour adultes. La parole des jeunes enfants est particulièrement difficile à reconnaître, et les données disponibles sont rares. En outre, pour notre application d'assistant de lecture pour les enfants de 5-7 ans, les modèles doivent s'adapter à une lecture lente, des disfluences et du bruit de brouhaha typique d'une classe. Nous comparons ici plusieurs modèles acoustiques pour la reconnaissance de phones sur de la parole lue d'enfant avec des données bruitées et en quantité limitée. Nous montrons que faire du *Transfer Learning* avec des modèles entraînés sur la parole d'adulte et trois heures de parole d'enfant améliore le taux d'erreur au niveau du phone (PER) de 7,6% relatifs, par rapport à un modèle enfant. La normalisation de la longueur du conduit vocal sur la parole d'adulte réduit ce taux d'erreur de 5,1% relatifs supplémentaires, atteignant un PER de 37,1%.

ABSTRACT

Transfer Learning based phone recognition on children learning to read, with speech recorded in a classroom environment

Current performance of speech recognition for children is below that of the state-of-the-art for adult speech. Young children's speech is particularly difficult to recognise, and substantial corpora are missing to train acoustic models. Furthermore, in the scope of our reading assistant for 5-7-year-old children learning to read, models need to cope with slow reading rate, disfluencies, and classroom-typical babble noise. In this paper, we compare acoustic models for phone recognition on child speech using data that is very noisy and limited in quantity. We show that transfer learning with adult-trained time-delay neural networks and three hours of child speech improves the phone error rate by 7.6% relative, over a model trained on child speech. The addition of vocal tract length normalisation on adult speech further reduces the error rate by 5.1% relative, reaching a PER of 37.1%.

MOTS-CLÉS : Reconnaissance de phones, parole d'enfant, apprentissage par transfert, normalisation de la longueur du conduit vocal, réseau de neurones à délai temporel.

KEYWORDS: Phone recognition, child speech, transfer learning, vocal tract length normalisation, time-delay neural network.

1 Introduction

Le corps humain, et en particulier l'appareil de production de la parole, évolue continuellement pendant les premières années de la vie. Entre 5 et 7 ans, les mécanismes articulatoires des enfants ne sont pas stables, ce qui implique une variabilité spectrale intra- et inter-locuteurs. La stabilisation

Volume 1 : Journées d'Études sur la Parole, pages 253–261.
HAL : hal-02798545.

du contrôle vocal ne se produit que vers l'âge de 8 ans (Lee *et al.*, 1999). En raison de la croissance lente du conduit vocal, leurs fréquence fondamentale et formants n'atteignent les niveaux d'un adulte mature qu'à l'âge de 15 ans (Mugitani & Hiroya, 2012). En outre, les erreurs phonologiques, comme la suppression de syllabes faibles ou la substitution de phones due à un mauvais positionnement de la langue et des lèvres (Fringi *et al.*, 2015), sont très courantes dans la parole des jeunes enfants, et ont tendance à disparaître avec l'âge. Ces différences morphologiques et phonologiques sont les principales causes des faibles performances des systèmes de Reconnaissance Automatique de la Parole (RAP) sur les voix d'enfants.

Les tuteurs numériques de lecture ont un fort impact pédagogique sur les enfants qui apprennent à lire, et plusieurs projets ont vu le jour au fil des ans (Mostow & Aist, 2001; Bolaños *et al.*, 2011; Proença, 2018). Le projet FLUENCE notamment travaille depuis quelques années sur l'évaluation automatique de la fluence auprès des apprenants lecteurs français (Godde *et al.*, 2017). Travailler sur la parole des lecteurs non-experts ajoute des difficultés dues à la présence de nombreuses disfluences.

Lalilo [1] propose un assistant de lecture pour les enfants de 5 à 7 ans, avec un exercice de lecture à voix haute. Pour cela, nous entraînons un classifieur qui donne une décision binaire sur la lecture correcte ou non du mot. Cet article présente nos travaux sur la modélisation acoustique, avec pour objectif l'amélioration de la précision de la reconnaissance automatique de phones, dont découle la précision du classifieur.

Dans la section 2, nous présentons nos motivations et les techniques utilisées. Le dispositif expérimental est détaillé dans la section 3, suivi des résultats dans la section 4. Enfin, nous analysons le comportement de nos modèles en fonction du voisement, et en présence d'erreurs de lecture.

2 Méthodes

En raison du peu de données de parole d'enfant disponibles, les systèmes fondés sur les réseaux neuronaux profonds n'ont commencé à être exploités que récemment pour la reconnaissance automatique de la parole d'enfant. Les architectures hybrides Deep Neural Network - Hidden Markov Model (DNN-HMM) de modèles acoustiques ont été largement utilisées dans la reconnaissance vocale au cours des deux dernières décennies. (Serizel & Giuliani, 2014a) utilisent un système DNN-HMM, entraîné conjointement sur des données adultes et enfants, puis adapté à la parole d'un groupe d'âge spécifique. Pour une application d'apprentissage d'une seconde langue chez les enfants, (Metallinou & Cheng, 2014) ont présenté un DNN-HMM qui, même entraîné sur très peu de données, a surpassé des systèmes basés sur des mélanges de lois gaussiennes (GMM-HMM).

Dans cet article, nous utilisons un système hybride où le DNN est un Time-Delay Neural Network (TDNN), architecture introduite par (Waibel *et al.*, 1989) pour la reconnaissance de phones. Ce type de réseau s'est révélé particulièrement adapté à la RAP, de par sa capacité à représenter les relations entre événements acoustiques dans le temps, ainsi qu'à fournir une invariance temporelle des paramètres appris par le réseau (Waibel *et al.*, 1989). Pour cela, la largeur du contexte varie selon les couches : les couches inférieures apprennent des caractéristiques acoustico-phonétiques de courte durée, tandis que les couches supérieures apprennent des caractéristiques plus complexes de plus longue durée. Les TDNN ont été utilisés avec succès pour la reconnaissance des voyelles sur la parole des enfants dans une langue peu dotée (Yong & Ting, 2011).

Le Transfer Learning (TL) permet de surmonter le manque de données dans un domaine spécifique : cette méthode consiste à transférer des connaissances préalables acquises sur un grand corpus hors

1. https://www.lalilo.com/

domaine vers un modèle acoustique entraîné sur un petit corpus correspondant à l'application. Le système bénéficie ainsi de connaissances de bases qui peuvent être adaptées à un domaine ou à une tâche spécifique. Cette méthode a montré une amélioration des performances dans des applications de transfert de langues (Shi *et al.*, 2018) et de transfert d'âge (Shivakumar & Georgiou, 2018).

Nous explorons ici l'adaptation de modèles acoustiques en utilisant un modèle TDNN entraîné sur un grand corpus de parole adulte, et deux méthodes de TL. La première consiste à ré-initialiser aléatoirement les deux dernières couches du réseau, et à les entraîner avec des données d'enfant, ce qui a pour objectif d'adapter fortement le modèle à la parole d'enfant. La seconde prend toutes les couches existantes du modèle source et les ré-entraîne avec des données d'enfant, et ainsi garde toutes les informations apprises par le modèle sur les voix d'adulte. Nous utilisons différents facteurs d'apprentissage pour équilibrer les connaissances pré-acquises sur la parole d'adulte et les caractéristiques acoustiques nouvellement acquises sur la parole d'enfant.

Une adaptation approfondie est réalisée avec la technique VTLN. Tandis que Serizel et al. (Serizel & Giuliani, 2014b) utilisent la VTLN sur un corpus mixte enfants-adultes en normalisant les caractéristiques de chaque locuteur, nous proposons d'utiliser la VTLN pour étendre les fréquences d'adulte vers des fréquences d'enfant et entraîner un TDNN adulte sur ces paramètres transformés. La VTLN n'est donc pas appliqué de façon individuelle à chaque locuteur, mais de façon globale à tous les locuteurs avec des facteurs de déformation fixes : un pour les femmes et un pour les hommes. Cela nous permet d'obtenir une gamme de fréquences proche de celle des enfants pour chaque locuteur, minimisant ainsi la différence entre parole d'adulte et parole d'enfant, avec pour objectif une meilleure efficacité du transfer learning. La principale contribution réside dans l'utilisation de ce TDNN adulte adapté aux voix d'enfants avec de la VTLN comme modèle source pour le transfer learning.

3 Dispositif expérimental

3.1 Données de parole

Nous utilisons deux jeux de données de parole en français : le corpus adulte *Commonvoice* [2], et un corpus enfant interne, appelé *Lalilo* par la suite. Le tableau 1 présente des informations sur ces deux corpus.

TABLE 1 – Informations sur les données de parole

Corpus	**Commonvoice**		**Lalilo**		
Set	Train	Test	Train	Test C	Test I
Nb locuteurs	78	268	562	69	153
Durée (h)	20,0	9,0	3,8	0,4	0,4
Durée moyenne (s)					
Par enregistrement	3,6	3,8	7,0	2,3	3,4
Par locuteur	127,4	-	22,0	-	-
RSB moyen (dB)	35,3 ± 16,1	32,9 ±14,2	25,6 ± 13,9	23,9 ±11,6	20,5 ± 11,9

Le jeu de données Commonvoice est composé de phrases lues par des adultes, tâche qui se rapproche de la lecture à voix haute des enfants. 72% des locuteurs sont masculins, et 7% féminin, les autres locuteurs n'ayant pas fourni cette information. Chaque enregistrement a été validé par 2 ou 3 annotateurs, le corpus ne contient donc que peu d'erreurs. En outre, il présente un rapport signal à bruit (RSB) moyen élevé.

2. Corpus disponible : https://voice.mozilla.org/fr

Le corpus Lalilo contient des enregistrements d'enfants de la grande section au CE1, âgés de 5 à 7 ans, lisant à haute voix des mots isolés, des phrases et des histoires courtes. Les enregistrements ont été recueillis soit directement dans les écoles, soit par le biais d'un exercice de lecture à voix haute sur la plateforme Lalilo. Dans le premier cas, les conditions environnementales sont relativement propres : un microphone de bonne qualité est utilisé, et le niveau de bruit est contrôlé. Dans le second cas, cependant, les enseignants laissent généralement un petit groupe d'élèves travailler en autonomie sur la plateforme, ce qui implique inévitablement la présence de bruit de brouhaha sur les enregistrements. Le tableau 1 affiche la moyenne et l'écart-type du RSB pour chaque ensemble de données. Par rapport au corpus Commonvoice, le RSB moyen est significativement plus faible.

Les données d'apprentissage ne contiennent que des histoires, phrases ou mots correctement prononcés et lus avec fluidité. En accord avec la tâche de l'assistant numérique qui vise à détecter les erreurs de déchiffrage et de fluence sur des mots isolés, l'ensemble de test est formé uniquement de mots isolés [3], la durée moyenne des enregistrements est donc plus faible que celle du corpus d'apprentissage. Les phones réellement lus par les élèves ont été transcrits par deux juges humains. Chaque mot a également été classé entre trois catégories :

— Correct : lecture correcte et fluide.
— Erreur de fluence (*Fluence*) : lecture correcte mais non fluide (hésitations, faux départs...)
— Erreur de déchiffrage (*Déchiffrage*) : lecture incorrecte, avec au moins un phone mal lu.

Deux sous-ensembles de test ont été créés à partir de ces catégories : Le test C contient les mots bien lus, et le test I contient les mots qui comportent des erreurs de lecture, c'est-à-dire de fluence ou de déchiffrage. Dans la catégorie *Déchiffrage*, les phones peuvent être soit substitués, soit supprimés, soit insérés. Les erreurs de déchiffrage prévalent sur les erreurs de fluence, car les premières sont plus répréhensibles que les secondes lors de l'évaluation du niveau de lecture d'un enfant. Ainsi, les mots classés comme erreurs de déchiffrage peuvent également contenir des erreurs de fluence.

3.2 Système de reconnaissance de phones

Toutes les expériences sont réalisées avec l'outil Kaldi (Povey *et al.*, 2011).

3.2.1 Paramètres acoustiques

Pour les modèles GMM-HMM servant à la génération des alignements pour l'entraînement des TDNN, les paramètres sont des Mel-frequency cepstral coefficients (MFCC) de dimension 13 avec une fenêtre de 25 ms et un décalage de 10 ms, auxquels nous ajoutons des dérivées première et seconde. Les TDNN sont alimentés par des MFCC haute résolution de dimension 40. Nous effectuons également de l'augmentation de données en déformant temporellement le son brut avec des facteurs de 0,9, 1,0 et 1,1. Dans la plupart des systèmes de la litérature, des i-vecteurs sont utilisés pour augmenter les paramètres avec des informations spécifiques au locuteur. Il a en effet été observé que cela améliorait les performances de modèles entraînés et testés sur de la parole d'adultes. Cependant, que nous entraînions l'extracteur de i-vecteurs sur des données d'adultes ou d'enfants, les performances étaient toujours détériorées par rapport à celles des modèles sans i-vecteurs. Dans le premier cas, les caractéristiques extraites sur la base de la parole d'adulte ne correspondaient pas à la parole d'enfant. Dans le second cas, les informations extraites à partir de la parole d'enfant n'étaient pas pertinentes de par la faible quantité de données et la faible durée moyenne de parole par locuteur (voir tableau 1). Les résultats présentés dans la section 4 sont donc obtenus sans l'utilisation de i-vecteurs.

3. Exemples audio disponibles : https://frama.link/JEP2020-exemples-audio

3.2.2 Modèles chain-TDNN

Les chain-TDNN, appelés TDNN par la suite, ont été implémentés avec l'architecture présentée dans (Peddinti *et al.*, 2015), et en s'appuyant sur la recette Kaldi du corpus Commonvoice [4]. Nous utilisons dans cet article une architecture de modèle chaîne (Povey *et al.*, 2016), qui diffère d'un DNN-HMM classique par l'utilisation d'une fonction de coût au niveau de la séquence plutôt qu'au niveau de la trame. La procédure d'entraînement est similaire à un entraînement par maximisation de l'information mutuelle sans graphe de phones (Veselý *et al.*, 2013).

Une autre caractéristique du modèle chaîne est sa fréquence de trame divisée par 3 à la sortie du réseau : cela accélère le calcul et permet d'effectuer une augmentation de données en appliquant un décalage de trame de 0, 1 et 2 trames.

Les chain-TDNN présentés dans cet article sont similaires à ceux spécifiés dans (Povey *et al.*, 2016). Ils contiennent une première couche affine LDA prenant comme trames d'entrée les indices -2,-1,0,1,2. Suivent ensuite 9 couches cachées avec ReLU, chacune avec 768 unités, parmi lesquelles les couches 2, 4, 6, 7, 8 sont configurées avec des indices de concaténation -1,0,1 -1,0,1 -3,0,3 -3,0,3 -6,-3,0. La couche de sortie comporte 496 unités.

Le HMM utilisé dans notre système hybride TDNN-HMM est un modèle monophone, car nous avons observé de meilleurs résultats qu'avec des modèles triphones. En effet, notre corpus de parole d'enfant étant très réduit, l'entraînement souffre du faible nombre d'occurrence de chaque triphone. De plus, les erreurs produites par les enfants pourraient correspondre à des triphones non représentés dans la langue française, et donc ne pas être reconnues.

3.2.3 Modèles Transfer Learning

Nous avons exploré deux méthodes de TL [5]. La première méthode consiste à supprimer les deux dernières couches du modèle TDNN source et à les remplacer par deux couches initialisées de façon aléatoire. La ré-initialisation doit permettre au réseau de s'adapter rigoureusement aux caractéristiques des enfants. La seconde méthode garde les couches existantes du modèle source, conservant ainsi certaines informations acoustiques de la parole d'adulte. Dans les deux méthodes, les couches finales, ainsi que les autres couches transférées du modèle source sont ré-entraînées avec des facteurs d'apprentissage respectifs de 1 et 0,25. Choisir un facteur d'apprentissage de 0 pour les couches transférées revient à utiliser uniquement les connaissances pré-acquises sur la parole d'adulte. Pour ré-entraîner les modèles TDNN avec des données d'enfants, nous pouvons fournir des alignements soit à partir d'un modèle GMM-HMM, soit à partir d'un modèle TDNN.

3.2.4 Vocal Tract Length Normalisation

Nous appliquons la normalisation VTLN aux MFCC extraits du corpus d'entraînement Commonvoice afin d'étendre la gamme de fréquences des adultes vers une gamme proche de celle des enfants. Les facteurs de déformation qui minimisent le PER obtenus par les modèles GMM-HMM sur l'ensemble Lalilo Test C sont de 1,2 pour les femmes et de 1,3 pour les hommes.

4 Évaluation

Dans cette section, nous testons plusieurs modèles acoustiques enfant (Lalilo), adulte (Commonvoice), TL et TL + VTLN sur le jeu de test C de Lalilo, c'est-à-dire sur des mots qui ont été correctement lus

4. Disponible au lien : https://frama.link/script-TDNN
5. Inspirées par les recettes Kaldi disponibles au lien : https://frama.link/scripts-TL

par des enfants. Les résultats sont affichés dans le tableau 2.

Nous ne visons pas ici à reconstituer et à corriger les mots en fonction des phones détectés, mais à repérer les substitutions, insertions et suppressions de phones chez les enfants qui lisent à voix haute. Par conséquent, nous ne mesurons pas les performances avec un WER, comme le font la plupart des études de RAP, mais avec un taux d'erreurs sur les phones (Phone Error Rate, PER). Le PER est défini comme le ratio entre le nombre d'erreurs (insertions, substitutions et suppressions) et le nombre de phones de la prononciation de référence. Dans cette même optique, nous utilisons un modèle de language unigramme appris sur le corpus d'entraînement de Lalilo pour le décodage.

4.1 Modèles chain-TDNN

Nous validons notre système de reconnaissance de phones en entraînant et testant un modèle TDNN sur le corpus Commonvoice. Les résultats sont affichés dans le tableau 2 : on atteint un PER de 28,4%.

TABLE 2 – Caractéristiques des différents modèles acoustiques et PER (%) obtenus

Nom du modèle	**Méthode de TL**	**Alignements générés par**	**VTLN**	**PER (%)**	
				Commonvoice	**Lalilo Test C**
Commonvoice	–	GMM-HMM	Non	28,4	72,5
Commonvoice + VTLN	–	GMM-HMM	Oui	38,6	69,6
Lalilo	–	GMM-HMM	Non	–	42,3
TL 1	1	GMM-HMM	Non	–	44,0
TL 2A	2	GMM-HMM	Non	–	43,0
TL 2B	2	TDNN	Non	–	39,1
TL 2B + VTLN	2	TDNN	Oui	–	37,1

Le TDNN enfant (Lalilo), atteignant un PER de 42,3%, est 41,4% relatifs plus performant que le TDNN adulte dans la tâche de reconnaissance de phones sur de la parole d'enfant (PER de 72,5%), alors qu'entraîné sur 5 fois moins de données. L'utilisation de la VTLN sur la parole d'adulte permet d'améliorer le PER sur la parole d'enfant, au prix d'une dégradation sur la parole d'adulte.

4.2 Modèles Transfer Learning

Les deux méthodes détaillées en section 3.2.3 sont appelées TL 1 et TL 2 dans le tableau 2. Les alignements utilisés sont générés avec les modèles enfant car ils ont montré de meilleures performances que les modèles adultes. Dans le tableau 2, les noms A et B correspondent respectivement aux alignements générés par GMM-HMM et TDNNF.

Le modèle TL 1 donne de moins bons résultats que le modèle TDNN enfant (Lalilo). La méthode TL 2A donne des résultats légèrement meilleurs que l'approche TL 1 précédente, mais toujours sans amélioration par rapport au modèle Lalilo. La meilleure performance de l'approche TL 2 n'était pas attendue, puisque les auteurs de la méthode ont obtenu de meilleurs résultats avec la méthode TL 1, en utilisant le corpus Wall Street Journal comme source et le corpus Resource Managment (3 heures) comme cible. Cela est dû au fait qu'ils adaptent les tâches de reconnaissance (de la parole radio aux commandes vocales) alors que nous adaptons les domaines de reconnaissance : lors de la ré-initialisation des couches proches de la sortie, ils ne perdent que les informations liées à la tâche, qu'ils peuvent remplacer grâce aux données cible. Dans notre cas, nous perdons de précieuses

informations acoustiques et de prononciation qui ne peuvent pas être retrouvées avec uniquement quatre heures de parole d'enfant.

Enfin, la méthode TL 2B, qui utilise des alignements générés par un TDNN enfant, apporte une amélioration substantielle, avec un PER de 39,1 % pour le modèle TL 2B, ce qui correspond à une amélioration relative de 7,6 % par rapport au TDNN enfant.

4.3 Impact de la VTLN

La VTLN apportant une amélioration sur le TDNN Commonvoice lorsque testé sur parole d'enfant, une tendance identique est attendue pour le modèle TL 2B +VTLN, obtenu par transfer learning avec le modèle TDNN Commonvoice + VTLN. Le PER diminue, de 39,1% pour le modèle TL 2B, à 37,1% pour le modèle TL 2B VTLN, ce qui correspond à une amélioration relative de 5,1%.

5 Analyses & discussion

5.1 Analyse d'erreurs en fonction du voisement

Le taux d'erreur de reconnaissance (TER) est défini dans l'équation (1), avec C, S, D référant respectivement aux nombres de détections correctes, substitutions et suppressions. Il mesure la capacité du système à reconnaître correctement un phone donné, et donc ne prend pas en compte les insertions comme le PER.

$$TER = \frac{S + D}{C + S + D} \tag{1}$$

La figure 1 montre la capacité des différents modèles à reconnaître les phones voisés et non-voisés.

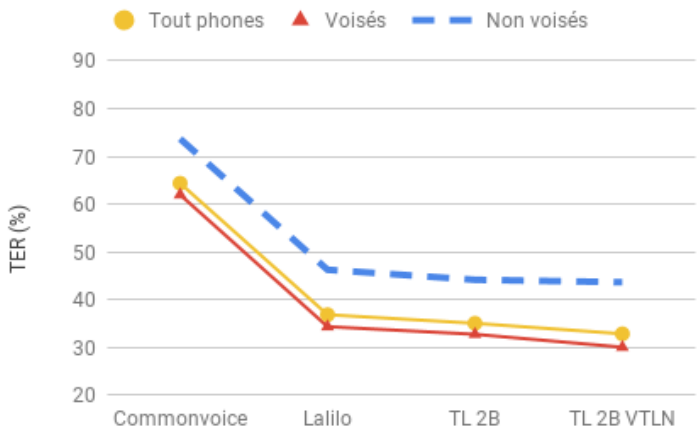


FIGURE 1 – Évolution du TER (%) sur les phones voisés et non-voisés du jeu de test C (Lalilo), pour les modèles adulte, enfant, TL 2B et TL 2B VTLN

La première observation est que les phones voisés sont, en moyenne, 24,6 % relativement mieux reconnus que les phones non-voisés. Les phones voisés, qui correspondent en français à 25 phones sur 31, représentent 79 % de notre corpus de test C, ce qui explique pourquoi la courbe "Tout phones" suit de si près celle des phones voisés. Le modèle enfant (Lalilo) réduit considérablement le TER par rapport au modèle adulte. De même, le modèle TL 2B améliore la reconnaissance des phones voisés et non voisés, avec des améliorations relatives respectives du TER de 4,7 % et 4,5 % par rapport au modèle enfant. Le dernier point, correspondant au modèle TL 2B VTLN, montre que la VTLN a une forte influence sur le TER pour les phones voisés (amélioration relative de 8,2%), mais n'apporte pas d'amélioration significative pour les phones non-voisés. Les phones non-voisés sont

en effet articulés sans aucune vibration des cordes vocales, ce qui signifie qu'ils ne possèdent pas de fréquence fondamentale ni d'harmoniques. Comme la VTLN agit sur les fréquences, il affecte les phones voisés qui sont caractérisés par leurs formants, mais pas les phones non-voisés. La très légère amélioration pourrait être due à la présence habituelle de pics de fréquence dans les consonnes, correspondant à la position des résonateurs, qui pourraient être modérément affectés par la VTLN.

5.2 Influence de la parole de lecteurs non-experts

Pour quantifier la difficulté apportée par des lecteurs non-experts par rapport à des lecteurs plus expérimentés dans le cadre d'un système de RAP, nous avons calculé le PER obtenu par le modèle monophone TL 2B VTLN sur les jeux de test C (lecture correcte) et I (lecture incorrecte) du corpus Lalilo, et obtenu les résultats affichés dans le tableau 3. Nous pouvons constater une détérioration relative drastique de 45,0% du PER pour le Test I, relativement au Test C.

TABLE 3 – PER (%) pour le modèle monophone TL 2B VTLN

Corpus	Test C	Test I (tous)	Test I (Fluence)	Test I (Déchiffrage)
PER (%)	37,1	53,8	51,1	56,9

Les catégories d'erreurs de lecture (Fluence et Déchiffrage) sont décrites en section 3.1. Nous observons que la catégorie Fluence obtient un PER relativement 10% meilleur que la catégorie Déchiffrage. En effet, les erreurs de déchiffrage prévalant sur les erreurs de fluence lors du classement des enregistrements, certains mots de la catégorie Déchiffrage contiennent les deux sortes d'erreurs, combinant ainsi les difficultés. Ces deux catégories correspondent à la réalité de notre application, et les résultats présentés démontrent la difficulté à reconnaître de la parole d'enfant apprenant lecteur.

6 Conclusion & perspectives

Dans le cadre d'une tâche de détection des erreurs pour l'évaluation de la lecture à haute voix chez les enfants, la précision du système de reconnaissance vocale a une grande incidence sur la pertinence des paramètres qui sont transmis à un classifieur. Dans cet article, nous améliorons la précision de la reconnaissance de phones sur parole d'enfant au moyen d'une méthode de *Transfer Learning*, où des modèles TDNN adultes sont adaptés avec quatre heures de parole d'enfant. Nous obtenons un PER de 39,1 %, ce qui correspond à un gain relatif de 7,6 % par rapport à un TDNN entraîné uniquement sur parole d'enfant. L'application de la VTLN sur le corpus d'entraînement de parole d'adulte pour l'extension de la gamme de fréquences de locuteurs, et l'utilisation de ce modèle adulte adapté au VTLN comme modèle source pour le transfer learning réduit le PER à 37,1%, ce qui apporte une réduction relative supplémentaire de 5,1%.

Nous utilisons actuellement une structure de TDNN factorisés qui ont montré de meilleurs résultats sur un petit corpus de parole d'enfant (Wu *et al.*, 2019). Nous pourrions également renforcer notre utilisation de la VTLN en utilisant un DNN pour adapter les facteurs de déformation spécifiquement à chaque locuteur, comme dans (Serizel & Giuliani, 2014b). De futurs travaux porteront sur la pertinence des paramètres MFCCs pour la parole d'enfant, avec l'étude d'une échelle Mel adaptée aux fréquences des enfants, et la recherche de paramètres spécifiques à nos données d'enfants. Enfin, une modélisation linguistique appropriée pourrait combler l'écart de performance entre les mots correctement lus et ceux contenant des erreurs de fluence ou de déchiffrage.

Références

BOLAÑOS D., COLE R., WARD W., BORTS E. & SVIRSKY E. (2011). FLORA : Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process.*, **7**(4), 16.

FRINGI E., LEHMAN J. F. & RUSSELL M. J. (2015). Evidence of phonological processes in automatic recognition of children's speech. In *INTERSPEECH*.

GODDE E., BAILLY G., ESCUDERO D., BOSSE M.-L. & ESTELLE G. (2017). Evaluation of reading performance of primary school children : Objective measurements vs. subjective ratings. p. 23–27.

LEE S., POTAMIANOS A. & NARAYANAN S. S. Y. (1999). Acoustics of children's speech : developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, **105**(3), 1455–1468.

METALLINOU A. & CHENG J. (2014). Using deep neural networks to improve proficiency assessment for children english language learners. In *INTERSPEECH*.

MOSTOW J. & AIST G. (2001). Evaluating tutors that listen : An overview of project listen.

MUGITANI R. & HIROYA S. (2012). Development of vocal tract and acoustic features in children. *The Journal of the Acoustical Society of Japan*, **68**(5), 234–240.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *ASRU*.

POVEY D., PEDDINTI V., GALVEZ D., GHAHREMANI P., MANOHAR V., NA X., WANG Y. & KHUDANPUR S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *INTERSPEECH*, p. 2751–2755.

PROENÇA J. D. L. (2018). *Automatic Assessment of Reading Ability of Children*. Thèse de doctorat.

SERIZEL R. & GIULIANI D. (2014a). Deep neural network adaptation for children's and adults' speech recognition. In *CLiC-it*.

SERIZEL R. & GIULIANI D. (2014b). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *SLT*, p. 135–140.

SHI L., BAO F., WANG Y. & GAO G. (2018). Research on transfer learning for Khalkha Mongolian speech recognition based on TDNN. In *IALP*, p. 85–89.

SHIVAKUMAR P. G. & GEORGIOU P. G. (2018). Transfer learning from adult to children for speech recognition : Evaluation, analysis and recommendations. *ArXiv*.

VESELÝ K., GHOSHAL A., BURGET L. & POVEY D. (2013). Sequence-discriminative training of deep neural networks. In *INTERSPEECH*.

WAIBEL A., HANAZAWA T., HINTON G., SHIKANO K. & LANG K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(3), 328–339.

WU F., GARCÍA-PERERA L. P., POVEY D. & KHUDANPUR S. (2019). Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network. In *INTERSPEECH*, p. 1–5.

YONG B. & TING H. N. (2011). Speaker-independent vowel recognition for malay children using time-delay neural network. *IFMBE*, **35**.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Informations segmentales pour la caractérisation phonétique du locuteur : variabilité inter- et intra-locuteurs

Cédric Gendrot[1] Emmanuel Ferragne[1] Thomas Pellegrini[2]
(1) LPP, UMR 7018 CNRS-Sorbonne Nouvelle, 19 rue des Bernardins, 75005 Paris, France
(2) IRIT, UMR 5505 CNRS-INP, UT1, UT3, UT2J, 118 Route de Narbonne, F-31062 Toulouse Cedex 9
cedric.gendrot@cnrs.fr, emmanuel.ferragne@u-paris.fr,
thomas.pellegrini@irit.fr

RÉSUMÉ

Nous avons effectué une classification automatique de 44 locuteurs à partir de réseaux de neurones convolutifs (CNN) sur la base de spectrogrammes à bandes larges calculés sur des séquences de 2 secondes extraites d'un corpus de parole spontanée (NCCFr). Après obtention d'un taux de classification moyen de 93,7 %, les différentes classes phonémiques composant chaque séquence ont été masquées afin de tester leur impact sur le modèle. Les résultats montrent que les voyelles orales influent avant toute autre classe sur le taux de classification, suivies ensuite par les occlusives orales. Ces résultats sont expliqués principalement par la représentation temporelle prédominante des voyelles orales. Une variabilité inter-locuteurs se manifeste par l'existence de locuteurs attracteurs qui attirent un grand nombre de faux positifs et qui ne sont pas sensibles au masquage effectué. Nous mettons en avant dans la discussion des réalisations acoustiques qui pourraient expliquer les spécificités de ces locuteurs.

ABSTRACT

An automatic classification task involving 44 speakers was performed using convolutional neural networks (CNN) on broadband spectrograms extracted from 2-second sequences of a spontaneous speech corpus (NCCFr). We obtained a mean classification rate of 93,7 % and carried out an occlusion experiment afterwards : different phonemic classes were hidden within each sequence so as to test their impact on classification rates. Results show that oral vowels influence the classification much more than the other classes. These results are mainly explained by the prevailing temporal representation of oral vowels. Substantial inter-speaker variability is observed and correlated with the presence of magnet speakers who 'attract' most false positive classifications from other speakers. In the discussion, we display acoustic measurements that may be relevant to explain these speakers' behaviour.

MOTS-CLÉS : caractérisation du locuteur, deep learning, segmental.

KEYWORDS: speaker characterization, deep learning, segmental.

1 Introduction

Il est fréquent de constater que des locuteurs de notre entourage présentent une production particulière de certains phonèmes : on pense par exemple à la réalisation de /ʃ-s/ et /ʒ-z/ en fricatives latérales

Volume 1 : Journées d'Études sur la Parole, pages 262–270.
HAL : hal-02798547.

[ɬ] et [ɮ]. Il est donc légitime de se poser la question de l'utilité potentielle de la prononciation spécifique de certains phonèmes dans la caractérisation du locuteur sur la base d'indices acoustiques. La caractérisation du locuteur peut avoir plusieurs objectifs : elle peut se faire dans une optique criminalistique pour la comparaison de voix (Morrison & Thompson, 2017; Ajili *et al.*, 2018). Elle peut également servir la recherche fondamentale en phonétique, qui vise à mieux comprendre la variabilité que l'on observe pour la description de phénomènes tels que la coarticulation, la réalisation des groupes prosodiques, etc. (Keating *et al.*, 2017, par exemple).

Notre travail se concentre sur l'influence du segmental pour la caractérisation du locuteur : certains phonèmes ou certaines classes de phonèmes sont-ils plus pertinents que d'autres ? Et est-ce que ces éventuels phonèmes discriminants sont partagés par l'ensemble des locuteurs ? Si nous nous focalisons sur le segmental, il ne faut pas oublier non plus que les informations prosodiques (valeurs de f0, d'intensité, etc.) peuvent également permettre de caractériser certains locuteurs ; citons notamment Dellwo *et al.* (2015) et Keating *et al.* (2017). Plusieurs études ont démontré que les caractéristiques segmentales des locuteurs peuvent être pertinentes pour la classification de ces derniers (Kahn, 2014; Amino *et al.*, 2006). Dans ce cadre, il est fréquent de trouver dans la littérature que les voyelles nasales ont un poids plus important dû à l'ajout de la cavité nasale dans la production (Shriberg & Stolcke, 2008). Mais Ajili *et al.* (2018), dans une étude plus récente, ont montré que les voyelles orales étaient les plus utiles dans une tâche de reconnaissance du locuteur pour une approche criminalistique, ces résultats contradictoires pouvant être interprétés comme une forte variabilité inhérente à la parole. Les auteurs notent également une forte variabilité inter-locuteurs, qu'il est nécessaire d'approfondir selon eux, et nous tentons ici d'analyser cette variabilité.

Dans une expérience s'appuyant sur l'apprentissage automatique comme pour la présente étude, il existe différents procédés pour déterminer quelles sont les informations mises à profit par le modèle afin de classifier les locuteurs (Ferragne *et al.*, 2019). Sur le même principe que Ajili *et al.* (2018), nous avons réalisé une expérience de classification avec masquage (occlusion), où une partie de l'information contenue dans le signal acoustique est cachée afin de comparer la classification avant et après masquage. Cette étude se rapproche également des travaux effectuées par Besacier & Bonastre (1998) dans lesquels des blocs temporels de signal sont sélectionnés pour améliorer les taux d'identification du locuteur. Cependant, contrairement aux études citées ci-dessus, nous utilisons des spectrogrammes à bandes larges en entrée, car notre objectif de phonéticien sera à terme de retracer la correspondance acoustique - articulatoire. Après une présentation de la méthode employée, et des résultats de l'influence du masquage, nous nous concentrerons sur la variabilité observée, en insistant notamment sur quelques locuteurs caractéristiques.

2 Méthode

Nous avons dans un premier temps mené une tâche de classification de segments de paroles en 44 classes de locuteurs en utilisant un réseau de neurones convolutif (CNN) qui prenait en entrée des spectrogrammes. Nous avons ensuite procédé à une occlusion partielle des spectrogrammes, d'abord par phonème (e.g. tous les phonèmes étaient remplacés tour à tour par un masque noir), puis par classe de phonèmes (e.g. toutes les voyelles orales étaient masquées simultanément) afin d'observer la dégradation du taux de classification conséquente au masquage de l'information.

2.1 Corpus et prétraitement

Des séquences de 2 secondes ont été utilisées pour la classification ; elles ont été extraites de 44 locuteurs du corpus NCCFr (Torreira *et al.*, 2010). Ce corpus est constitué de conversations spontanées d'environ une heure entre deux (voire trois) amis ; il a été annoté par des transcripteurs professionnels, puis aligné phonémiquement par le système du LIMSI. Le corpus est composé d'enregistrements de 23 hommes et 21 femmes ; les séquences contenant un minimum de 18 et un maximum de 43 phonèmes ont été retenues, sans autre type de contrainte. Ces séquences, échantillonnées à 16 kHz, ont été converties en spectrogrammes à bandes larges avec des trames de 5 ms, un chevauchement de 90 % et une taille de FFT de 512 points. La dynamique a été fixée à 70 dB et quantifiée sur 8 bits de niveaux de gris dans les images finales. La résolution en fréquence, 257 points pour 8 kHz, a été laissée telle quelle dans les images fournies en entrée du modèle. En revanche, nous avons réduit la résolution temporelle des spectrogrammes (de 3 991 à 400 points) pour des raisons évidentes de mémoire. Nous disposons donc de 15 400 images de spectrogrammes : 350 pour chacun des 44 locuteurs.

2.2 Modèle initial

Un réseau de neurones profond de type ResNet-18 (He *et al.*, 2016) a été utilisé pour la classification automatique des spectrogrammes en 44 classes de locuteurs. L'ensemble d'apprentissage contenait 70 % des données ; 10 % servaient pour la validation et 20 % pour l'évaluation. Nous avons utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec une valeur initiale du taux d'apprentissage de $1e$-4, divisé par deux après huit itérations complètes sur les données d'apprentissage. Un maximum de 10 itérations en tout a été effectué avec des mini-lots (*mini-batchs*) de 32 exemples, ce qui fait que le modèle a convergé en 28 minutes sur une carte GPU NVIDIA GTX 1080.

2.3 Protocole d'occlusion

Dans un premier temps, afin de quantifier la pertinence des phonèmes pour la caractérisation du locuteur, nous avons effectué le masquage phonème par phonème tout au long de chaque séquence de 2 secondes. L'objectif était d'identifier un ou plusieurs phonèmes susceptibles de faire basculer l'identification du locuteur (i.e. engendrer une classification erronée). Ce type de changement dans la classification n'a été obtenu que pour les séquences dont la probabilité de classification dans la classe correcte avant masquage était faible, inférieure à 50 %. Les taux d'identification étant supérieurs à 90 % avec des probabilités d'identification très élevées, le masquage d'un seul phonème pouvait suffire que très rarement à engendrer une erreur de classification. Au total, à l'issue du masquage par phonème, seules 2.5 % des séquences présentaient un changement de classe de locuteur, ce qui est insuffisant pour effectuer une analyse quantitative. Notons tout de même que les phonèmes /s/ et /ʃ/ ont été identifiés pour deux locuteurs comme particulièrement pertinents, notamment parce que -après écoute des séquences concernées- ceux-ci réalisés avec un chuintement. Pour quelques autres locuteurs, le masquage d'une hésitation longue (>300 ms) pouvait faire basculer la classification en locuteurs sur la séquence testée. Mais ces cas étaient par trop rares et nous avons donc procédé dans un second temps à une occlusion par classes phonémiques, où tous les phones correspondant à une classe phonémique ont été masqués simultanément. Cette occlusion représente ainsi une durée plus importante (293 ms en moyenne, toutes classes confondues) et nous espérions voir augmenter le

nombre de changements d'identifications du locuteur après occlusion.

Les classes phonémiques ont été regroupées de la façon suivante : voyelles orales (ORAVO), consonnes/voyelles nasales (NASAL), occlusives (OCCLU), fricatives (FRICA) et sonantes (SONOR). Les consonnes et voyelles nasales ont été regroupées puisque la nasalité est estimée comme un critère prépondérant mais également afin d'obtenir des groupes plus équilibrés en termes de fréquences pour chaque catégorie.

Dans l'étude de Ajili *et al.* (2018), un masque est également appliqué de façon aléatoire sur un échantillon de signal de durée équivalente à celle correspondant aux phonèmes de la classe phonémique masquée. Nous avons choisi de ne pas procéder de cette façon mais plutôt de prendre en compte la variation de durée a posteriori dans nos analyses statistiques. Les résultats que nous présenterons dans les sections suivantes seront basés sur les séquences dont la classification du locuteur passe de correcte à incorrecte (3 821 sur 15 029).

3 Résultats

Les résultats présentés ci-dessous ont été calculés sur les séquences d'évaluation exclusivement. Le taux moyen des bonnes classifications avant masquage est de 93.7 % (14100/15029). Après masquage, il passe à 68.3 % (10279 / 15029) : pour les occlusives 78.0 % (2236/2867), les fricatives 83.6 % (2347/2805), les nasales 83.9 % (2291/2729), les voyelles orales 36.2 % (1046/2887) et les sonantes 83.9 % (2359/2812). Ces confusions vont vers un autre locuteur du même sexe dans 96.6 % des cas pour les femmes et seulement 61.5 % des cas pour les hommes.

TABLE 1 – Résumé des résultats de classification (taux de bonne classification en %, score de probabilité de classification dans la classe correcte en %, taille du masque en ms) pour les séquences sans masquage puis en fonction chaque classe phonémique masquée.

	séquence sans masque	ORAVO	OCCLU	NASAL	FRICA	SONOR
classif. correct.	93.8	36.2	78.0	83.9	83.6	83.9
probabilité	84.9	27.7	60.8	65.4	66.4	67.9
durée masque	0	490	304	223	254	190

En analysant les scores d'identification après masquage, nous déduisons que les voyelles orales ont un effet important sur la classification, loin devant les occlusives puis les autres catégories. Notons également que le niveau de probabilité indiqué par le réseau lors de la classification, qui donne un indice de la certitude du résultat, passe de 84.9 % avant masquage à 57 % en moyenne (et 28 % pour les voyelles orales). Pour la suite de cette étude, nous avons cependant décidé de nous concentrer sur les cas de changements de classification (de correcte à erronée). Ce résultat est surprenant au premier abord puisque la classes des nasales n'est que peu pertinente dans la caractérisation du locuteur ici. Cela pourrait être en partie expliqué par le fait que nous avons combiné voyelles et consonnes nasales dans la même classe (pour des raisons d'homogénéité des classes). En effet, les voyelles nasales sont plus particulièrement mentionnées comme pertinentes dans la littérature, et les consonnes nasales dans une moindre mesure seulement. Nous avons comptabilisé le nombre de voyelles nasales présentes dans chaque séquence afin d'estimer si un nombre plus important de

voyelles nasales pouvait être corrélé à des changements de classifications plus fréquents. Les résultats montrent que lorsque le masquage de la classe des nasales implique 2 voyelles nasales et plus (1.78 en moyenne dans l'ensemble des séquences), le taux de classifications correctes chute à 79 %, contre 90 % quand il y a moins de 2 voyelles nasales dans les séquences. Ce résultat corrobore l'idée que les voyelles nasales sont plus pertinentes que les consonnes nasales pour la caractérisation du locuteur.

Il est nécessaire de pondérer ces résultats en considérant les fréquences d'apparition inhérentes à chaque classe, la classe des voyelles orales étant par exemple beaucoup plus fréquente que les autres classes. La durée masquée moyenne – en effectuant le calcul pour toutes les occurrences – est de 300 ms, celle des voyelles orales monte à 490 ms. Les occlusives ont une durée masquée (304 ms) supérieure aux nasales (223 ms), puis pour les plus courtes les fricatives (253 ms) et les sonantes (190 ms).

Pour ce faire, nous avons utilisé un modèle mixte linéaire généralisé (GLMM) afin d'évaluer la probabilité d'un changement de classe en fonction de la classe phonémique masquée et de sa durée. Les variables aléatoires utilisées dans ce modèle sont les locuteurs et les différentes séquences testées. Le modèle indique un effet significatif avec une valeur F de 391.4 ($p=0.0013$) pour la classe phonémique et 1284.6 ($p<0.0001$) pour la durée, l'interaction entre la classe et la durée est quant à elle non significative avec une valeur F de 3.9 ($p=0.11$). Ces résultats indiquent que l'importance de la classe phonémique sur la caractérisation du locuteur est bien pertinente, parallèlement à l'influence de la durée occultée au sein de la séquence.

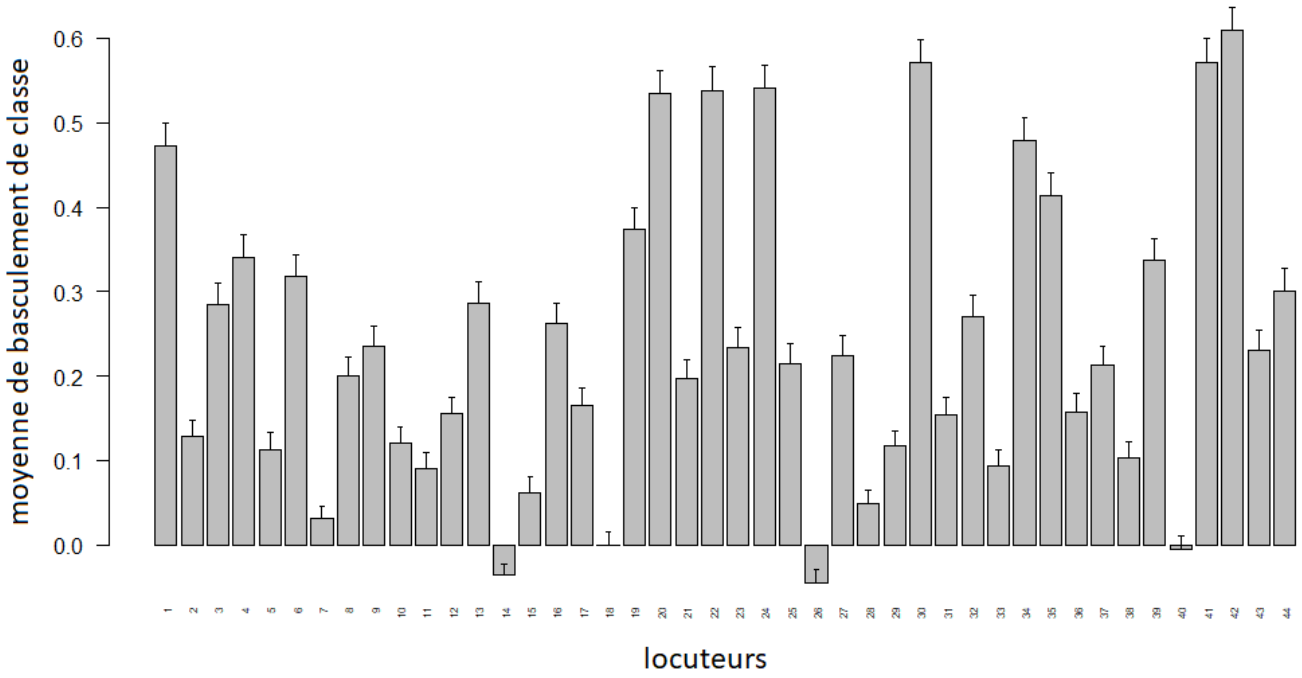


FIGURE 1 – Sensibilité des locuteurs au masquage : en abscisse les identifiants des locuteurs et en ordonnée la moyenne des changements de classification

4 Variation inter-locuteurs

Nous observons que pour environ 20 % des locuteurs, le masquage n'a que peu voire pas d'effet sur le résultat de la classification. Pour quantifier ce phénomène, nous avons calculé la moyenne des occurrences où un changement de classe a lieu (0 si l'identification du locuteur est passée de correcte à incorrecte, 1 si l'identification est restée correcte, et -1 si l'identification passe d'incorrecte à correcte). Comme illustré sur la Figure 1, les locuteurs 14, 18, 26 et 40 ont une moyenne nulle à négative, les locuteurs 7, 15 et 28 ont une moyenne inférieure à 0.05, tandis que les locuteurs 11 et 33 ont une moyenne située entre 0.05 et 0.1. Ces 9 locuteurs ont un score moyen de bonne classification à 92.0 %, contre 94.1 % pour les 35 autres locuteurs, et 93,7 % pour la moyenne, avec des taux de probabilité de 84.9 % identiques à ceux des 35 autres locuteurs. Ces résultats montrent que les locuteurs insensibles au masquage ne sont pas des locuteurs plus difficiles (ou faciles) à classifier, et que la source de cette insensibilité au masquage doit être cherchée ailleurs. En considérant les différentes classes phonémiques dans ce résultat sur la Figure 2, on observe qu'après masquage, le taux moyen des bonnes classifications descend seulement à 88 % lorsque la classe des voyelles orales est masquée et ne bouge pas pour les autres classes phonémiques, ce qui conforte l'insensibilité de ces locuteurs au masquage pour toutes les classes phonémiques, les différences observées entre celles-ci étant considérablement réduites (Figure 3).

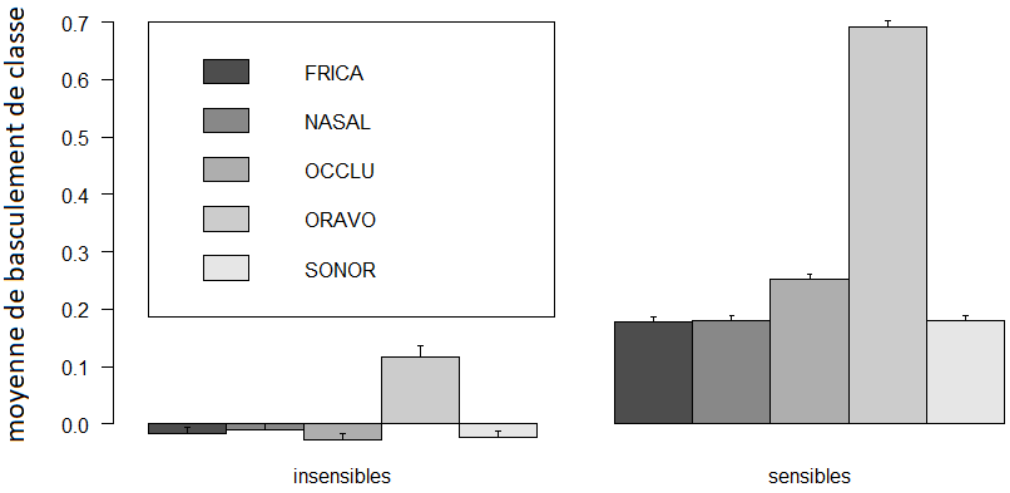


FIGURE 2 – Sensibilité au masquage en fonction de la classe phonémique pour les 9 locuteurs peu sensibles au masquage comparés aux 35 autres locuteurs

Nous tentons ici d'approfondir la compréhension de cette variabilité inter-locuteurs en observant la matrice de confusion. Les lignes de la matrice nous renseignent sur les vrais positifs et faux positifs obtenus pour les 44 locuteurs, et les 9 locuteurs mentionnés plus haut reçoivent un nombre important de faux positifs de la part des autres locuteurs. La non sensibilité au masquage d'un locuteur serait donc liée au nombre de faux positifs reçus, et nous donnerons à ces locuteurs le terme de locuteurs attracteurs.

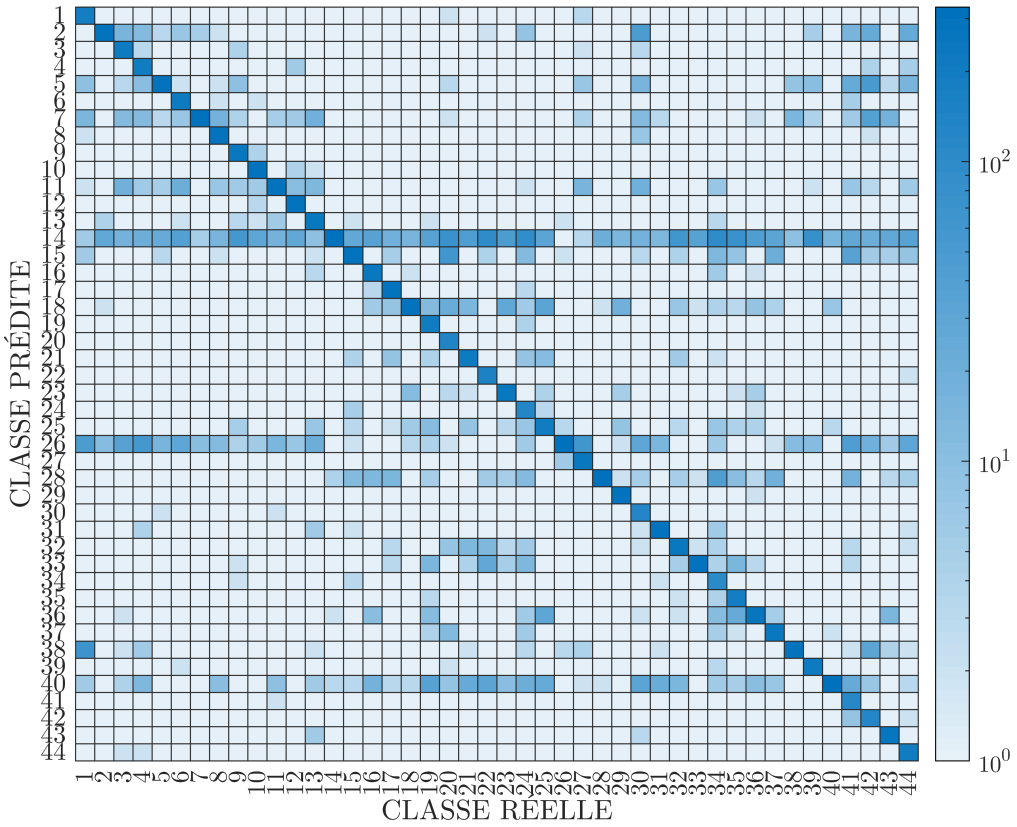


FIGURE 3 – Matrice de confusion de la classification des locuteurs après masquage

Afin de valider ce résultat, nous avons calculé une corrélation de Spearman entre le nombre de faux positifs que chaque locuteur a reçu et sa sensibilité au masquage (calculée comme la moyenne du nombre de changements de classes de correct à incorrect). Le coefficient obtenu est de -0.67 ($t = -5.8966; p = 5.594e - 07$) ce qui montre un lien très net entre ces deux variables. Ce taux est plus élevé si l'on considère uniquement la classe phonémique des occlusives (-0.8), des nasales (-0.8), des fricatives (-0.77) et des sonantes (-0.79), mais il est moins élevé que la moyenne pour les voyelles orales (-0.53). Ces résultats confirment un statut particulier de la classe des voyelles orales, qui lorsqu'on les masque, modifie considérablement la classification des locuteurs, mais les erreurs de classification sont moins focalisées sur certains locuteurs que pour les autres classes. Enfin, si l'on effectue une corrélation entre la sensibilité au masquage de chaque locuteur et le nombre de faux positifs qu'il a reçus avant la procédure de masquage, celle-ci tombe à 0.088 ($t = -0.57278; p = 0.5698$). Cette dernière corrélation montre que le masquage permet de mettre en avant ces locuteurs attracteurs qui n'apparaissent pas lorsque le spectrogramme entier est utilisé (avant la procédure de masquage). Les locuteurs qui au contraire sont très sensibles à l'occlusion ne reçoivent qu'un nombre très limité de faux positifs quelle que soit la classe phonémique masquée.

5 Discussion et conclusion

Notre étude a montré que lorsque le masquage est effectué phonème par phonème, certains segments très spécifiques s'avèrent être pertinents pour la caractérisation du locuteur.Il ne s'agit cependant en général que de phonèmes très spécifiques avec des prononciations atypiques tels que /s/ et /ʃ/ ou des

hésitations.

Lorsqu'on procède au masquage par classes de phonèmes, les résultats montrent que les voyelles orales, notamment du fait de leur durée plus importante, jouent un rôle dans la bonne classification des locuteurs puisque leur absence détériore considérablement les résultats. Mais environ 20 % des locuteurs ne sont pas sensibles au masquage, ces locuteurs attirant à eux les prédictions dont le score de probabilité est plus faible. Ces locuteurs que nous avons qualifiés d'attracteurs pourraient être considérés comme les agneaux (lambs) selon la terminologie de Doddington et al. George Doddington & Reynolds (1998) car ces locuteurs pourraient apparaitre comme faciles à imiter. Afin de comprendre pourquoi ces locuteurs recueillent un nombre important de faux positifs, nous avons effectué des mesures acoustiques sur les différentes séquences testées de ces locuteurs et avons pu constater qu'ils étaient caractérisés par une variation acoustique plus importante que les autres locuteurs, notamment pour leurs valeurs de f0 et d'intensité. Nous avons également pu faire ressortir des locuteurs qui se distinguent par leur caractère moyen sur l'ensemble des mesures acoustiques, plutôt qu'extrêmes sur une seule. Une classification de ces 44 mêmes locuteurs a été réalisée sur la base d'indices prosodiques seuls par Chignoli *et al.* (2020) (soumis à cette conférence) et montre que les informations de f0 et d'intensité peuvent être complémentaires au spectrogramme dans près d'un tiers des classifications des locuteurs.

Il est à noter que lorsqu'une classe phonémique permet de faire basculer la classification du locuteur de correcte à erronée, il est très fréquent que les autres classes phonémiques testées fassent également basculer la classification (25 % de cas où une seule classe phonémique est impliquée dans un changement de catégorie pour une séquence, 24 % de cas où il y a 2 classes, 51 % de cas où il y a entre 3 et 5 classes) Ce résultat indique qu'au delà de la pertinence de la classe phonémique pour la classification du locuteur, c'est la séquence dans son ensemble qui joue un rôle dans le résultat de la classification. Nous avons remarqué après écoute des séquences mal identifiées que celles-ci contenaient des rires, chuchotements, éclats de voix, etc., ce qui rend plus complexe la classification par le réseau dans une expérience de masquage qui a pour effet de faire baisser la probabilité d'appartenance à la classe (certitude du résultat), et donc d'atteindre plus facilement un seuil critique d'identification.

Pour la suite de cette étude, nous envisageons de prendre en compte le vecteur de probabilité d'appartenance à chaque classe afin d'obtenir une analyse plus fine de l'évaluation du masquage en calculant par exemple un classement entre locuteurs. Il est également envisagé de travailler sur la base d'une tâche de discrimination (et non plus de classification) dans un ensemble ouvert de locuteurs afin de pouvoir généraliser les conclusions proposées dans ce travail.

Références

AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & KAHN J. (2018). Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique. In *Proc. XXXIIe Journées d'Études sur la Parole*, p. 28–36. DOI : 10.21437/JEP.2018-4.

AMINO K., SUGAWARA T. & ARAI T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Sciences and Technology*, **27**, 233–235.

BESACIER L. & BONASTRE J.-F. (1998). Time and frequency pruning for speaker identification. In *Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*.

CHIGNOLI G., GENDROT C. & FERRAGNE E. (2020). Caractérisation du locuteur par CNN à l'aide des contours d'intensité et d'intonation : comparaison avec le spectrogramme. In *soumis aux Journées d'Etude de la Parole 2020*.

DELLWO V., LEEMANN A. & KOLLY M.-J. (2015). Rhythmic variability between speakers : Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, **137**(3), 1513–1528.

FERRAGNE E., GENDROT C. & PELLEGRINI T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In *Proc. ICPhS*.

GEORGE DODDINGTON, WALTER LIGGETT A. M. M. P. & REYNOLDS D. (1998). Sheep, goats, lambs and wolves : a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *NIST 1998 Speaker Recognition Evaluation*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778, Las Vegas, NV, USA : IEEE. DOI : 10.1109/CVPR.2016.90.

KAHN J. (2014). *Parole de locuteur : performance et confiance en identification biométrique vocale*. Thèse de doctorat, Avignon.

KEATING P., KREIMAN J. & VASSELINOVA N. (2017). Acoustic similarities among voices. part 2 : Male speakers. *The Journal of Acoustic Society of America*, **142**.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

MORRISON G. S. & THOMPSON, C. W. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, **18**, 326–434.

SHRIBERG E. & STOLCKE A. (2008). The case for automatic Higher-Level features in forensic speaker recognition. In *Proc. International Conference on Speech Communication and Technology (Interspeech)*, p. 1509–1512.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, **52**(3), 201–212. DOI : 10.1016/j.specom.2009.10.004.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Evaluation de l'intelligibilité de patients avec traitement du cancer des cavités orales et pharyngales

Alain Ghio[1], Muriel Lalain[1], Marie Rebourg[1],
Corinne Fredouille[2], Virginie Woisard[3]

(1) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Laboratoire d'Informatique d'Avignon, Avignon, France
(3) Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France

alain.ghio@lpl-aix.fr

RÉSUMÉ

La perte d'intelligibilité représente une plainte importante des patients atteints de troubles de la parole. Plusieurs batteries de test d'intelligibilité existent mais leurs limitations résident dans la capacité des auditeurs à restaurer les séquences distordues. Nous proposons un nouveau test fondé sur l'utilisation de pseudo-mots en grande quantité afin de complètement neutraliser les effets perceptifs indésirables. Nous avons appliqué ce test à une population de 39 sujets sains et 78 patients post traitement de cancers de la cavité buccale et de l'oropharynx. Chaque locuteur a produit 52 pseudo-mots tirés aléatoirement. 40 auditeurs ont retranscrit ces productions. Les transcriptions orthographiques ont été phonétisées et comparées aux formes phonétiques attendues. Un algorithme fournit un score de déviation phonologique perçue (PPD) fondée sur le nombre de traits différents entre la forme attendue et celle transcrite. Les résultats montrent qu'il existe un seuil PPD de 0.6 traits/phonème au-dessus duquel, la parole produite est dysfonctionnelle. De plus, le score de PPD est bien corrélé au jugement subjectif de la sévérité obtenue auprès d'experts. Ce test semble donc efficace pour mesurer la performance articulatoire des locuteurs.

ABSTRACT

Intelligibility Assessment of patients in the Context of Head and Neck Cancers

The loss of intelligibility is a major complaint from patients with speech impairments. Several intelligibility test batteries exist, but their limitations lie in the ability of listeners to restore distorted sequences. We propose a new test based on the use of pseudo-words in large quantities in order to completely neutralize the unwanted perceptual effects. We applied this test to a population of 39 healthy subjects and 78 post-treatment patients for cancers of the oral cavity and pharynx. Each speaker produced 52 pseudo-words drawn at random. 40 listeners transcribed these productions. The orthographic transcriptions were phoneticized and compared to the expected phonetic forms. An algorithm provides a perceived phonological deviation score (PPD) based on the number of different features between the expected form and the transcribed one. The results show that there is a PPD threshold of 0.6 features / phoneme above which the speech produced is dysfunctional. In addition, the PPD score is well correlated with the subjective judgment of severity obtained from experts. This test therefore seems effective in measuring the articulatory performance of speakers..

MOTS-CLÉS : phonétique clinique, intelligibilité, troubles de la parole, cancer des VADS
KEYWORDS: clinical phonetic, intelligibility, speech disorders, head and neck cancer

Volume 1 : Journées d'Études sur la Parole, pages 271–279.
HAL : hal-02798549.

1 Une mesure d'intelligibilité privilégiant l'information acoustique

La perception de la parole est une intégration complexe d'informations provenant du signal de parole et d'informations de haut niveau détenues par l'auditeur. L'information extraite du signal de parole est traitée de façon ascendante et correspondant au décodage acoustico-phonétique. L'information de haut niveau est manipulée de manière descendante et correspondant à l'exploitation des connaissances linguistiques, encyclopédiques, situation de communication, contexte psychosocial... Lindblom (1990) définit comme « information dépendante du signal » le résultat de processus ascendant tandis que « l'information indépendante du signal » est liée aux processus descendants. Dans ce cadre, Keintz et al. (2007) définissent l'intelligibilité comme « la quantité de parole comprise à partir du seul signal acoustique ». Inversement, Fontan et al. (2015) définissent la compréhension comme « l'intégration à la fois des informations acoustico-phonétiques et de toutes les informations pertinentes indépendantes du signal afin de comprendre un message parlé dans une situation de communication particulière ». Nous adoptons clairement la définition de Keintz et dans ce cadre, le terme qui devrait être utilisé pour l'intelligibilité est la reconnaissance de la parole et non la compréhension du message. Dans notre définition de l'intelligibilité, la question est la suivante : les auditeurs reconnaissent-ils les sons de la parole qui sont prononcés ? Pour illustrer clairement cette position, nous pouvons rappeler la célèbre phrase composée par Noam Chomsky (1957) « Les idées vertes incolores dorment furieusement » comme exemple syntaxiquement correct, mais sémantiquement absurde. Si cette phrase est clairement produite oralement par un locuteur natif sans trouble de la production de la parole et correctement transcrite par un auditeur natif sans trouble auditif dans un environnement calme, il est raisonnable de dire que l'énoncé est intelligible mais incompréhensible.

Dans son modèle de processus de communication, Lindblom (1990) explique que lorsque les informations dépendant du signal sont précises, l'auditeur est capable de comprendre le message sans information indépendante du signal. En revanche, lorsque les informations dépendant du signal sont insuffisantes, les informations indépendantes du signal deviennent cruciales pour comprendre le message du locuteur. Si un locuteur présentant un trouble de la production de la parole fournit aux auditeurs un signal imprécis, il tentera, avec la complicité de son interlocuteur, de compenser en augmentant les informations indépendantes du signal pour combler les lacunes laissées par des informations dépendantes du signal incomplètes ou compromises. Dans la vie de tous les jours, ces processus sont essentiels et doivent être largement utilisés dans la communication par et avec les patients. Mais dans le cadre d'une évaluation de la gravité d'un trouble de la parole, il est nécessaire de concentrer la mesure sur le locuteur lui-même en minimisant les effets liés à l'auditeur dont les variations peuvent être vues comme un bruit de mesure.

De façon classique, les tests d'intelligibilité sont effectués à partir de phrases ou de mots issus de listes de référence. Les limitations de ce type de test résident dans la capacité des auditeurs à restaurer les séquences distordues (Warren et al., 1970). Cet effet est d'autant plus fort que les auditeurs ont une connaissance forte des mots utilisés dans le test et que ces mots sont peu ambigus et donc fortement prédictibles. C'est généralement le cas des orthophonistes qui peuvent faire un usage si important de ces listes qu'ils/elles finissent par les connaître par cœur. On peut citer par exemple la BECD (Auzou et al., 2006) qui ne comporte que 50 mots. Le biais lié à cette connaissance (Rebourg et al, 2020) et donc à la forte influence des mécanismes perceptifs descendants est un score d'intelligibilité surévalué car la restauration phonémique de l'auditeur rend « transparentes » les distorsions de production.

La solution que nous proposons consiste à utiliser des pseudo-mots, c'est-à-dire des logatomes respectant les structures phonotactiques fréquentes du français, en grande quantité de façon à complètement neutraliser les effets de lexicalité ou d'apprentissage des items par les auditeurs. Au final, les auditeurs sont confrontés à une tâche de décodage acoustico-phonétique suivie d'une transcription écrite. Les détails de la construction du test sont donnés dans (Ghio et al., 2018). Le principe du test est de faire prononcer 52 pseudo-mots tirés aléatoirement d'une liste de 89346 formes possibles, sachant que chaque liste est, par construction, phonétiquement équilibrée. Les pseudo-mots ont été construits avec les formes $C(C)_1V_1C(C)_2V_2$ où $C(C)_i$ est une consonne isolée ou un groupe consonantique. Par exemples: stoumo, vurtant, muja, charou, leba, ranto…

Dans le cadre des troubles de la production de la parole, nous partons de l'hypothèse que l'information dépendante du signal est dégradée en raison de l'imprécision articulatoire et/ou phonatoire et que cette imprécision contribue à diminuer l'intelligibilité. Dans ce contexte, si le locuteur a l'intention de dire quelque chose mais que l'auditeur entend autre chose, nous émettons le postulat qu'il s'agit d'une erreur de production de parole. Nous supposons que le canal de communication est parfait (pièce silencieuse, lecture audio performante) et que l'auditeur est normo entendant.

2 Matériel et méthodes

Le projet C2SI (Carcinologic Speech Severity Index) est un projet financé par l'Institut National du Cancer dont l'objectif est d'obtenir une mesure de l'impact des traitements des cancers de la cavité buccale et du pharynx sur la production de la parole. Cette mesure est explorée à la fois par des méthodes perceptives et par traitement automatique de la parole (Astésano et al., 2018). Dans ce cadre, un certain nombre de patients ont été enregistrés et l'objectif de ce travail est de faire part des résultats de la mesure d'intelligibilité sur cette cohorte de locuteurs.

2.1 Locuteurs

117 locuteurs (39 sujets sains et 78 patients) ont été enregistrés dans le service d'oncoréhabilation de l'Oncopole à Toulouse. Les patients devaient répondre aux critères d'inclusion suivants:

- avoir un cancer T1 à T4 de la cavité buccale et / ou du pharynx;

- avoir bénéficié d'un traitement par chirurgie et / ou radiothérapie et / ou chimiothérapie;

- être à plus de 6 mois après la fin du traitement pour assurer la stabilité du trouble de la parole, qu'il soit audible ou non.

De même, les critères de non-inclusion étaient de présenter une autre source de troubles de la parole (par exemple le bégaiement) ou de présenter des problèmes cognitifs ou visuels incompatibles avec la conception du protocole d'évaluation. Ces critères de non-inclusion ont également été utilisés pour le recrutement de la population témoin.

2.2 Le corpus

Pour enregistrer le corpus, les locuteurs étaient confortablement installés dans une salle anéchoïque devant un écran d'ordinateur sur lequel était automatiquement affichée la forme orthographique du pseudo-mot à prononcer et une version audio produite en même temps. Cette double modalité, visuelle et auditive, a permis de limiter les erreurs de lecture ou d'éventuelles difficultés auditives et attentionnelles. Les enregistrements ont été effectués avec un microphone à condensateur cardioïde Neumann TLM 102 connecté à un enregistreur numérique FOSTEX. La fréquence d'échantillonnage était de 48 kHz.

Chaque locuteur prononçait une liste différente de pseudo-mots, tirés aléatoirement du dictionnaire des 89346 formes possibles, tout en intégrant des contraintes phonétiques identiques dans chaque liste. En effet, comme décrit dans (Ghio et al., 2016), le même nombre de phonèmes apparaissent dans chaque liste mais avec des combinatoires différentes, ce qui nous conduit à faire l'hypothèse que les listes sont équivalentes. Une fois le locuteur enregistré, le signal de parole était segmenté afin d'obtenir un fichier audio par pseudo-mot. Le corpus était donc composé de 117 locuteurs x 52 items = 6084 stimuli

2.3 Le test de perception

40 auditeurs francophones natifs, sans trouble de l'audition, non spécialistes des troubles de la parole, ont retranscrit ces productions via le logiciel LANCELOT (André et al., 2003). Ils recevaient l'instruction suivante: " Vous allez entendre des non-mots. Un non-mot est une combinaison de sons de la langue française qui n'a pas de signification (ex: gloutu). En respectant les règles de l'orthographe du français, vous devrez transcrire ce que vous entendrez. Certaines prononciations seront difficiles à identifier, mais dans tous les cas, vous devrez fournir une transcription ".

Les différents stimuli étaient répartis dans plusieurs blocs et présentés dans un ordre aléatoire. Chaque stimulus a été transcrit par 3 auditeurs différents, ce qui représente finalement 18 252 réponses (6084 stimuli x 3). Ces tests de perception ont eu lieu au Centre d'Expérimentation sur la Parole (http://cep.lpl-aix.fr/) du Laboratoire Parole et Langage à Aix-en-Provence. Chaque auditeur, portant un casque audiophonique Superlux HD 681B a transcrit sur ordinateur environ 456 stimuli en 3 blocs. L'intensité de la présentation a été préréglée par l'auditeur pour être confortable et optimale pour la tâche. Chaque test a commencé avec quatre stimuli d'entraînement. Chaque élément était présenté une fois automatiquement mais l'auditeur pouvait répéter la lecture deux fois.

2.4 Le prétraitement des réponses

Au terme du test de perception, nous avons récupéré 18 252 réponses. Comme détaillé dans (Ghio et al, 2018), ces transcriptions orthographiques ont été phonétisées et comparées aux formes phonétiques attendues des pseudo-mots. Cette comparaison utilise un algorithme de Wagner-Fisher fondé sur un calcul de traits distinctifs déviants entre la forme cible et la forme transcrite. Notre mesure, que nous baptisons PPD (Perceived Phonological Deviation), représente le nombre moyen de traits mal perçus par phonème. Etant donné que le test part du postulat que le canal de communication est parfait (pièce silencieuse, lecture audio performante) et que l'auditeur est normo entendant, l'erreur perçue est directement mise en lien avec une erreur de production.

Puisque chaque pseudo-mot a été transcrit par 3 auditeurs différents, nous obtenons 3 scores pour chaque pseudo-mot. Pour chaque pseudo-mot, nous avons calculé la moyenne des réponses des 3 auditeurs, ainsi que la médiane. Afin d'assurer la cohérence des réponses et l'accord entre les auditeurs, nous avons appliqué une détection des valeurs aberrantes. Nous considérons comme aberrant un résultat qui s'écarte de 2.5 traits de la médiane. Comme +/- 2.5 traits donne une étendue de 5, c'est la moitié de la différence maximale entre 2 phonèmes qui est de 10. À la suite de ce filtrage, nous avons supprimé 1,2% des transcriptions. À cette étape, nous avons un score par pseudo-mot par locuteur.

Dans un deuxième temps, nous avons calculé la moyenne de ces 52 scores par locuteur. Nous obtenons ainsi, pour chacun des 117 locuteurs, un score PPD qui reflète le nombre moyen de traits déviants par phonème.

3 Résultats et discussion

Tous les tests statistiques ont été effectués dans l'environnement logiciel R version 3.4.4 (R Core team, 2017). Nous rappelons que le score PPD (Perceived Phonological Deviation) rend compte d'une dégradation et que plus il est bas, plus nous considérons que l'intelligibilité est bonne.

3.1 Le score PPD en fonction du groupe de locuteur

Le score PPD des sujets sains est en moyenne à une distance de 0,48 trait / phonème (écart-type = 0.23) alors que cette distance atteint 1,29 pour les patients (écart-type = 0.63) (Figure 1). Les données n'étant pas gaussiennes, nous avons effectué une transformation logarithmique du score. Nous avons obtenu des distributions gaussiennes (test de Shapiro, $p > 0,05$) et des variances homogènes (test de Bartlett, $p > 0,05$). L'analyse de la variance (ANOVA) a été réalisée avec le log-score comme variable et le « groupe de locuteurs » comme facteur. La différence entre les deux groupes était significative ($F(1,115) = 127.2$; $p < 0,001$).

On remarque que même les sujets contrôle ont un score PPD non nul, ce qui s'explique par le fait que les auditeurs perçoivent des exemplaires de phonèmes légèrement altérés y compris par des locuteurs non pathologiques. Nous mesurons là les distorsions « normales » qui se produisent dans la production de la parole et qui sont en général rectifiées par l'accès au lexique et au sens, mécanismes inhibés dans notre tâche de décodage acoustico-phonétique.

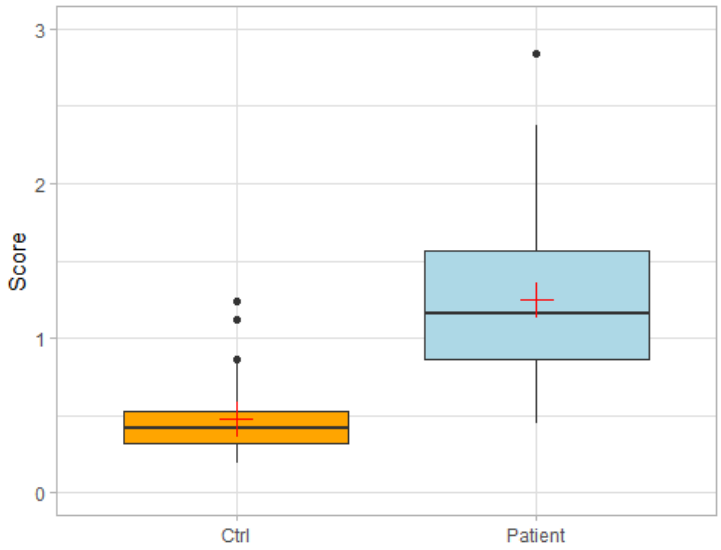


Figure 1 : score PPD pour le groupe contrôle (Ctrl) et pour les groupe des Patients

3.2 Le score PPD par locuteur et son pouvoir classificateur

La Figure 2 illustre la répartition du score PPD par locuteur en fonction du groupe contrôle (CTRL) ou des patients. On remarque une répartition assez distincte entre les sujets sains (faibles scores) et la majorité des patients, ce qui laisse présager un bon pouvoir discriminant de ce test. Certains patients ont un score bas, ce qui peut traduire un faible impact fonctionnel sur la parole du traitement du cancer sur ces sujets. Inversement, quelques locuteurs du groupe contrôle s'illustrent par des scores élevés. En analysant les raisons de ces scores, nous avons observé que ces sujets avaient eu des difficultés ou un manque d'attention à produire les pseudomots. Par exemple, le sujet TTT88 prononce distinctement [mjozo] le pseudomot « minso » ou encore [plɔkso] la séquence « plouco ». Il est donc normal que les auditeurs aient transcrits ces deux items « miozo » ou « plocso », ce qui a généré un PPD important car la cible était « minso » et « plouco ».

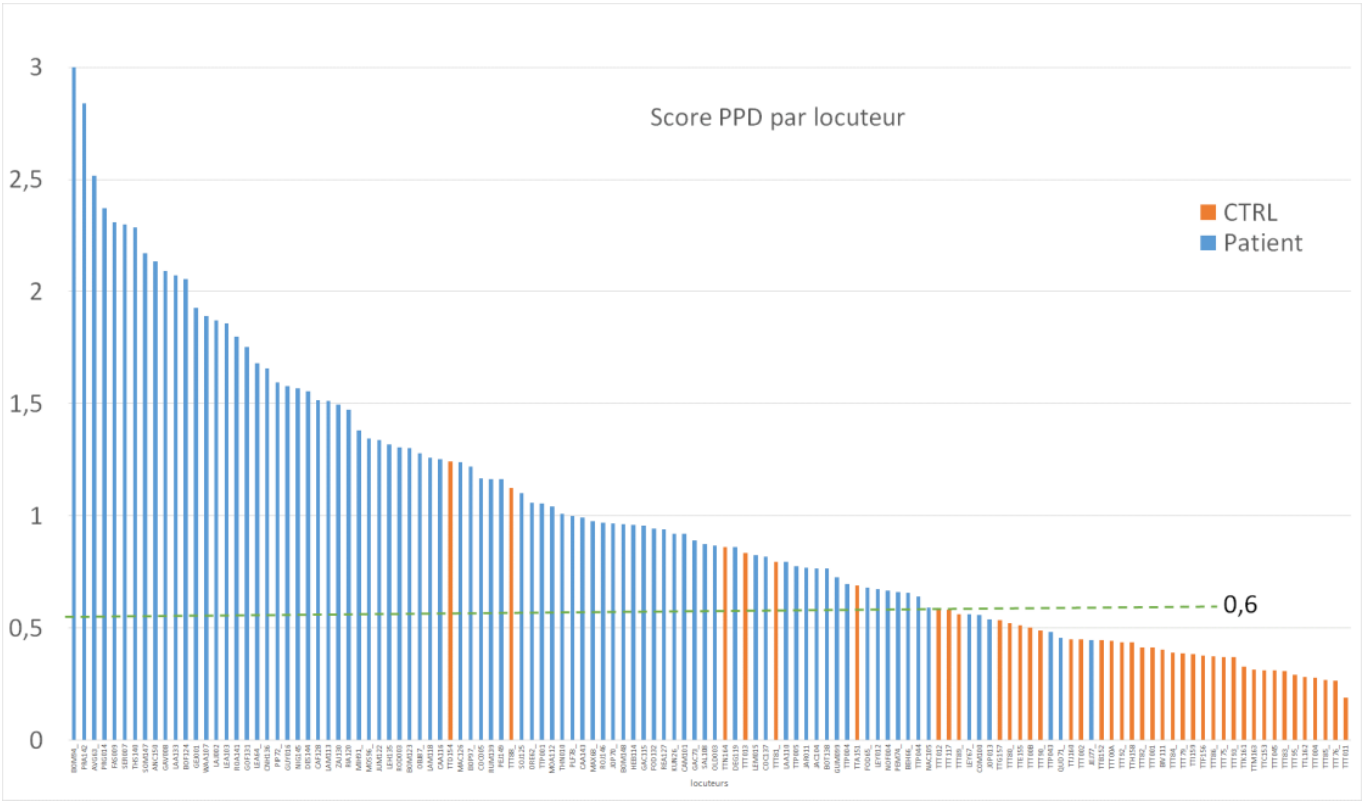


Figure 2 : score PPD par locuteur.
La ligne pointillée horizontale indique le seuil optimal de distinction patient/ contrôle

Afin de mesurer le pouvoir discriminant du score PPD, nous avons établi une courbe de sensibilité/spécificité, qui permet de mesurer la performance du classificateur binaire, c'est-à-dire le pouvoir de catégorisation des deux groupes de locuteurs sur la base du score PPD. La fonction ROC (Receiver Operating Characteristic) se présente sous la forme d'une courbe qui trace le Taux de Vrais Positifs (patients détectés comme patients) en fonction du Taux de Faux Positifs (fraction des sujets sains qui sont incorrectement détectés comme des patients) pour tous les seuils de classification (Delacour et al., 2005).

Cette courbe (Figure 3) est intéressante car elle permet de prédire la performance de classification par la mesure de l'aire sous la courbe (Area Under the Curve ⇔ AUC). L'AUC, qui mesure l'intégralité de l'aire à deux dimensions située sous l'ensemble de la courbe ROC, indique la probabilité pour que la fonction PPD place un patient devant un sujet contrôle (dans le meilleur des cas, l'AUC vaut 1). Elle permet d'évaluer l'intérêt diagnostique d'un test. Dans notre cas, l'AUC est égal à 0.94, ce qui correspond à une haute précision. Nous pouvons donc affirmer que nous avons obtenu une validité de construit du test fondé sur le score PPD car il permet de distinguer avec une haute précision les locuteurs sains du groupe des patients.

La courbe ROC permet également de déterminer la valeur seuil qui va optimiser le test. Dans notre cas, la question se pose sur la valeur seuil du PPD en dessous duquel se situe la normalité et au-dessus duquel on entre dans le dysfonctionnement. Intuitivement, celle-ci peut être identifiée comme étant le point de la courbe le plus éloigné de la diagonale représentant le test d'apport nul. Ce point correspond également au maximum de l'indice de Youden (Se + Sp –1) » (Delacour et al., 2005). Dans notre cas, le maximum de l'indice de Youden vaut 0.783, qui correspond à un seuil PPD égal à 0.6, indiqué sur la Figure 2 par la ligne pointillée horizontale..

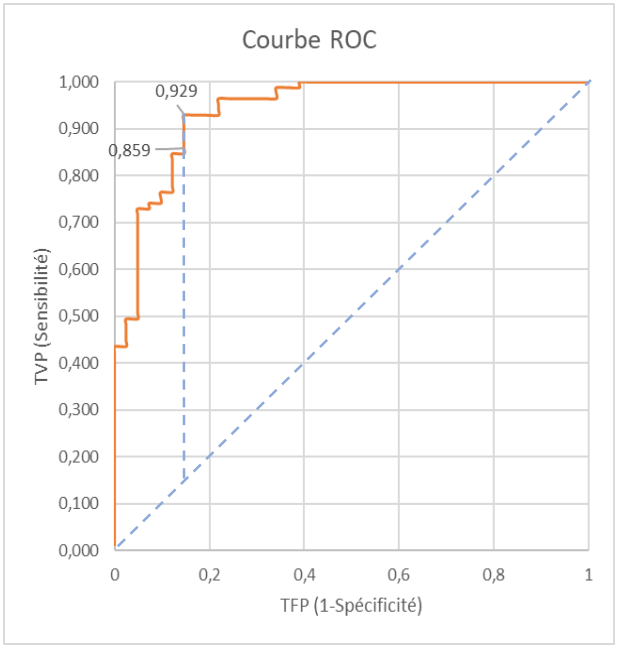


Figure 3 : courbe ROC établie sur le pouvoir classificateur du PPD pour distinguer les sujets sains des patients

3.3 La corrélation avec une mesure clinique de jugement subjectif de la sévérité

Dans le cadre du projet C2SI, les patients et les locuteurs sains ont produit de la parole spontanée obtenue avec une tâche de description d'image (Astesano et al., 2018). La première minute de chaque enregistrement a été utilisée pour une évaluation clinique subjective telle que définie par Kent et al. (1989). Cette évaluation globale consiste en un jugement de la sévérité du trouble sur une échelle ordinale entre 0 (altération sévère) et 10 (parole normale). Le score relatif à un locuteur est obtenu par la moyenne de six orthophonistes considérés comme experts en troubles de la parole. Afin d'évaluer la fiabilité inter-juges, un coefficient de corrélation inter-classes (ICC) a été calculé. Le degré de concordance entre les notes du jury est bon (r = 0,77). Le jury est homogène et fait office de « gold standard » (Balaguer et al., 2019). Afin de vérifier la validité concourante du test PPD, nous avons examiné la corrélation entre le score PPD et le jugement clinique de la sévérité (Figure 4). Ces deux grandeurs sont bien corrélées avec un $R_{spearman}$ = -0.85

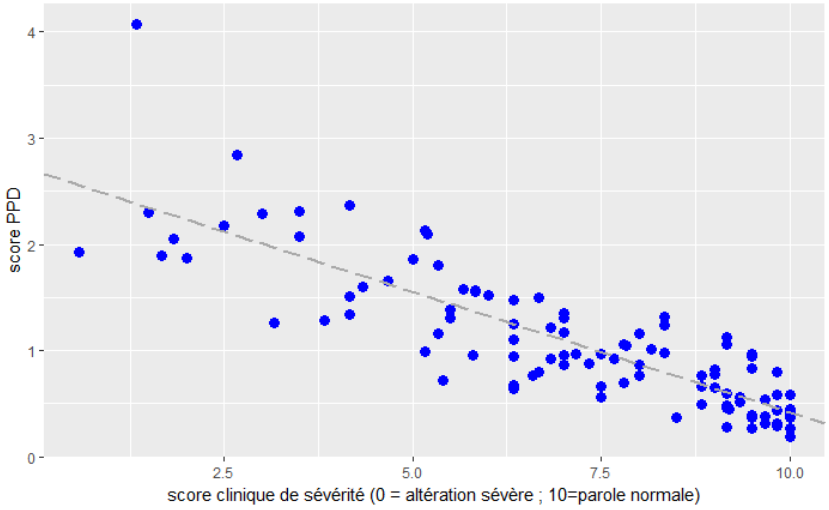


Figure 4 : corrélation entre score PPD et score clinique de sévérité

4 Conclusion

Le test d'intelligibilité que nous proposons est fondé sur des pseudomots ainsi que sur une métrique organisée autour du comptage de traits phonétiques mal identifiés par des auditeurs. Si le test est correctement administré, ces erreurs d'identification peuvent être directement associées aux troubles de la production de la parole. Nous avons conscience de notre définition stricte de l'intelligibilité vue comme « la quantité de parole comprise à partir du seul signal acoustique » Keintz et al. (2007). Un tel test doit être complété par des évaluations de la compréhensibilité comme cela a été fait dans le cadre du projet C2SI par (Nocaudie et al, 2018), pour rendre compte de la réalité du déficit fonctionnel de parole chez ces patients.

Dans ce travail, nous considérons que le score PPD a été validé dans le construit du test dans la mesure où il remplit sa fonction discriminante pour distinguer avec précision les patients des sujets sains. Nous observons aussi une validité concourante puisque cette mesure est corrélée avec le gold standard du jugement subjectif clinique de la sévérité du trouble de la production de la parole. Comparée à cette évaluation subjective de la sévérité, la tâche que nous proposons est une véritable tâche linguistique de décodage et non un processus d'interprétation subjectif. Elle correspond donc plus au processus normal de la communication orale. Grâce au matériel linguistique utilisé (les pseudo-mots), l'évaluation par décodage acoustico-phonétique est moins dépendante des mécanismes top down de la perception et donc moins dépendante de l'auditeur. Cette réduction de la dépendance aux spécificités de l'auditeur pourrait permettre de réduire les phénomènes de variabilité qui fragilisent les résultats des tests subjectifs. De plus, l'utilisation de pseudo-mots présente l'avantage de disposer d'un matériel linguistique bien maîtrisé, standardisé et en très grande quantité. Par conséquent, le score PPD obtenu est moins sujet à un biais d'évaluation. Comme le montrent les résultats, l'évaluation de l'intelligibilité obtenue avec des auditeurs naïfs lors de la tâche de décodage acoustico-phonétique est cohérente avec celle des experts. Bien qu'artificiel dans son matériel, ce test est finalement écologique car il est pertinent avec des auditeurs non experts. Cela représente aussi un avantage supplémentaire, d'un point de vue économique car le recours à des évaluateurs naïfs est moins contraignant que le recrutement de spécialistes.

Au terme de ces travaux, de nouvelles pistes émergent de ces résultats prometteurs. L'obtention de matrices de confusion entre phonèmes permettrait de dépasser la seule valeur scalaire du score PPD. Une analyse plus fine des traits altérés pourrait revêtir une importante valeur d'orientation thérapeutique. Un certain nombre de questions devra être réglé : le tirage aléatoire des listes permet-il une équivalence de résultat ? Autrement dit, un même locuteur confronté à deux listes aura-t-il des résultats identiques avec les 2 listes ? Est-il possible de réduire le nombre d'items (52) de façon à réduire le temps de passation tout en maintenant des résultats robustes ? L'un des axes de recherche prévus dans un avenir proche sera d'appliquer cette méthode sur d'autres pathologies comme par exemple, dans les dysarthries (Projet ANR ANR-18-CE45-0008 Rugbi). Nous travaillons également à évaluer la contribution de cette mesure en la comparant à des méthodes automatiques qui pourraient être utilisées comme un auditeur robot dont les caractéristiques déterministes réduiraient un peu plus la variabilité perceptive humaine. Mais l'approbation de ces techniques passera là encore par une validation concourante.

Remerciements

Ce travail a été soutenu par la subvention n ° 2014-135 de l'Institut National pour le Cancer (INCA) projet C2SI et par la subvention ANR-18-CE45-0008 de l'Agence Nationale de la Recherche en 2018 Projet RUGBI

Références

ANDRÉ C, GHIO A, CAVÉ C, TESTON B . (2003) PERCEVAL: a Computer-Driven Systemfor Experimentation on Auditory and Visual Perception. International Congress of Phonetic Sciences (ICPhS), Barcelona, Spain. pp.1421-1424.

ASTÉSANO C. , BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., GIUSTI L. et al. (2018), Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer, LREC, 7-12 May 2018, Miyazaki (Japan)

AUZOU P, ROLLAND-MONNOURY V. (2006), Batterie d'évaluation de la dysarthrie, 1st ed. Isbergues:

BALAGUER, M., BOISGUÉRIN, A., GALTIER, A., GAILLARD, N., PUECH, M., & WOISARD, V. (2019). Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *136*(5), 355–359

CHOMSKY, N (1957). Syntactic Structures. The Hague/Paris: Mouton. p. 15

DELACOUR H, SERVONNET A, PERROT A, VIGEZZI JF, RAMIREZ JM (2005) La courbe ROC (receiver operating characteristic) : principes et principales applications en biologie clinique. Annales de Biologie Clinique. 2005;63(2):145-154.

FONTAN L, TARDIEU J, GAILLARD P, WOISARD V, RUIZ R. (2015) Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise, Journal of Speech Language and Hearing Research. 2015 Jun;58(3):977-86. https://doi.org/10.1044/2015_JSLHR-H-13-0335

GHIO A., GIUSTI L., BLANC E., PINTO S., LALAIN M, ROBERT D., FREDOUILLE C., WOISARD V. (2016) Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ?. *Journées d'Etude sur la Parole*, Paris, France, p.589-596

GHIO A, LALAIN M, GIUSTI L, POUCHOULIN G, ROBERT D, et al.(2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. XXXIIe Journées d'Etudes sur la Parole, LPL, 2018, Aix-en-Provence, France. pp.285-293,

KEINTZ, C. K., BUNTON, K., & HOIT, J. (2007). Influence of visual information on the intelligibility of dysarthric speech. American Journal of Speech-Language Pathology, 16, 222–234.

KENT RD, WEISMER G, KENT JF, ROSENBEK JC.(1989) Toward phonetic intelligibility testing in dysarthria. Journal of Speech and Hearing Disorders. 1989 Nov;54(4):482-99.

LAARIDH I, FREDOUILLE C, GHIO A, LALAIN M, , et al (2018). Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers. *Interspeech,* 2943-2947

LINDBLOM, B. (1990). On the communication process: Speaker listener interaction and the development of speech. Augmentative and Alternative Communication, 6, 220–230.

NOCAUDIE O, ASTÉSANO C, GHIO A, LALAIN M, WOISARD V (2018). Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx. 32e JEP, pp.196-204,

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

REBOURG M, LALAIN M, GHIO A, FREDOUILLE C, FAKHRY N, WOISARD V, (2020), Évaluer l'intelligibilité, mots ou pseudo-mots ? Comparaison entre deux groupes d'auditeurs, JEP, ce volume

WARREN, R.M., WARREN, R.P., (1970). Auditory illusions and confusions. Scientific American. 223(6), 30–36.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal

Adrien Gresse Mathias Quillot Richard Dufour Jean-François Bonastre
LIA, Avignon Université, France
prénom.nom@univ-avignon.fr

RÉSUMÉ

La recherche d'acteurs vocaux pour les productions audiovisuelles est réalisée par des directeurs artistiques (DA). Les DA sont constamment à la recherche de nouveaux talents vocaux, mais ne peuvent effectuer des auditions à grande échelle. Les outils automatiques capables de suggérer des voix présentent alors un grand intérêt pour l'industrie audiovisuelle. Dans les travaux précédents, nous avons montré l'existence d'informations acoustiques permettant de reproduire des choix du DA. Dans cet article, nous proposons une approche à base de réseaux de neurones pour construire une représentation adaptée aux personnages/rôles visés, appelée p-vecteur. Nous proposons ensuite de tirer parti de données externes pour la représentation de voix, proches de celles d'origine, au moyen de méthodes de distillation de la connaissance. Les expériences menées sur des extraits de voix de jeux vidéo montrent une amélioration significative de l'approche p-vecteur, avec distillation de la connaissance, par rapport à une représentation x-vecteur, état-de-l'art en reconnaissance du locuteur.

ABSTRACT

Learning voice representation using knowledge distillation for automatic voice casting

The search for voice actors for audiovisual productions is carried out by artistic directors (DA). DA are constantly on the lookout for new vocal talent, but are unable to conduct large-scale auditions.Automatic tools able to suggest the most suited voices are of a great interest for audiovisual industry. In previous work, we have shown the existence of acoustic information allowing us to reproduce DA choices. In this article, we propose a neural network-based approach to construct a representation adapted to targeted characters/roles, called p-vector. We then propose to take advantage of external data, close the origin one, for the representation of voices, using knowledge distillation methods. Experiments carried out on voice extracts from video games show a significant improvement in the p-vector representation, including knowledge distillation, compared to x-vectors, state-of-the-art representation in speaker recognition.

MOTS-CLÉS : distillation de la connaissance, p-vecteur, similarité perceptive, réseaux de neurones profonds.

KEYWORDS: knowledge distillation, p-vector, perceptual similarity, deep neural network.

1 Introduction

Les entreprises visant la diffusion internationale d'oeuvres audiovisuelles (films, séries, jeux vidéo...) atteignent alors un public multilingue et multiculturel. Ainsi, les producteurs de ces créations audiovisuelles accordent de plus en plus d'attention aux voix qu'ils attribuent à un personnage ou à un rôle

Volume 1 : Journées d'Études sur la Parole, pages 280–288.
HAL : hal-02798550.

particulier afin de renforcer le sentiment d'immersion du public. Ce processus de changement de la voix originale dans une langue par une nouvelle voix dans un autre langage est appelé *doublage vocal*. Il consiste à remplacer l'intégralité des dialogues de la création originale par de nouveaux acteurs vocaux dans le contexte linguistique et culturel ciblé. Dans ce contexte, la sélection des voix appropriées dans une langue cible en fonction à la fois de la voix d'origine et du rôle, est une tâche cruciale, appelée *casting vocal*. Habituellement, un expert humain, appelé *directeur artistique* (DA), effectue la tâche de casting de voix dans des sociétés de doublage.

Le problème majeur du doublage vocal réside dans le fait que la "similitude" recherchée entre une voix originale et une voix doublée est loin d'être une simple ressemblance acoustique. Il comprend les caractéristiques socioculturelles des langues et des pays sources et cibles. De plus, il n'y a pas de vocabulaire bien établi pour décrire les voix, les personnages et les effets immersifs. Il y a deux limites à la façon dont les DA effectuent la tâche de casting vocal : les choix des DA intègrent un certaine subjectivité, liée à leurs propres caractéristiques socioculturelles, et 2), les DA ne peuvent pas écouter et mémoriser un nombre très élevé de voix. Par conséquent, un DA travaille généralement avec une liste réduite d'acteurs qu'il a écoutés et/ou avec lesquels il a déjà travaillé.

Les outils automatiques capables de mesurer l'adéquation potentielle entre une voix originale dans une langue source et une voix doublée dans une langue et un contexte cibles, présentent un grand intérêt pour l'industrie audiovisuelle. Ils aideront les DA à remédier aux problèmes susmentionnés et à ouvrir la porte à de nouveaux talents de voix, par exemple en pré-sélectionnant un nombre raisonnable de candidats au sein d'un très large ensemble de voix.

La similarité vocale dans le contexte du doublage de voix a été étudié dans (Obin *et al.*, 2014; Obin & Roebel, 2016). Les auteurs ont montré l'importance de certaines caractéristiques para-linguistiques (*e.g.* âge, genre, état du locuteur, qualité de la voix...). Dans (Gresse *et al.*, 2017), les auteurs proposent d'estimer la proximité de "doublage" entre deux voix (une dans la langue source, et une dans la langue cible) au moyen d'une approche i-vecteur/PLDA, inspirée du domaine de la reconnaissance du locuteur. (Gresse *et al.*, 2019) supposent que des informations, ou a minima des indices, liées au casting réalisé par les DA sont présentes dans les voix de doublage choisies. L'approche proposée permet de distinguer les paires *cible* (*i.e.* une voix dans une langue source associée à la voix du personnage correspondant dans la langue cible) de *non-cible* (*i.e.* voix qui ne correspond pas au bon personnage). Une limite de ce travail est que l'utilisation de l'apprentissage supervisé binaire donne de faibles capacités de généralisation au modèle, étant donné que l'interpolation ne peut s'appuyer que sur des contre-exemples.

Des travaux récents en reconnaissance du locuteur (Variani *et al.*, 2014; Snyder *et al.*, 2016, 2017, 2018) ont montré que des représentations au moyen de réseaux de neurones profonds, et d'apprentissage bout-en-bout (end-to-end), surpassent la représentation de référence i-vecteur. Dans cet article, nous proposons d'apprendre une représentation latente originale du personnage/rôle, appelée p-vecteur, à partir d'une approche fondée sur les réseaux de neurones. Ces p-vecteurs sont conçus pour aider le système à avoir une meilleure assimilation de la dimension du personnage, et par conséquent à mieux gérer les voix inconnues. Cette approche constitue la première contribution de cet article.

Néanmoins, un frein à l'utilisation d'une telle approche s'appuyant sur les réseaux de neurones est la nécessité d'une grande quantité de données dans le domaine considéré. Dans notre contexte, la seule information que nous pouvons utiliser est la sélection vocale de l'opérateur humain (DA). Dans les travaux que nous avons initiés, seul un petit nombre de personnages est mis à notre disposition. Dans cet article, nous proposons de remédier à ce problème en appliquant des méthodes de distillation de la connaissance en utilisant des données supplémentaires, provenant d'un domaine proche, pour

extraire les informations spécifiques au personnage/rôle. Plus généralement, nous pensons que les connaissances extraites, par exemple des jeux vidéo, pourraient être transférées à d'autres contextes, tels que les personnages de voix d'émissions de télévision.

Cet article est organisé comme suit. Nous présentons d'abord l'approche p-vecteur et le cadre général de distillation de la connaissance dans la partie 2. Ensuite, nous détaillons le corpus et nous décrivons le protocole expérimental que nous avons mis en place dans la partie 3. Nous présentons nos résultats et les discutons dans la partie 4. Enfin, les conclusions et perspectives sont données dans la partie 5.

2 Approche

2.1 Représentation dédiée au personnage

Ces dernières années, des architectures à base de réseaux de neurones profonds ont été proposées pour apprendre des espaces de représentation des données (Bengio *et al.*, 2013). Nous proposons d'apprendre une représentation dédiée aux voix jouées, appelée p-vecteur. L'espace p-vecteur (p signifiant "personnage") est optimisé sur une tâche de discrimination personnage/rôle. Il permet de projeter des segments de voix d'une manière qui maximise la variabilité des personnages.

En général, la représentation des données en entrée de méthodes d'apprentissage automatique a un impact fort sur les performances des applications. Ici, nous adoptons la représentation x-vecteur, initialement introduite en reconnaissance automatique du locuteur (Snyder *et al.*, 2018). Une grande quantité de données provenant de nombreux locuteurs sont utilisées pour créer l'espace de plongement des locuteurs (*speaker embeddings*). Des segments audio sont projetés dans cet espace et caractérisés par des x-vecteurs. Les x-vecteurs sont considérés ici comme une représentation compacte et de taille fixe d'une séquence vectorielle de paramètres acoustiques de longueur variable. Nous faisons l'hypothèse que les plongements de locuteurs contiennent des informations intriquées correspondant à la dimension personnage/rôle. Nous proposons donc de construire un nouvel espace de représentation (p-vecteur) capable de discriminer les différents personnages.

2.2 Distillation de la connaissance

Dans ce travail, nous devons traiter un nombre relativement restreint de données. Nous proposons d'utiliser la distillation de la connaissance afin d'exploiter des données supplémentaires d'un domaine proche pour pallier ce problème.

La distillation (Lopez-Paz *et al.*, 2016) unifie deux techniques qui introduisent toutes deux un maître pour guider un modèle d'élève tout au long de son processus d'apprentissage. La première technique introduit le concept d'information privilégiée (*Privileged Information*) (Vapnik & Izmailov, 2015) en ajoutant un nouvel élément $x_i^\star$ à la paire caractéristique-étiquette $(x_i,\, y_i)$, avec $i \in [1...N]$ où N correspond au nombre d'exemples. La deuxième technique, appelée distillation de la connaissance (*Knowledge Distillation*) (Hinton *et al.*, 2015), permet à un réseau neuronal simple de résoudre une tâche compliquée en distillant les connaissances à partir d'un modèle "lourd". Plus généralement, le maître offre au modèle étudiant la possibilité d'apprendre à partir d'une décision qui n'est pas contenue dans l'échantillon d'entraînement (Lopez-Paz *et al.*, 2016). En règle générale, un réseau neuronal utilisant une fonction d'activation *softmax* fournit une probabilité pour chaque classe obtenue avec la formule suivante :

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)}$$

où T fait référence à la température et z_i désigne la sortie calculée pour chaque classe de la couche finale. Une valeur plus élevée de T donne une distribution de probabilité plus progressive sur toutes les classes. Le fait est que le vecteur de probabilité q_i contient beaucoup plus d'informations qu'un simple codage à chaud (*one-hot encoding*). La distillation consiste à élever la température jusqu'à ce que le modèle du maître produise des cibles souples (*soft-targets*) appropriées. Ces dernières correspondent à l'information privilégiée. Comme illustré dans la figure 1, nous adaptons le modèle

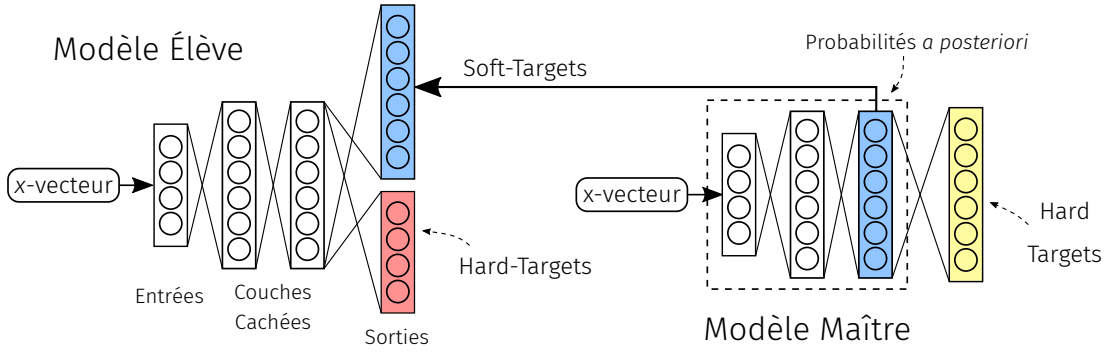


FIGURE 1 – Le modèle maître est entraîné à prédire des soft-targets afin que le modèle étudiant puisse les utiliser. Les deux modèles peuvent être entraînés sur le même corpus ou sur un corpus différent.

de l'élève aux cibles fixes (*hard-targets*, *i.e.* les étiquettes de personnages) et aux cibles souples (*soft-targets*) provenant du maître. Pour ce faire, nous utilisons un paramètre d'imitation noté λ qui contrôle la priorité entre l'imitation des probabilités *soft* et les prédictions habituelles des étiquettes *hard* pendant l'entraînement du modèle étudiant. Ceci est rendu possible en minimisant la perte :

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} [(1-\lambda) l(y_i,\ q_i) \ +\ \lambda l(s_i,\ q_i)]$$

où l désigne la perte d'entropie croisée et s_i fait référence aux *soft-targets* du modèle du maître. Le cadre maître-élève a été utilisé dans plusieurs travaux (Price *et al.*, 2016; Markov & Matsui, 2016; Li *et al.*, 2017; Watanabe *et al.*, 2017; Asami *et al.*, 2017; Joy *et al.*, 2017) pour une grande variété de tâches telles que la reconnaissance de la parole robuste au bruit, l'adaptation à un domaine, et la normalisation de locuteurs. L'approche proposée étend à l'origine ce cadre aux voix jouées et spécifiquement à la représentation des personnages/rôles.

Étant donné le nombre limité de personnages dans notre corpus, nous entraînons le modèle du maître sur un jeu de données supplémentaires contenant plus d'étiquettes de personnages. Nous supposons que cela pourrait aider le modèle étudiant à apprendre une représentation robuste et plus générale en s'adaptant aux *soft-targets* du maître.

Dans cet article, nous avons tout d'abord proposé, pour le casting vocal, un nouvel espace de représentation appelé p-vecteur obtenu à partir d'une couche d'embeddings d'un réseau de neurones profonds. Afin de pallier le problème de limitation des données, nous avons proposé une approche par distillation de la connaissance pour améliorer la représentation personnage/rôle.

3 Protocole expérimental

3.1 Corpus

Les extraits vocaux des personnages du jeu de rôle *Mass Effect 3* constituent notre corpus principal. Initialement édité en anglais, ce jeu a été traduit et doublé dans d'autres langues. Dans nos expériences, nous utilisons les versions anglaise et française des séquences audio, représentant $7,5$ d'heures de parole dans chaque langue. Les segments vocaux durent en moyenne 3 secondes, où chaque segment correspond à une interaction vocale unique. Un personnage est alors défini par un couple unique de voix jouées français-anglais. Pour éviter tout biais en termes d'identité du locuteur, nous considérons uniquement un sous-ensemble restreint de 31 personnages différents (13 féminins et 18 masculins), où nous sommes certains qu'aucun des acteurs ne joue plus d'un personnage. Chaque jeu de données anglais et français contient 10 000 segments vocaux.

Afin de remédier au nombre limité de personnages dans le corpus *Mass Effect 3*, nous utilisons des données supplémentaires d'un autre jeu vidéo multilingue appelé *Skyrim*. Nous limitons ce corpus aux dialogues anglais et français, pour un total de 120 heures de discours. Pour chaque langue, nous avons 50 000 segments annotés avec 30 étiquettes de personnages différents (7 féminins et 23 masculins). Comme nous n'avons pas suffisamment de garantie sur la correspondance français-anglais des segments et que nous ne sommes pas certains qu'un acteur joue un rôle unique, nous n'utilisons pas ce corpus dans l'étape d'évaluation. Il ne sert donc qu'à transférer les connaissances du maître au modèle étudiant dans le processus de distillation. Notons qu'il n'y a pas d'intersection entre les acteurs de *Skyrim* et *Mass-Effect 3*, ce qui évite un biais de locuteur dans l'ensemble de test. Enfin, tous les segments vocaux sont des fichiers audio enregistrés en studio de haute qualité. Tous les segments d'une durée inférieure à 1 seconde ont été supprimés.

3.2 Représentation des données

Nous réalisons une paramétrisation acoustique classique des segments audio que nous transformons en une séquence de caractéristiques de dimension 60 contenant 20 MFCCs incluant le log énergie et les dérivées de premier et second ordre ($\Delta + \Delta\Delta$). Nous utilisons une fenêtre glissante de Hamming de 20 ms (chevauchement de 10 ms), pour calculer les paramètres. Nous effectuons une normalisation des moyennes cepstrales et une détection d'activité vocale (VAD) pour supprimer les trames de faible énergie qui correspondent principalement au silence. Un système x-vecteur a été construit avec la boîte à outils Kaldi (Povey *et al.*, 2011) et entraîné sur le corpus Voxceleb (Chung *et al.*, 2018).

3.3 Protocole d'apprentissage

Le nombre de segments de voix dans le corpus *Mass Effect 3* n'est pas très bien réparti entre les différents personnages, ceux-ci n'ayant pas la même importance au sein du jeu. En conséquence, nous ne sélectionnons que 16 personnages (5 féminins, 11 masculins) qui ont tous, au moins, 90 segments vocaux dans les deux langues (anglais et français). Les segments sont tous choisis au hasard. De plus, nous créons une validation croisée en k-parties sur cet ensemble de personnages afin d'en avoir 4 dans chaque pli. Ainsi, nous avons $k = 4$ cas distincts, notés A, B, C et D qui couvrent tous les personnages, chaque cas impliquant 12 personnages pour l'apprentissage et 4 pour l'évaluation. Ces 4 personnages sont donc complètement absents du corpus d'apprentissage (ils ne partagent aucune étiquette ni aucun locuteur avec les données d'apprentissage), ce qui rend la tâche d'appariement de voix décrite dans 3.4 extrêmement difficile. Enfin, 20 % des données d'entraînement sont utilisées

pour la validation. En ce qui concerne le corpus additionnel, nous avons choisi le même nombre de segments pour chacun des 30 personnages et divisé en deux parties avec le même ratio affecté à la validation. Comme nous l'avons dit précédemment, aucune donnée de *Skyrim* n'est utilisée pour le test.

Les deux modèles maître et étudiant suivent la même architecture de réseau de neurones. Nous créons un Perceptron multicouche (MLP) en utilisant la boîte à outils Keras (Chollet *et al.*, 2015). Nous connectons une couche d'entrée avec 512 dimensions à 3 couches cachées de dimension 256 plus une couche d'embedding (*i.e.* correspondant aux p-vecteurs) de dimension 64, enfin une couche de sortie finale avec une fonction d'activation *softmax* . Les couches cachées sont combinées à une fonction d'activation tangente hyperbolique. Nous appliquons un dropout dans les 4 couches cachées avec les taux suivants : $0,25$, $0,25$, $0,25$, $0,5$. Nous utilisons une initialisation *Xavier* (Glorot & Bengio, 2010) et nous utilisons l'optimiseur *Adadelta* avec sa configuration par défaut pour résoudre la minimisation de la fonction de perte d'entropie croisée. De plus, nous utilisons une taille de batch de 12 exemples et nous entraînons le modèle sur 300 époques pendant que nous surveillons la fonction de perte sur l'ensemble de validation pour éviter le sur-apprentissage.

Le modèle du maître est entraîné sur les caractéristiques et étiquettes de l'ensemble de données supplémentaire (*Skyrim*), considéré comme une information privilégiée. Le modèle du maître peut être considéré comme un discriminateur de personnages/rôles. Ensuite, nous utilisons le maître pour calculer les *soft-targets Mass Effect 3* et former le modèle de l'élève sur les *hard-* et soft-targets de ce corpus. L'élève apprend à ajuster les 12 *hard-targets* et les 30 *soft-targets* en fonction du paramètre λ qui contrôle l'influence entre l'imitation des *soft-* et *hard-targets* pendant la phase d'apprentissage. Enfin, les p-vecteurs sont extraits de la couche d'embeddings du modèle de l'élève.

3.4 Evaluation

Pour évaluer la qualité de la représentation apprise, nous effectuons d'abord une analyse de clustering avec l'algorithme des k-moyennes sur les embeddings extraits (p-vecteurs). Nous avons expressément défini $k = 4$ pour refléter le nombre de personnages se trouvant dans l'ensemble de test. Tous les segments de voix qui sont ensuite rassemblés dans le même cluster sont affectés au personnage le plus représenté, de sorte qu'un cluster ait une seule étiquette. Ainsi, un score de F-mesure est calculé sur les segments sachant l'hypothèse de chaque cluster. Notons que plusieurs clusters peuvent être affectés au même personnage, ce qui pourrait être un problème. Mais nous considérons que cela reste un cas particulier indiquant un mauvais résultat.

De plus, nous évaluons l'approche sur une tâche d'appariement de voix avec le corpus *Mass Effect 3* en utilisant le système proposé dans (Gresse *et al.*, 2019). Ici, nous testons la capacité de faire une distinction significative entre les paires *cible* (*i.e.* paire de personnages identique en anglais-français) et *non-cible* (*i.e.* paire de personnages différente en anglais-français) lorsque nous entraînons le modèle de similarité avec les p-vecteurs.

4 Résultats

Nous utilisons différentes valeurs pour la température de distillation $T \in [1..5]$, les meilleurs résultats étant observés avec $T = 4$ en moyenne ($T = 1$ revient à apprendre sans distillation). De plus, nous vérifions les différentes valeurs dans la plage $[0, 1]$ pour le paramètre d'imitation λ : nous obtenons les meilleurs résultats en utilisant $\lambda = 0,3$ lors de la moyenne sur A, B, C et D.

4.1 Analyse par clustering

Le tableau 1 présente les résultats, en termes de F-mesure, de l'analyse par clustering utilisant avec la représentation p -vecteur. Nous observons que les p-vecteurs ont des scores de F-mesure bien meilleurs que les x-vecteurs (baseline dans ce travail), ce qui n'est pas surprenant puisque les x-vecteurs sont conçus pour se concentrer sur les identités vocales des voix des acteurs anglais et français, plus que sur leur personnage/rôle. Nous observons des résultats relativement bons, jusqu'à $0,78$ dans le meilleur des cas, ce qui indique la capacité des p-vecteurs à décrire automatiquement des personnages/rôles inconnus. En ce qui concerne le cas C, nous émettons l'hypothèse que les faibles scores de F-mesure peuvent résulter de la similitude inhérente entre les personnages – tous sont des soldats masculins – impliqués dans ce test particulier. Étonnamment, le système p-vecteur sans distillation fonctionne mieux dans ce cas spécifique.

	A	B	C	D
baseline (x-vecteur)	0,54	0,52	0,36	0,71
p-vecteur (sans distillation)	0,66	0,72	**0,59**	0,66
p-vecteur + distillation	**0,78**	**0,78**	0,40	**0,77**

TABLE 1 – F-mesures obtenues pour l'analyse par clustering sur les donnés de test.

La figure 2 illustre une projection à 2 dimensions de l'espace des p-vecteurs grâce à l'algorithme t-SNE. Sans surprise, nous voyons une distinction claire entre les personnages masculins et féminins dans les cas A, B et D (C ne contient que des soldats masculins). Les personnages de même sexe sont également correctement séparés. En considérant D, nous observons que chaque voix d'acteur des deux personnages *Hackett* (bleu) et *Illusive Man* (orange) ont une identité vocale forte, ce qui pourrait faciliter l'analyse par clustering et expliquer le score de F-mesure élevé inattendu (0.71) avec le système de référence x-vecteur.

4.2 Tâche de similarité

Nous évaluons également l'approche p-vecteur avec le système de similarité de voix dans le tableau 2. Les résultats sont présentés en termes d'exactitude et de test de Student (t-test). Le test statistique confirme la différence significative entre les scores de similitude des paires *cible* et *non-cible* puisque toutes les p-valeurs associées sont sous le seuil de rejet. En moyenne, la représentation p-vecteur surpasse le système de référence x-vecteur sur la tâche de similarité, avec une précision moyenne de $57\ \%$ et un t-score moyen de de 44,79 sur les quatre cas. De plus, nous constatons des variations plus faibles entre les différents cas de test. Compte tenu de la difficulté de cette tâche, nous pensons qu'ils constituent une preuve solide que les p-vecteurs contiennent une information personnage/rôle.

5 Conclusion

Dans cet article, nous avons tout d'abord proposé, pour le casting vocal, un nouvel espace de représentation appelé p-vecteur obtenu à partir d'une couche d'embeddings d'un réseau de neurones profond. Afin de pallier le problème de limitation des données, nous avons proposé une approche par distillation de la connaissance pour améliorer la représentation personnage/rôle. Nous avons observé une amélioration substantielle des résultats au travers de cette représentation p-vecteur, en comparaison de l'approche classique x-vecteur. Ces résultats démontrent que les p-vecteurs contiennent des informations dédiées à la dimension personnage/rôle.

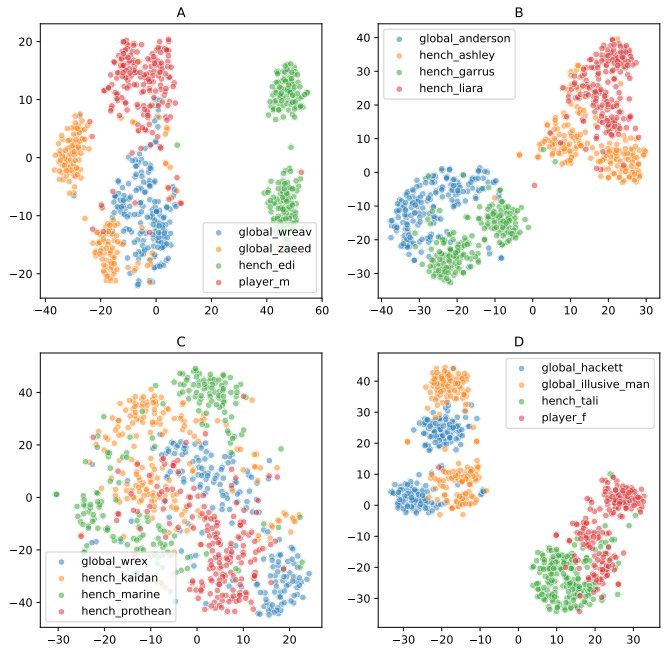


FIGURE 2 – Représentation dans l'espace des p-vecteurs pour chaque personnage dans A, B, C et D.

		exactitude	t-score
baseline x-vecteur	A	0,60	64,58
	B	0,52	20,63
	C	0,54	26,86
	D	0,49	-6,19
	moyenne	0,54	26,47
p-vecteur (sans distillation)	A	0,58	53,82
	B	0,54	20,70
	C	**0,57**	**49,86**
	D	0,54	23,34
	moyenne	0,55	36,93
p-vecteur + distillation	A	**0,63**	**80,00**
	B	**0,55**	**36,46**
	C	0,55	28,33
	D	**0,55**	**34,24**
	moyenne	0,57	44,79

TABLE 2 – Performance sur la tâche d'appariement de voix sur le corpus de test. L'exactitude sur la validation est généralemnt en dessous de 85 %.

Les paramètres T et λ, pour lesquels nous avons obtenu les meilleurs résultats en moyenne, offrent selon nous des perspectives d'analyses intéressantes. En particulier la pondération λ entre *soft-* et *hard-targets*, qui pourrait aider à mieux cerner la spécificité vocale des personnages de notre corpus par comparaison à un modèle de voix générique.

En raison des limites de notre corpus et malgré le protocole rigoureux que nous avons conçu, une certaine prudence doit être prise. Les résultats doivent être confirmés sur un corpus plus grand, avec plus de personnages, avant d'être capable de pouvoir généraliser les observations, par exemple à une autre culture. Le cadre maître-élève, pour être plus efficace, pourrait aussi être étendu sur de plus grands ensembles de données d'apprentissage, avec de nombreuses étiquettes de personnages et plusieurs acteurs par étiquette. De plus, les p-vecteurs permettent d'initier de nouvelles recherches sur les questions d'explicabilité, notamment dans le cadre des choix des directeurs artistiques. Nous souhaitons confronter les p-vecteurs à une simple décision binaire pour observer l'impact potentiel d'une caractéristique particulière sur la dimension du personnage. Les travaux futurs remplaceront le système de similitude, qui fait la distinction entre les paires de caractères identiques et différents, avec des caractéristiques explicatives (par exemple, le genre, la qualité de la voix, le timbre, la prosodie...).

Références

ASAMI T., MASUMURA R., YAMAGUCHI Y., MASATAKI H. & AONO Y. (2017). Domain adaptation of dnn acoustic models using knowledge distillation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

BENGIO Y., COURVILLE A. & VINCENT P. (2013). Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 1798–1828.

CHOLLET F. *et al.* (2015). Keras.

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. In *INTERSPEECH*.

GLOROT X. & BENGIO Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics*.

GRESSE A., QUILLOT M., DUFOUR R., LABATUT V. & BONASTRE J.-F. (2019). Similarity metric based on siamese neural networks for voice casting. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

GRESSE A., ROUVIER M., DUFOUR R., LABATUT V. & BONASTRE J.-F. (2017). Acoustic pairing of original and dubbed voices in the context of video game localization. In *INTERSPEECH*.

HINTON G., VINYALS O. & DEAN J. (2015). Distilling the knowledge in a neural network.

JOY N. M., KOTHINTI S. R., UMESH S. & ABRAHAM B. (2017). Generalized distillation framework for speaker normalization. In *INTERSPEECH*.

LI J., SELTZER M. L., WANG X., ZHAO R. & GONG Y. (2017). Large-scale domain adaptation via teacher-student learning.

LOPEZ-PAZ D., BOTTOU L., SCHÖLKOPF B. & VAPNIK V. (2016). Unifying distillation and privileged information. In *International Conference on Learning Representations*.

MARKOV K. & MATSUI T. (2016). Robust speech recognition using generalized distillation framework. In *INTERSPEECH*.

OBIN N. & ROEBEL A. (2016). Similarity search of acted voices for automatic voice casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**, 1642–1651.

OBIN N., ROEBEL A. & BACHMAN G. (2014). On automatic voice casting for expressive speech : Speaker recognition vs. speech classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 ASRU*.

PRICE R., ISO K.-I. & SHINODA K. (2016). Wise teachers train better dnn acoustic models. *EURASIP Journal on Audio, Speech, and Music Processing*, **2016**.

SNYDER D., GARCIA-ROMERO D., POVEY D. & KHUDANPUR S. (2017). Deep neural network embeddings for text-independent speaker verification. In *INTERSPEECH*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embeddings for speaker recognition. In *ICASSP* : IEEE.

SNYDER D., GHAHREMANI P., POVEY D., GARCIA-ROMERO D., CARMIEL Y. & KHUDANPUR S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Spoken Language Technology Workshop (SLT)* : IEEE.

VAPNIK V. & IZMAILOV R. (2015). Learning using privileged information : similarity control and knowledge transfer. *Journal of machine learning research*, **16**, 2023–2049.

VARIANI E., LEI X., MCDERMOTT E., MORENO I. L. & GONZALEZ-DOMINGUEZ J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP*.

WATANABE S., HORI T., LE ROUX J. & HERSHEY J. R. (2017). Student-teacher network learning with enhanced features. In *Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Lénition et fortition des occlusives en coda finale dans deux langues romanes : le français et le roumain

Mathilde Hutin[1], Adèle Jatteau[2], Ioana Vasilescu[1], Lori Lamel[1], Martine Adda-Decker[1,3]
(1) Université Paris-Saclay, CNRS, LIMSI, Bât. 507, rue du Belvédère, 91405 Orsay, France
(2) Université de Lille, CNRS, UMR 8163, STL, Lille, France
(3) Université Paris 3 Sorbonne Nouvelle, CNRS, UMR 7018, LPP, 19 rue des Bernardins, 75005 Paris, France
{mathilde.hutin, ioana.vasilescu, lori.lamel, martine.adda}@limsi.fr,
adele.jatteau@univ-lille.fr

RÉSUMÉ

L'exploration automatisée de grands corpus permet d'analyser plus finement la relation entre motifs de variation phonétique synchronique et changements diachroniques : les erreurs dans les transcriptions automatiques sont riches d'enseignements sur la variation contextuelle en parole continue et sur les possibles mutations systémiques sur le point d'apparaître. Dès lors, il est intéressant de se pencher sur des phénomènes phonologiques largement attestés dans les langues en diachronie comme en synchronie pour établir leur émergence ou non dans des langues qui n'y sont pas encore sujettes. La présente étude propose donc d'utiliser l'alignement forcé avec variantes de prononciation pour observer les alternances de voisement en coda finale de mot dans deux langues romanes : le français et le roumain. Il sera mis en évidence, notamment, que voisement et dévoisement non-canoniques des codas françaises comme roumaines ne sont pas le fruit du hasard mais bien des instances de dévoisement final et d'assimilation régressive de trait laryngal, qu'il s'agisse de voisement ou de non-voisement.

ABSTRACT

Lenition and fortition of word-final stops in two Romance languages: French and Romanian.
Automatic transcription and speech-text alignment of large corpora enables a fine investigation of the relationship between synchronic phonetic variation and diachronic change: errors in automatic transcripts can provide valuable insights about indicative possible systemic change that is in the process of appearing. It ensues that such automatic processing can be useful to investigate cross-linguistic attested phonological phenomena to establish whether or not they are occurring in languages that are not yet subject to them. The present study proposes to use forced alignment to explore voicing alternation in word-final position in two Romance languages: French and Romanian. It will be shown that non-canonical voicing and devoicing of French and Romanian coda stops are not random but are instances of final devoicing and regressive assimilation of voicing and voicelessness.

MOTS-CLÉS : Grands corpus, alignement automatique, alignement forcé, lénition, fortition, voisement, dévoisement, français, roumain

KEYWORDS: Large corpora, automatic alignment, forced alignment, lenition, fortition, voicing, devoicing, French, Romanian

Volume 1 : Journées d'Études sur la Parole, pages 289–298.
HAL : hal-02798551.

1 Introduction

Les études sur la variation phonétique et sur la relation entre motifs de variation synchronique et changements diachroniques connaissent récemment un nouvel essor grâce à l'exploration automatisée de corpus toujours plus grands. D'autres études ont montré que les erreurs dans les transcriptions automatiques sont riches d'enseignements sur la variation contextuelle en parole continue, et plus généralement sur les possibles mutations systémiques sur le point d'apparaître (Adda-Decker, 2006 ; Ohala, 1997). Dans ce cadre, il est intéressant de se pencher sur des phénomènes phonologiques largement attestés dans les langues du monde en diachronie comme en synchronie pour établir leur émergence ou non dans des langues qui n'y sont pas encore sujettes.

Les processus de lénition et de fortition figurent parmi ces phénomènes. Un segment est considéré comme subissant une lénition si sa transformation suit un chemin aboutissant à sa syncope pure et simple ; un segment est considéré comme subissant une fortition s'il suit le chemin inverse (bien que cette définition soit débattue ; Honeybone, 2008). On sait grâce à l'observation du changement diachronique dans les langues romanes (Brandão de Carvalho, 2008) qu'un segment non-voisé devra d'abord devenir voisé, par exemple, avant de disparaître, et qu'ainsi le premier est « plus fort » que le second. Il s'ensuit qu'un segment non-voisé qui devient voisé est un segment lénifié, tandis qu'un segment voisé qui devient non-voisé est un segment fortifié. En synchronie, de tels phénomènes de lénition et fortition consonantiques peuvent avoir lieu à divers degrés dans les langues romanes occidentales (Ryant & Liberman, 2016 ; Vasilescu & al., 2018 pour l'espagnol ; Hualde & Prieto, 2014 pour l'espagnol et le catalan ; Hualde & Nadeu, 2011 pour l'italien ; Jatteau et al. 2019a,b,c pour le français). Pour les langues romanes orientales en revanche, la situation est plus délicate : Chitoran et al. (2015), observant la lénition en roumain chez 8 locuteurs natifs, ne fait état que de quelques rares cas d'affaiblissement consonantique ; Niculescu et al. (soumis), étudiant lénition et fortition dans des grands corpus du roumain, ne montre que peu de variation, si ce n'est en coda, mais n'entre pas davantage dans le détail. La position finale de mot, désormais coda, est en effet, à travers les langues du monde, connue pour être une position sensible aux phénomènes d'assimilation ou de neutralisation (Keating et al. 1983 ; Blevins, 2006 ; Myers, 2012), et c'est pourquoi elle est au centre de cette étude.

Le français et le roumain sont toutes deux des langues romanes, la première appartenant à la branche occidentale, la seconde à la branche orientale. Ces deux langues sont notables, dans leur famille, pour permettre à la totalité de leurs occlusives, d'apparaître en coda, y compris en toute fin de mot, comme le montrent les bandes noires dans les tableaux 1.a et 1.b.

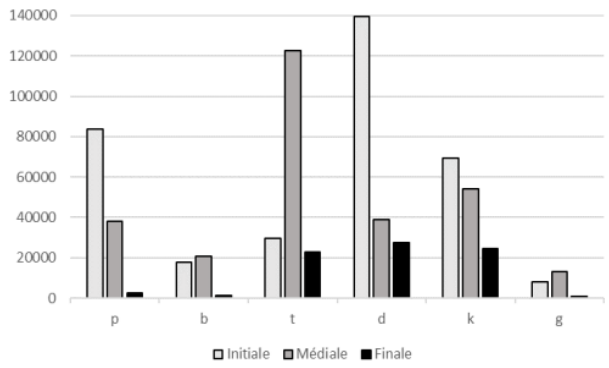


FIGURE 1.a : Répartition des occlusives du français selon leur position dans le mot (Adda-Decker, 2019 ; Vasilescu et al., soumis)

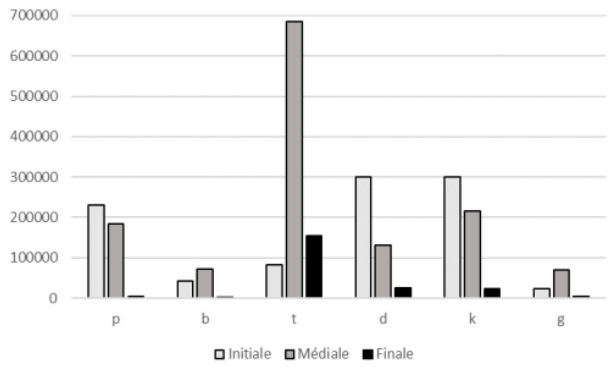


FIGURE 1.b : Répartition des occlusives du roumain selon leur position dans le mot (Adda-Decker, 2019 ; Vasilescu et al., soumis)

De plus, français (1) comme roumain (2) ont une opposition phonologique de voisement, c'est-à-dire que les oppositions p~b, t~d, k~g… donnent lieu à des paires minimales.

(1) Français : /pu/, *pou* ~ /bu/, *boue* ; /tu/, *tout* ~ /du/, *doux* ; /ku/, *cou* ~ /gu/, *goût*

(2) Roumain : /palə/, *lame* ~ /balə/, *bave* ; /talə/, *rassemblement* ~ /dalə/, *dalle* ; /kalə/, *cale* ~ /galə/, *gala*

Dans la mesure où cette opposition est canoniquement maintenue en coda, ces deux langues sont idéales pour observer les alternances de voisement dans cette position.

La présente étude se propose donc d'observer les alternances de voisement en coda dans deux langues romanes : le français et le roumain. Dans la Section 2, nous présenterons les données et la méthode utilisées. La Section 3 sera consacrée aux phénomènes de fortition, la Section 4 aux phénomènes de lénition. La Section 5 conclut et propose un regard critique sur les résultats.

2 Données et méthode

Pour examiner la variation de voisement en coda, nous proposons ici une étude des occlusives non-voisées et voisées dans des grands corpus du français et du roumain. Des phénomènes aussi spécifiques et multifactoriels que la variation accidentelle de voisement peuvent être observés de façon plus fiable grâce à de très grands corpus (Coleman et al. 2016). Outre l'observation à grande échelle qui permet des résultats statistiquement significatifs, l'avantage des grands corpus repose sur le fait que la parole y est plus naturelle que dans des enregistrements opérés lors d'expériences ou d'enquêtes car récoltée à partir de conversations réelles.

Pour ce qui est du français, nous avons utilisé trois grands corpus transcrits manuellement. Le corpus ESTER (Galliano et al. 2005) comprend originellement 80 heures de discours semi-préparé récoltées en 2003, 2005 et 2009, mais nous l'avons filtré pour enlever les données en français non-métropolitain (RFI et RTM) et ne retenir que les quelque 40 heures en français standard. Le corpus bipartite ETAPE 1 et 2 (Gravier et al. 2012) contient 13,5 heures de radio et 29 heures de télévision en français, notamment des débats et des conversations, récoltées dans les années 2010. Enfin, NCCFr (Torreira et al. 2010) contient 31 heures d'interaction en face-à-face, spontanée, entre amis, enregistrées entre 2007 et 2008.

Pour ce qui est du roumain, nous avons utilisé le corpus bipartite Quaero qui illustre la variante standard (dialecte daco-roumain du sud de la Roumanie). Il consiste en 3,5 heures de parole journalistique et 3,5 heures de débats et d'interviews du début des années 2010. La première partie a été récoltée sur des chaînes de télévision et des émissions radiophoniques roumaines provenant essentiellement des antennes radio RFI Journal et RRA (Radio România Actualităţi) et de l'agence de presse Euranet, et est représentative d'un discours plutôt préparé, voire lu. La deuxième partie consiste en débats et interviews récoltés sur la chaîne de télévision nationale roumaine Antena 3 et est plutôt représentative d'un discours spontané, non-préparé. Les émissions contenant trop de chevauchement de parole ou de bruit de fond ont été enlevées des données.

Comme ces données ne sauraient être segmentées à la main en raison de leur quantité, elles ont été traitées suivant la méthode décrite dans Gauvain et al. (2002, 2005) et Hallé & Adda-Decker (2007, 2011). Deux systèmes de reconnaissance automatique de la parole, un pour le français et un pour le roumain, ont été utilisés pour forcer un alignement pour lequel des variantes de prononciation ont été introduites pour les codas voisées et non-voisées (cf. Jatteau et al, 2019a,c pour le français ; Vasilescu et al. 2014 pour le roumain). Les systèmes ont été autorisés à sélectionner la coda dite

canonique (celle qu'on trouve dans la forme sous-jacente) ou sa variante altérée si la réalisation acoustique correspondait mieux à cette dernière. Par exemple /p/ peut être étiqueté [p] ou [b], et /b/ [b] ou [p]. Ainsi, selon ce que le système estimait correspondre le mieux à l'audio, un mot français comme *soupe* pouvait être aligné avec les transcriptions [sup], [supə], [sub] ou [subə] et un mot roumain comme *grup* avec les transcriptions [grup] ou [grub] ; de même, un mot français comme *tube* pouvait être aligné avec les transcriptions [tyb], [tybə], [typ] ou [typə] et un mot roumain comme *dialog* avec les transcriptions [dialog] ou [dialok].

Dans la mesure où il a été montré que schwa avait tendance à bloquer les effets d'adjacence (Hutin et al., 2020, soumis), nous avons décidé de ne pas conserver les items réalisés avec un schwa final en français dans nos analyses. Cela permet aussi une comparaison plus fiable avec le roumain, où il n'y a pas de schwa final optionnel. De plus, dans les 7 heures de parole en roumain, 37 items étaient considérés comme étant suivis du mot *-ul*, qui est en fait un morphème ici mal segmenté pour des raisons graphiques après les sigles (ex. *SMURD-ul*, « le SAMU » segmenté [smurd + ul] et non [smurdul] à cause du tiret). Ces 37 items ont été enlevés de la base de données finale.

Ainsi, les corpus du français utilisés dans la présente analyse comprennent 58911 occlusives en coda et les corpus du roumain 4529. Leur répartition est donnée dans le Tableau 1.

	p	t	k	b	d	g	Total
Français	3178	23113	21701	2024	7827	1068	58911
Roumain	110	3288	486	60	487	98	4529

TABLEAU 1. Répartition des occlusives en coda en français et en roumain.

On constate entre autres que /t/ est sur-représenté dans les deux langues. On remarque aussi que les données du roumain sont beaucoup moins nombreuses que celles du français, ce qui signifie que les analyses pour cette langue risquent de ne pas toujours être significatives.

3 Phénomènes de fortition

Le dévoisement en position finale de mot peut être considéré comme une instance de fortition. Elle peut typiquement avoir deux manifestations : l'assimilation régressive de non-voisement due à un segment non-voisé dans l'environnement immédiat (dans le cas des codas, à droite), ou le dévoisement en fin de domaine prosodique, attendu notamment en fin d'énoncé devant pause (Myers 2012, Jatteau et al. 2019b). Dans nos données, 22,29% des occlusives voisées /b, d, g/ sont réalisées dévoisées en français et 51,78% en roumain. De telles proportions de dévoisement méritent une étude plus poussée.

3.1 Dévoisement final *vs* assimilation de non-voisement

Ce que nous entendons par dévoisement final ici consiste en la réalisation non-voisée d'une obstruante canoniquement voisée lorsque celle-ci apparaît devant pause ; l'assimilation de non-voisement consiste en la réalisation non-voisée d'une obstruante canoniquement voisée lorsqu'elle est dans l'environnement immédiat d'une autre obstruante non-voisée. Il a été démontré que le français a une tendance variable au dévoisement des obstruantes devant pause (Jatteau et al. 2019a,b,c) comme à l'assimilation de non-voisement (Snoeren et al., 2006) : étant donnée la proportion de dévoisement des codas roumaines, il est intéressant de voir si ces deux phénomènes pourraient avoir lieu aussi en roumain.

Pour démontrer qu'il y a bien une tendance au dévoisement final, il faut montrer qu'il y a significativement plus de codas dévoisées devant pause que dans tout autre contexte ; pour démontrer qu'il y a bien de l'assimilation de trait laryngal, il faut montrer qu'il y a significativement plus de codas dévoisées devant obstruante non-voisée. Pour ce faire, les données ont été réparties en cinq catégories, selon que le mot contenant la coda d'intérêt était suivi d'une pause ou d'un autre mot commençant par une voyelle, une sonante, une obstruante voisée ou une obstruante non-voisée.

Les Figures 2.a et 2.b présentent les proportions de codas non-canoniquement dévoisées selon leur contexte de droite en français et en roumain respectivement.

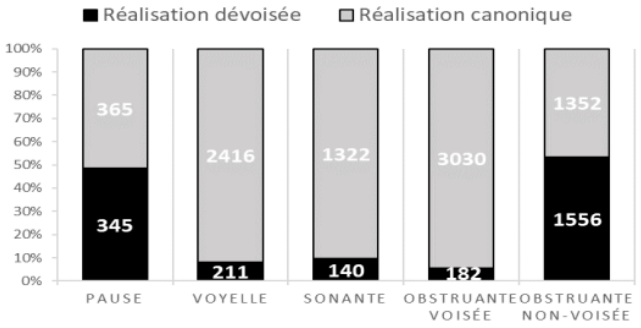


FIGURE 2.a : Proportion de /b, d, g/ en fin de mot réalisés dévoisés selon le contexte de droite en français.

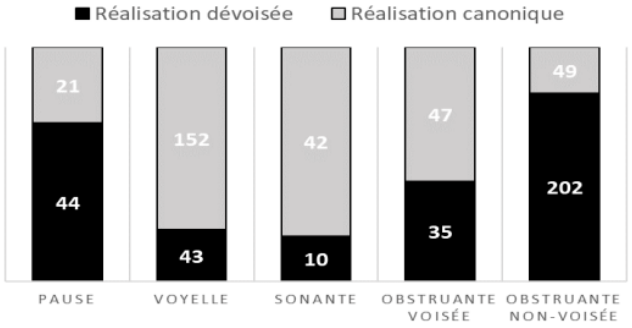


FIGURE 2.b : Proportion de /b, d, g/ en fin de mot réalisés dévoisés selon le contexte de droite en roumain.

Il est intéressant de constater que le français et le roumain montrent les mêmes tendances. Dans les deux langues, les occlusives voisées /b, d, g/ ont tendance à être largement dévoisées devant pause (48,59% en français, 67,69% en roumain) et devant obstruante non-voisée (53,51% en français et 80,48% en roumain). Inversement, dans les deux langues, le dévoisement est en moins grande proportion devant voyelle (8,03% en français, 22,05% en roumain), sonante (9,58% en français, 19,23% en roumain) et obstruante voisée (5,67% en français et 42,68% en roumain), avec néanmoins des différences de proportions qui pourront faire l'objet d'investigations ultérieures.

Les résultats semblent moins nets pour le roumain, où on trouve des proportions importantes de dévoisement dans des contextes où ce dernier n'est pas du tout attendu – ce qui est peut-être dû à la présence avant la coda d'une sonante, voisée par défaut, qui propagerait son trait laryngal à droite (ex. *scurt*, « court » ; *presi**m**t*, « pressentiment » ; *mome**n**t*, « moment ») ou encore à l'accent lexical, qui n'a pas été pris en compte ici. Cependant, les différences de voisement selon le contexte de droite sont significatives en roumain (χ^2=183.19, p<.0001) comme en français (χ^2=2876.7, p<.0001), ce qui nous permet de conclure que ces deux langues sont sujettes à dévoisement final proprement dit et à assimilation de non-voisement.

3.2 L'effet du lieu d'articulation

La littérature existante pointe du doigt les facteurs phonétiques en jeu dans le dévoisement final. Entre autres, comme il est plus difficile de maintenir le différentiel de pression nécessaire à la vibration des plis vocaux dans un tractus vocal plus court (Ohala, 1997), on suppose que les vélaires auront davantage tendance à être dévoisées.

Les Figures 3.a et 3.b montrent les proportions d'occlusives canoniquement voisées qui se trouvent dévoisées devant pause (cas de dévoisement final) selon leur lieu d'articulation.

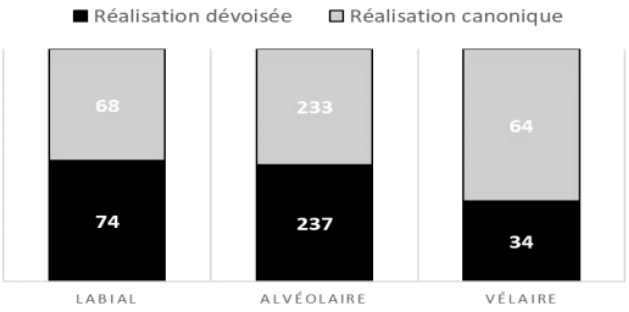


FIGURE 3.a : Proportion de réalisation dévoisée de /b, d, g/ devant pause en français

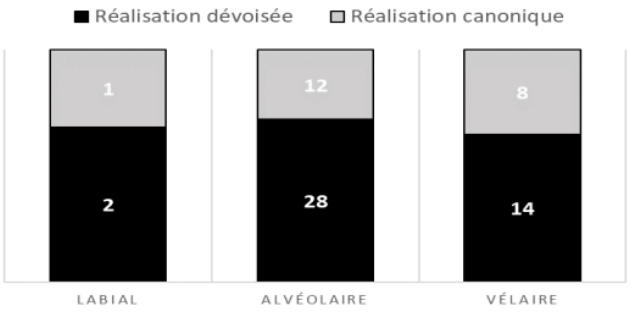


FIGURE 3.b : Proportion de réalisation dévoisée de /b, d, g/ devant pause en roumain

Le dévoisement final en roumain semble toucher davantage les alvéolaires (70,00%) que les labiales (66,67%) et les vélaires (63,64%). La tendance attendue des vélaires à favoriser le dévoisement est ici compromise mais la différence n'est pas significative (χ^2=0.26, p=0.88). Pour le français, labiales et alvéolaires sont également plus souvent dévoisées (52,11% et 50,43% respectivement) que les vélaires (34,69%). L'hypothèse que les vélaires devraient dévoiser plus spontanément que les autres occlusives est mise à mal cette fois-ci de façon significative (χ^2=8.91, p=.01).

Les Figures 4.a et 4.b montrent les proportions d'occlusives canoniquement voisées qui se trouvent dévoisées devant obstruante non-voisée (cas d'assimilation de non-voisement) selon leur lieu d'articulation.

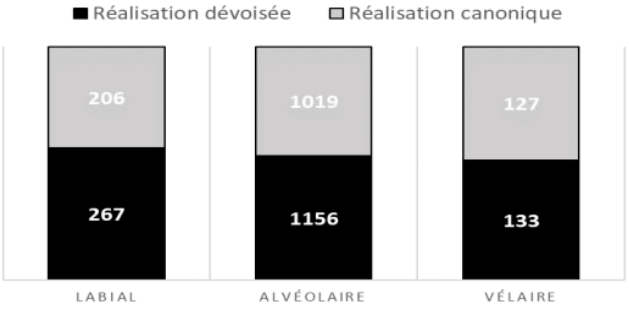


FIGURE 4.a : Proportion de réalisation dévoisée de /b, d, g/ devant obstruante non-voisée en français

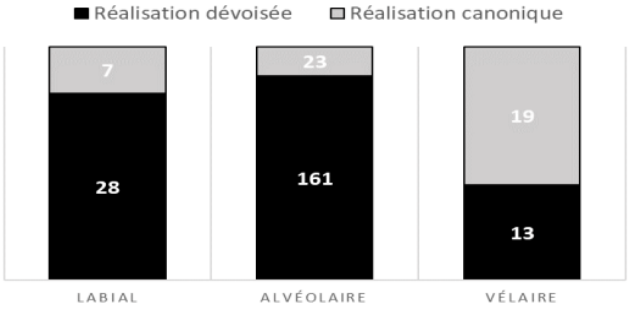


FIGURE 4.b : Proportion de réalisation dévoisée de /b, d, g/ devant obstruante non-voisée en roumain

Dans le cas de l'assimilation régressive de non-voisement, c'est-à-dire quand les codas /b, d, g/ sont réalisées dévoisées devant une obstruante non-voisée, on constate qu'en roumain, le dévoisement touche en priorité, comme pour le dévoisement final, les alvéolaires (87,50%) puis les labiales (80,00%) et en dernier lieu les vélaires (40,63%). Cette fois cependant, les résultats sont statistiquement significatifs (χ^2=38.13, p<.0001). En français, là aussi, les tendances sont les mêmes que pour le dévoisement final : les labiales sont les plus dévoisées (56,45%), puis les alvéolaires (53,15%) et enfin les vélaires (51,15%), mais ces résultats ne sont pas significatifs (χ^2=2.33, p=.31).

L'assimilation régressive de non-voisement semble donc corroborer les tendances observées pour le dévoisement final dans chacune des langues étudiées. Cependant, les effets du point d'articulation sont peu significatifs pour le dévoisement final en roumain et pour l'assimilation de non-voisement en français. Dans le premier cas, on peut supposer que le peu d'occurrences d'occlusives labiales en coda en roumain soit en cause. Dans le second, on peut supposer qu'il n'y a simplement pas de

différence selon le lieu d'articulation. En tous cas, l'hypothèse des vélaires favorisant le dévoisement est mise à mal dans les deux langues.

4 Phénomènes de lénition

Le voisement non-canonique d'une obstruante sous-jacemment non-voisée peut être considéré comme une instance de lénition. Si le voisement final spontané, c'est-à-dire devant pause, est largement controversé, l'assimilation régressive de voisement, due à un segment voisé dans son environnement immédiat (dans le cas des codas, à droite) est largement attesté. Dans nos données, 23,87% des occlusives non-voisées /p, t, k/ sont réalisées voisées en français et 16,45% en roumain.

4.1 L'assimilation régressive de voisement

L'assimilation de voisement consiste en la réalisation voisée d'une obstruante canoniquement non-voisée lorsqu'elle est dans l'environnement immédiat d'une obstruante voisée, ou éventuellement d'une sonante, voisée par défaut. Il a été démontré que le français pouvait être sujet à assimilation de voisement (Snoeren et al. 2006, Hallé & Adda-Decker 2007, 2011) : étant donnée la proportion de voisement des codas roumaines et leur propension, par ailleurs, à l'assimilation de non-voisement, il est intéressant de voir si cette langue est également sujette à assimilation de voisement.

Pour démontrer qu'il y a bien de l'assimilation de trait laryngal, en l'occurrence de voisement, il faut montrer qu'il y a significativement plus de codas non-canoniquement voisées devant obstruante voisée ou encore sonante. Pour ce faire, les données ont été réparties en cinq catégories, selon que le mot contenant la coda d'intérêt était suivi d'une pause ou d'un autre mot commençant par une voyelle, une sonante, une obstruante voisée ou une obstruante non-voisée.

Les Figures 5.a et 5.b présentent les proportions de codas non-canoniquement voisées selon leur contexte de droite en français et en roumain respectivement.

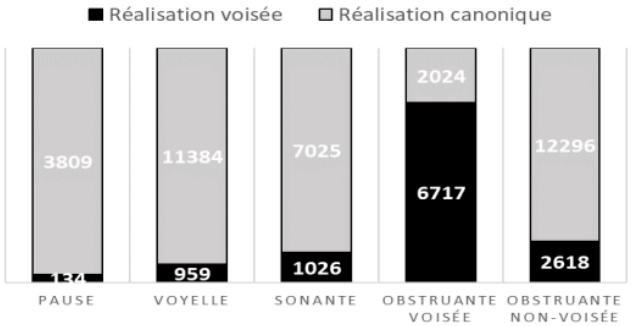


FIGURE 5.a : Proportion de /p, t, k/ en fin de mot réalisés voisés selon le contexte de droite en français.

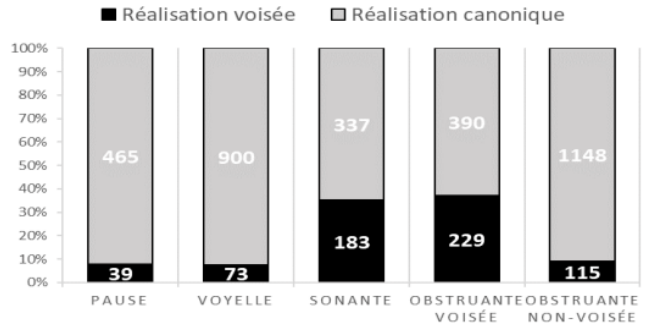


FIGURE 5.b : Proportion de /p, t, k/ en fin de mot réalisés voisés selon le contexte de droite en roumain.

Dans ces figures, on voit que le voisement de /p, t, k/ en coda de mot a lieu essentiellement devant obstruante voisée, en français (76,84%) comme en roumain (37,00%). L'effet est moins marqué en roumain, mais dans cette langue on trouve aussi du voisement devant sonante (35,19%), qui est un autre contexte favorisant l'assimilation de voisement. De plus, la différence n'est pas seulement significative pour le français (χ^2=17046, p<.0001) mais aussi pour le roumain (χ^2=456.57, p<.0001),

ce qui nous permet de conclure que le voisement non-canonique de /p, t, k/ en fin de mot est bien, dans ces deux langues, une instance d'assimilation régressive de voisement.

4.2 L'effet du lieu d'articulation

En français, les études sur l'assimilation de voisement proposent que les labiales seraient les plus propices au voisement, puis les alvéolaires et enfin seulement les vélaires (Hallé & Adda-Decker 2007). Les Figures 6.a et 6.b indiquent les proportions de voisement de /p, t, k/ dans nos corpus.

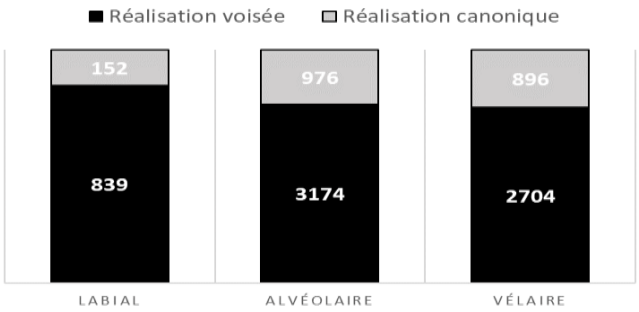


FIGURE 6.a : Proportion de réalisations voisées de /p, t, k/ devant obstruante voisée en français

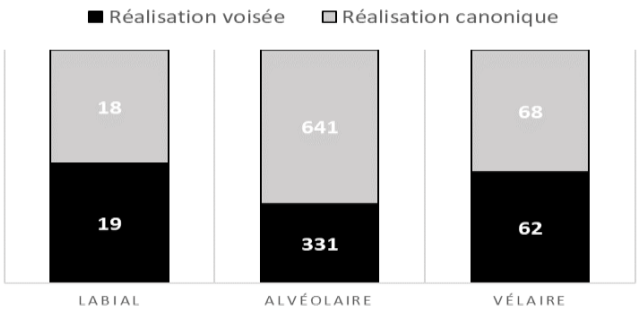


FIGURE 6.b : Proportion de réalisations voisées de /p, t, k/ devant obstruante voisée et sonante en roumain

Les figures 6a et 6b montrent que les labiales sont les plus enclines à subir une assimilation de voisement en roumain (51,35%) comme en français (84,66%). En revanche, les vélaires sont, en roumain, assez sujettes à l'assimilation de voisement (47,69%) tandis qu'en français, elles y sont le moins sujettes (75,11%). Nos données confirment donc l'échelle proposée par Hallé & Adda-Decker (2007).

Enfin, il est intéressant de voir qu'en roumain, la coda la moins couramment voisée est l'alvéolaire (34,05%), alors que sa contrepartie voisée était la plus couramment dévoisée. Il s'agit là peut-être d'un effet de la grande fréquence de /t/ en coda en roumain. Dans tous les cas, les différences sont significatives pour le roumain (χ^2=13.05, p=.002) comme pour le français (χ^2=40.42, p<.0001).

5 Conclusion et discussion

Dans cette étude, il a été mis en évidence que le voisement et le dévoisement non-canoniques des codas du français comme du roumain ne sont pas le fruit du hasard mais bien des instances de dévoisement final et d'assimilation régressive de trait laryngal, qu'il s'agisse de voisement ou de non-voisement, surtout en français. Le point d'articulation, pour sa part, a un effet significatif pour l'assimilation de voisement dans les deux langues, l'assimilation de non-voisement en roumain et le dévoisement final en français. De plus, plusieurs points intéressants sont à noter. Tout d'abord, en français comme en roumain, le dévoisement final semble défavoriser la vélaire, alors que l'état de l'art nous invitait à postuler le contraire. Par ailleurs, les deux phénomènes de dévoisement, le dévoisement final et l'assimilation de non-voisement, ne se comportent pas de la même façon selon la langue, mais montrent les mêmes tendances à l'intérieur d'une même langue – tendances qui s'opposent systématiquement à celles observées pour l'assimilation de voisement.

Malheureusement, certaines des observations proposées reposent sur des différences statistiquement non-significatives. La cause en est le plus souvent un effectif de données roumaines trop peu élevé. Il serait intéressant de confirmer les résultats avec de plus grands corpus du roumain.

Maintenant que la présence de dévoisement final et d'assimilation de trait laryngal a été montrée en roumain et en français, et l'effet du lieu d'articulation de la coda exploré, de plus amples recherches s'imposent sur le comportement des fricatives, ou encore sur divers paramètres sociolinguistiques comme le style de parole ou le sexe du locuteur.

Remerciements

Cette recherche a été en partie financée par le labex DigiCosme (projet ANR-11-LABEX-0045DIGICOSME) opéré par l'ANR dans le cadre du programme « Investissement d'Avenir » Idex Paris Saclay (ANR-11-IDEX-0003-02).

Références

ADDA-DECKER, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique des corpus oraux. Papier présenté aux Journées d'Étude sur la Parole (JEP2006), Dinard, France, 12–16 juin

ADDA-DECKER, M. (2019). Variation in Romance languages: insights from large corpora. Invited speech at *49th Linguistic Symposium on Romance Languages – LSRL 49*.

BLEVINS, J. (2006). Theoretical synopsis of Evolutionary Phonology, *Theoretical Linguistics*, vol.32, no. 2, pp.117–166

BRANDÃO DE CARVALHO, Joaquim. (2008). Western Romance. In BRANDÃO DE CARVALHO, J., SCHEER, T. & SÉGÉRAL, P. (eds) *Lenition and Fortition*. Berlin: Mouton de Gruyter.

CHITORAN, I., HUALDE, J. & NICULESCU, O. (2015). Gestural undershoot and gestural intrusion – from perceptual errors to historical sound change. *Proceedings of 2nd ERRARE Worskhop* (Sinaia, Romania)

COLEMAN, J., RENWICK, M. E.L. & TEMPLE, R.A.M. (2016). Probabilistic underspecification in nasal place assimilation. *Phonology*, 33(3), 425-458

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J.-F. & GRAVIER, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *9th European Conference on Speech Communication and Technology*

GAUVAIN, J.-L., LAMEL, L. & ADDA, G. (2002). The LIMSI broadcast news transcription system. *Speech communication* 37 (1-2), 89–108

GAUVAIN, J.-L., ADDA, G., ADDA-DECKER, M., ALLAUZEN, A., GENDNER, V., LAMEL, L. & SCHWENK, H. (2005). Where Are We in Transcribing French Broadcast News? *Proceedings of ISCA Eurospeech'05*, Lisbon, Sep 2005.

GRAVIER, G., ADDA, G., PAULSON, N., CARRÉ, M., GIRAUDEL, A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *LREC –8th International Conference on Language Resources and Evaluation*

HALLÉ, P. & ADDA-DECKER, M. (2007). Voicing assimilation in journalistic speech. *16th International Congress of Phonetic Sciences*, 2007, 493–496.

HALLÉ, P. & ADDA-DECKER, M. (2011). Voice assimilation in French obstruents: A gradient or a categorical process? *Tones and features: A festschrift for Nick Clements*, De Gruyter, 149–175

HONEYBONE, P. (2008). Lenition, weakening and consonantal strength: tracing concepts through the history of phonology. In BRANDÃO DE CARVALHO, J., SCHEER, T. & SÉGÉRAL, P. (eds) *Lenition and Fortition*. Berlin: Mouton de Gruyter.

HUALDE, J. & NADEU, M. (2011). Lenition and phonemic overlap in Rome Italian. *Phonetica*, 68, 215–242

HUALDE, J. & PRIETO, P. (2014). Lenition of intervocalic alveolar fricatives in Catalan and Spanish. *Phonetica*, 71, 109–127

HUTIN, M., JATTEAU, A., VASILESCU, I., LAMEL, L. & ADDA-DECKER, M. (Soumis). Overview of word-final schwa in Standard French and its shielding effect from fortition and lenition. *Linguistic Vanguard*.

HUTIN, M., JATTEAU, A., VASILESCU, I., LAMEL, L. & ADDA-DECKER, M. (2020). Le schwa final en français standard est-il un lubrifiant phonétique ? *Actes du Congrès Mondial de Lignuistique française, CMLF 2020*.

JATTEAU, A., VASILESCU, I., LAMEL, L. & ADDA-DECKER, M. (2019a). "Gra[f] ! Le dévoisement final dans les grands corpus de français". Paper presented at the SRPP seminar, Université Sorbonne Nouvelle, Feb. 15th 2019

JATTEAU, A., VASILESCU, I., LAMEL, L., ADDA-DECKER, M. & AUDIBERT, N. (2019b). "Gra[f]e!" Word-final devoicing of obstruents in Standard French: An acoustic study based on large corpora. *Proceedings of Interspeech*, 1726–1730. Graz, Austria.

JATTEAU, A., VASILESCU, I., LAMEL, L. & ADDA-DECKER, M. (2019c). Final devoicing of fricatives in French : Studying variation in large corpora with automatic alignment. *International Congress of Phonetic Sciences*, Melbourne, Australie.

KEATING, P., LINKER, P. & HUFFMANN, M.-K. (1983). Patterns in allophone distribution for voiced and voiceless stops. *Journal of Phonetics* 11(3). 277–290.

MYERS, S. (2012). Final devoicing: Production and perception studies. In T. BOROWSKY, S. KAWAHARA, T. SHINYA, M. SUGAHARA (Eds.), *Prosody matters: Essays in honor of Elisabeth Selkirk*, Equinox, pp.148–180.

NICULESCU, O., VASILESCU, I., CHITORAN, I., ADDA-DECKER, M. & LAMEL, L. (Soumis). Romanian obstruents still strong after all these years: Obstruent voicing and devoicing in a large corpus study of Romanian. *Labphon 17*, 2020.

OHALA, J. J. (1997). Aerodynamics of phonology. *Proceedings of the Seoul International Conference on Linguistics*. Seoul: Linguistic Society of Korea, pp. 92–97

RYANT, N. & LIBERMAN, M. (2016). Large-scale analysis of Spanish /s/-lenition using audiobooks. *Proceedings of the 22nd International Congress on Acoustics* (Buenos Aires, Argentina)

SNOEREN, N., HALLÉ, P. & SEGUI, J. (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, vol.34, pp. 241–268

TORREIRA, F., ADDA-DECKER, M. & ERNESTUS, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, vol. 52, no. 3, pp.201–212.

VASILESCU, I., HERNANDEZ, N., VIERU, B. & LAMEL, L. (2018). Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s. *Proceedings of Interspeech* (Hyderabad, India)

VASILESCU, I., VIERU, B. & LAMEL, L. (2014). Exploring Pronunciation Variants for Romanian Speech-to-text Transcription. *Proceedings of SLTU* (St. Petersburg, Russia)

VASILESCU, I., WU, Y., JATTEAU, A., ADDA-DECKER & LAMEL, L. (Soumis). Alternances de voisement et processus de lénition et de fortition : une étude automatisée de grands corpus en cinq langues romanes. *Revue TAL*.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Sur l'utilisation de la reconnaissance automatique de la parole pour l'aide au diagnostic différentiel entre la maladie de Parkinson et l'AMS

Imed Laaridh[1] Julie Mauclair[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
imed.laaridh@irit.fr, julie.mauclair@irit.fr

RÉSUMÉ

Cet article présente une étude concernant l'apport du traitement automatique de la parole dans le cadre du diagnostic différentiel entre la maladie de Parkinson et l'AMS (Atrophie Multi-Systématisée). Nous proposons des outils de reconnaissance automatique de la parole pour évaluer le potentiel d'indicateurs de la parole dysarthrique caractérisant ces deux pathologies. Dans ce cadre, un corpus de parole pathologique (projet ANR Voice4PD-MSA) a été enregistré au sein des Centres Hospitaliers Universitaires (CHU) de Toulouse et Bordeaux. Les locuteurs sont des patients atteints de stades précoces de la maladie de Parkinson et d'AMS ainsi que des locuteurs témoins.
Des mesures automatiques caractérisant la qualité de la reconnaissance automatique de la parole ainsi que la prosodie des patients ont montré un intérêt pour la caractérisation des pathologies étudiées et peuvent être considérées comme un outil potentiel pour l'aide à leur diagnostic différentiel.

ABSTRACT

On using automatic speech recognition for the differential diagnosis of Parkinson's Disease and MSA

This article presents a study regarding the contribution of automatic speech processing in the differential diagnosis between Parkinson's disease and MSA (Multi-System Atrophies). We propose to adapt automatic speech recognition tools for the evaluation and study of the dysarthric speech of these two pathologies. A corpus of pathological speech (Voice4PD-MSA) has been recorded at both Toulouse and Bordeaux university hospitals. Speakers are patients with early cases of Parkinson's disease and MSA as well as healthy controls.
Automatic measures characterizing the quality of the automatic speech recognition as well as the prosody of the patients speech have shown an interest for the characterization of the studied pathologies and can be considered as a potential tool for automatic assistance in their differential diagnosis.

MOTS-CLÉS : Dysarthrie, traitement automatique de la parole, parole pathologique, diagnostic différentiel, aide au diagnostique.

KEYWORDS: Dysarthria, automatic speech processing, pathological speech, differential diagnosis, Assistive diagnosis.

Volume 1 : Journées d'Études sur la Parole, pages 299–307.
HAL : hal-02798552.

1 Introduction

La Maladie de Parkinson (MP) est la maladie neuro-dégénérative la plus courante après la maladie d'Alzheimer, touchant environ 1,5% de la population de plus de 65 ans et environ 170 000 Français (Tison *et al.*, 1994). Son diagnostic clinique se base sur la présence d'une lenteur de mouvement associée à une manifestation motrice de rigidité, tremblement au repos et instabilité posturale (Hughes *et al.*, 1992). Il est souvent confirmé par une réponse positive à la thérapie de remplacement de la dopamine et à l'absence de signes qui suggèrent la présence d'un autre trouble parkinsonien tel que l'Atrophie Multi-Systématisée (AMS). Cependant, le diagnostic définitif ne peut être confirmé que par la découverte de certaines protéines suite à une autopsie (*post-mortem*). L'apparition d'une dysarthrie de type hypokinétique (parole monotone avec des difficultés à initier l'articulation, un débit variable et une voix rauque et soufflée), est souvent associée à la MP (Darley *et al.*, 1975). Ces troubles de la parole apparaissent au début de la MP et même pendant la période pré-symptomatique ; cette observation peut aider à un diagnostic précoce, comme l'a démontré (Harel *et al.*, 2004) dans une étude analysant rétrospectivement des enregistrements de patients atteints de la MP.

L'AMS est une maladie neuro-dégénérative rare d'étiologie inconnue. Elle se caractérise par une combinaison variable de parkinsonisme, atteinte cérébelleuse, troubles dysautonomiques et syndromes pyramidaux. Les premiers symptômes de l'AMS apparaissent généralement à partir de 60 ans (Tison *et al.*, 2000) et le pronostic vital des patients est rapidement engagé. La survie médiane des patients se situe entre 5,8 et 9,5 ans après le diagnostic (Schrag *et al.*, 2008). Le diagnostic clinique de l'AMS présente un degré de confiance limité et il nécessite une confirmation *post-mortem* (Gilman *et al.*, 2008). Comme pour la MP, les patients atteints d'AMS souffrent assez tôt d'une dysarthrie de type mixte présentant des composantes hypokinétiques, spastiques (parole lente et laborieuse avec une articulation imprécise) et ataxiques (parole brusque, ralentie, irrégulière, explosive et scandée) (Darley *et al.*, 1975).

Dans le domaine médical, plusieurs méthodes ont été étudiées pour le diagnostic de l'AMS. L'imagerie par résonance magnétique (IRM) du cerveau aide le clinicien en révélant la présence d'anomalies distinctes chez les patients atteints d'AMS (Barbagallo *et al.*, 2016). Cependant, l'IRM du cerveau peut être normale, en particulier chez les patients dont le diagnostic différentiel entre la MP et l'AMS est difficile. D'autres techniques d'imagerie telles que la tomographie permettent d'identifier des schémas métaboliques distincts entre la MP et l'AMS (Niethammer & Eidelberg, 2012). Mais cette technique est très coûteuse et n'est pas disponible en routine clinique. Au-delà de l'imagerie, plusieurs études sont en cours pour comparer les niveaux de certains marqueurs de dégénérescence axonale dans le plasma des patients atteints de la MP et d'AMS.

L'analyse de voix et de la parole est souvent utilisée pour aider le diagnostic médical des troubles affectant les patients atteints de MP ou d'AMS. L'analyse acoustique automatique, méthode non invasive par excellence, a fait l'objet de plusieurs études afin de fournir une évaluation plus objective de la pathologie et de la voix dysphonique aux orthophonistes. Ont ainsi été étudiés les prononciations de voyelles soutenues (Dibazar *et al.*, 2002), de mots (Fredouille *et al.*, 2019) ou du texte (Laaridh *et al.*, 2015) au travers de mesures telles que le rapport harmonique/bruit, l'énergie (Shama *et al.*, 2006), des mesures aérodynamiques (J. Holmes *et al.*, 2000), la durée des phonèmes, l'étranglement de la voix et l'altération de la prosodie (Ma *et al.*, 2010; Whitehill, 2010).

Peu de travaux se concentrent sur l'identification de la pathologie et la majorité porte sur l'évaluation de la qualité de la parole et la mesure automatique de l'intelligibilité. Dans (Rusz *et al.*, 2015), la comparaison entre les patients atteints de MP et d'AMS a révélé des différences significatives

dans les dimensions ataxiques, hypokinétiques et spastiques de la dysarthrie. Ces dimensions offrent une évaluation objective possible des symptômes de la dysarthrie à l'aide de plusieurs mesures acoustiques, telles que le tremblement vocal, la durée des voyelles, les variations d'intensité, le rapport harmonique/bruit. Cependant, seules les voyelles soutenues et la répétition des syllabes /pataka/ ont été utilisées comme tâches de production et l'étude ne ciblait pas particulièrement le diagnostic différentiel précoce des pathologies.

Contexte et Objectifs En stade précoce, les symptômes de la MP et de l'AMS sont très similaires, en particulier chez les patients atteints d'AMS-P où le parkinsonisme prédomine. Par conséquent, le diagnostic différentiel entre AMS et MP est souvent très difficile à établir aux premiers stades de la maladie, alors que la certitude d'un diagnostic précoce est importante pour le patient en raison du pronostic divergent et de la gravité du pronostic de l'AMS. Aucun marqueur objectif validé n'est actuellement disponible pour guider le clinicien dans ce diagnostic et le besoin de tels marqueurs est donc très élevé dans la communauté neurologique. Comme la dysarthrie est un symptôme commun et précoce dans les deux maladies, mais d'origine différente, notre approche consiste à rechercher comment la caractériser, par le biais du traitement automatique de la parole et du signal, et rechercher des différences entre les patients atteints de MP et d'AMS aux premiers stades des maladies.

Cette étude s'intègre dans un projet pluridisciplinaire financé par l'ANR (ANR VOICE4PD-MSA), auquel participent des professionnels de la médecine, de l'orthophonie (CHU de Bordeaux et Toulouse), de l'informatique et la statistique (INRIA Bordeaux, Université Toulouse 3). Le but de ce projet est de proposer des marqueurs extraits à partir de la parole pour l'assistance dans le diagnostic différentiel entre la MP et l'AMS.

Cette publication donne un état d'avancement de l'étude. L'article est organisé comme suit. La section 2 décrit le corpus construit dans le cadre de ce projet. La section 3 décrit la méthodologie utilisée dans ce travail, basée sur un alignement contraint par le texte ainsi qu'une reconnaissance automatique des phonèmes. Dans la section 4, différentes mesures sont analysées en fonction des populations étudiées et de leurs intérêts pour la caractérisation de chaque pathologie alors que la section 5 fournit quelques conclusions et directions pour de futurs travaux.

2 Corpus

Afin de répondre aux objectifs du projet ANR Voice4PD-MSA, un corpus de parole pathologique est en cours de construction. Les patients enregistrés dans ce corpus souffrent de stades précoces d'une des deux pathologies affectant le parole : la MP et l'AMS. À ce stade du projet, ont été enregistrés 39 patients (20 atteints de la MP, 14 de l'AMS) et 5 témoins sains avec comme but de réunir à terme les enregistrements de 60 patients et 30 locuteurs témoins.
Tous les enregistrements ont été réalisés dans les CHU de Toulouse et de Bordeaux sous le contrôle des phoniâtres en respectant le même protocole expérimental.

Les sessions d'enregistrements ont été réalisées dans les salles de consultation caractérisées par un environnement sonore silencieux. Un enregistreur numérique de haute qualité (de marque Zoom H4n) a été utilisé avec deux microphones, un microphone-casque et un microphone Rode NT1 cardioïde à environ 10 centimètres du locuteur et un microphone AKG C1000S cardioïde branché sur une deuxième station ; les deux stations pour l'enregistrement et l'analyse de la parole pathologique sont

des stations "état de l'art" commercialisées par la société SQ-lab. Tous les microphones réalisent leurs enregistrements en parallèle afin d'éviter au patient de répéter les mêmes tâches plusieurs fois.

Durant la session d'enregistrement, tous les locuteurs ont réalisé les mêmes tâches de production de parole : lecture de texte, parole spontanée, voyelle /a/ tenue, répétition de syllabes /pataka/ et /badaga/, lecture de logatomes. La durée d'une session d'enregistrement est d'environ 15 minutes par locuteur.

Dans ce travail, nous nous concentrons sur la tâche de lecture du texte "La chèvre de monsieur Seguin" d'environ 70 mots. Les résultats présentés dans ce travail se basent sur les enregistrements réalisés par le microphone-casque. Le tableau 1 regroupe les informations liées à ce corpus.

Population	**# de locuteurs**	**Durée moyenne ± écart-type (sec)**
Maladie de Parkinson (MP)	20	22.22 ± 3.56
Atrophie Multi-Systématisée (AMS)	14	23.86 ± 4.51
Témoins	5	22.0 ± 1.82

TABLE 1: Informations liées au corpus Voice4PD-MSA incluant le nombre de locuteurs par type de population et la durée des enregistrements en seconde par locuteur.

3 Méthodologie basée sur le traitement automatique

Afin de pouvoir identifier le potentiel et la pertinence d'indices acoustiques de manière la plus objective possible, nous avons privilégié leur extraction à partir de deux traitements automatiques :

— Le premier est un alignement forcé du signal de parole sur la suite phonétique contrainte par le texte prononcé.

— Le deuxième est une tâche de reconnaissance automatique du signal en une suite de phonèmes, sans connaissance a priori du texte prononcé et sans lexique.

Les indices acoustiques étudiés sont extraits des sorties de chacun de ces traitements ; ils peuvent ensuite être analysés indépendemment du traitement ou comparés.

Dans cette publication, nous proposons des indicateurs temporels basés sur la durée des voyelles et la vitesse d'élocution, ainsi qu'un indicateur spectral issu de la reconnaissance des phones.

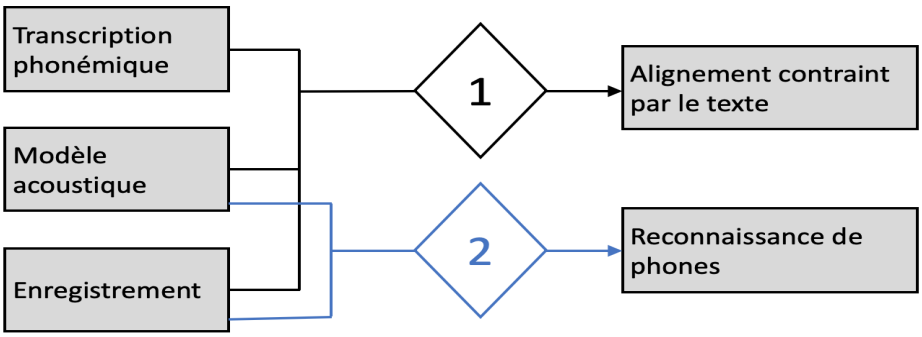


FIGURE 1: Diagramme des deux traitements automatiques utilisés, l'alignement contraint par le texte et la reconnaissance de phones.

Les deux traitements automatiques utilisés sont réalisés grâce à la boîte à outils Kaldi (Povey *et al.*, 2011). Les lexiques phonétisés des deux traitements utilisent le même jeu de phonèmes, à savoir les 37 phonèmes de la langue française. Les modèles acoustiques des phonèmes, des HMM indépendants

du contexte, sont appris sur environ 200 heures d'enregistrements radiophoniques du corpus ESTER (Galliano *et al.*, 2005). La différence entre ces deux traitements réside dans le modèle de langage utilisé comme nous le verrons par la suite.

Alignement contraint par le texte Pour réaliser l'alignement forcé des enregistrements de parole sur la suite de phonèmes a priori prononcée, le système prend en entrée la séquence de mots prononcés lors de chaque enregistrement ainsi que le lexique phonétisé. La séquence de mots est le résultat d'une transcription manuelle, après écoute, où sont introduits les ajouts, suppressions et substitutions de mots ainsi que les pauses, réalisés par les patients par rapport au texte de référence. Des variantes de prononciation pour chaque mot sont introduites dans le lexique, conformément aux règles de prononciation standard et aux règles de liaison potentielles.

L'alignement automatique repose sur une mise en correspondance à l'aide de l'algorithme de Viterbi du signal de parole avec les modèles statistiques associés. De ce processus d'alignement résultent deux segmentations temporelles des enregistrements : l'une correspond à la localisation des phonèmes avec, pour chacun, ses frontières de début et de fin dans le signal ; l'autre permet de localiser les mots et les pauses.

Reconnaissance automatique de phones Le système de reconnaissance automatique de phones utilise comme élément de base le phonème et non le traditionnel "mot". La parole produite par les locuteurs est transcrite en une suite de phonèmes. Pour le modèle de langage, il est supposé que tout phonème peut suivre n'importe quel phonème. En dehors du fait qu'il s'agit de l'hypothèse la plus simple, le choix d'un modèle unigramme équiprobable s'explique par deux raisons principales :

— Utiliser l'information acoustique uniquement lors de la reconnaissance de phones sans correction éventuelle introduite par le lexique de mots et/ou le modèle de langage associé.
— Pouvoir utiliser le même modèle sur la tâche de production de logatomes (pseudo-mots) où lexique et modèle de langage n'ont pas de sens.

Il résulte de ce processus une localisation temporelle des phones, indépendante de la transcription du texte lu, ainsi que leur labellisation.

4 Étude des indicateurs acoustiques et temporels

Comme annoncé, les différentes segmentations permettent d'extraire un grand nombre d'indices, mesures acoustiques (d'ordre spectral) ou temporelles sur les différentes populations du corpus. Pour cette première étude, nous nous sommes attardés sur les mesures suivantes :

— la qualité acoustique comparant la réalisation acoustique d'un son à celle attendue, en prenant en compte les segmentations issues de l'alignement forcé et de la reconnaissance phonémique ;
— la durée des voyelles observées ;
— les mesures globale et locale de la vitesse d'élocution.

4.1 Qualité acoustique au niveau phonème

L'alignement forcé sur le texte lu tente de localiser tous les phones au risque de pénaliser par endroit la qualité acoustique. À l'inverse, le système de reconnaissance phonétique est basé entièrement sur la reconnaissance d'une information acoustique indépendante du locuteur. Une approche pour mesurer la qualité de la réalisation acoustique de chaque enregistrement consiste à aligner temporellement les sorties de l'alignement contraint par le texte et de la reconnaissance automatique phonétique, et à comparer les phones ainsi associés trame à trame (toutes les 10 ms). Le pseudo taux de reconnaissance

phonémique, à savoir le ratio de trames bien reconnues par la reconnaissance automatique, s'apparente à un indicateur de qualité *TR* :

$$TR = 100 * \frac{\#\ trames\ bien\ reconnues\ par\ la\ reconnaissance\ automatique}{\#\ de\ trames\ de\ référence} \quad (1)$$

Les mesures de l'indicateur *TR* pour chaque individu des trois populations du corpus d'étude sont rapportées sur la figure 2 : un point représente un individu. Les trois populations sont isolées les unes des autres afin de pouvoir visualiser la moyenne et la variance pour chacune d'elles. Plus ce taux est important, mieux la parole a été reconnue par rapport à une réalisation acoustique standard.

Pour la population "témoin", le taux de reconnaissance phonémique est de 62%, un taux normal pour un système indépendant du locuteur, ce qui valide la qualité des modèles acoustiques appris et justifie l'examen des résultats sur les deux autres populations.

Pour les patients atteints d'AMS, le système a beaucoup plus de difficulté à reconnaître les phonèmes prononcés : seulement 38% de trames sont bien reconnus sur l'ensemble des enregistrements contrairement aux patients atteints de la MP pour qui le *TR* moyen est de 57%, proche du groupe "témoin". Une différence significative entre ces deux pathologies est confirmée statistiquement ($p<0.001$) (Anova à un facteur). Ce comportement reflète une information importante que nous retrouvons dans la littérature caractérisant ces deux pathologies : la dysarthrie liée à l'AMS est souvent plus sévère que celle associée à la MP. Cette sévérité est reflétée dans la qualité de la reconnaissance automatique des phonèmes et l'indicateur *TR* présente un potentiel pour son utilisation pour la différenciation entre les deux pathologies.

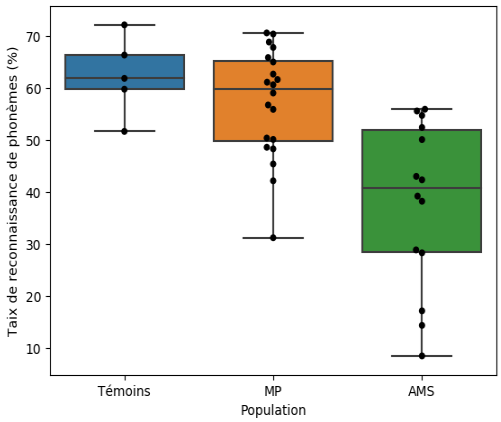


FIGURE 2: Valeur de l'indicateur *TR* (%) par individu et par population.

4.2 Durée moyenne des voyelles

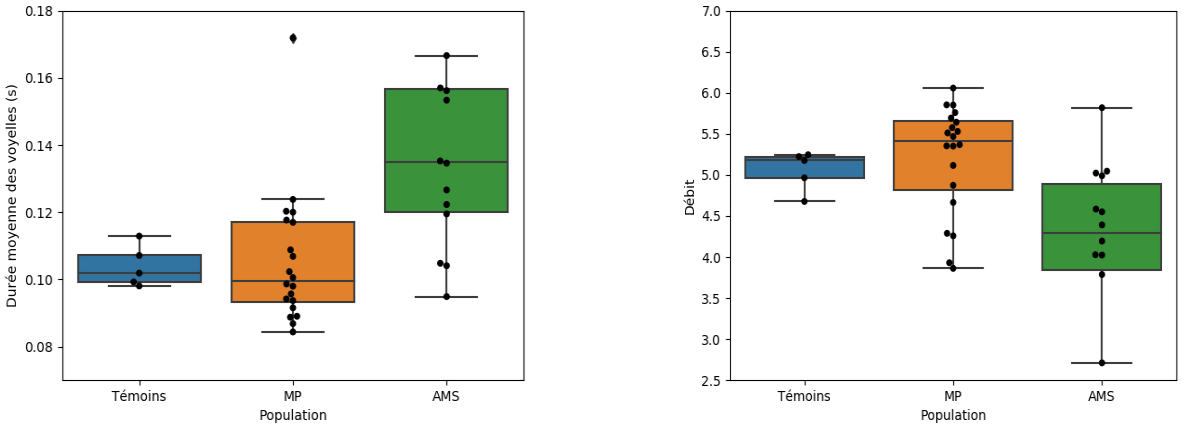
L'allongement des phonèmes, plus particulièrement celui des voyelles, est souvent associé à la dysarthrie ataxique (Darley *et al.*, 1975). Vu que la dysarthrie mixte associée à l'AMS comporte une composante ataxique, absente de la dysarthrie associée à la MP, l'étude de la durée moyenne des voyelles chez les patients des deux populations est pertinente. Ainsi, l'indicateur Dur_{voy} est obtenu en moyennant sur chaque enregistrement, les longueurs des segments reconnus comme une voyelle :

$$Dur_{voy} = \frac{Durée\ totale\ des\ voyelles}{\#\ de\ voyelles} \quad (2)$$

Cet indicateur Dur_{voy} est calculé sur chaque enregistrement à l'issue de la reconnaissance automatique de phonèmes, des travaux antérieurs de segmentation automatique du signal de parole ayant

montré que les frontières détectées par les systèmes de reconnaissance phonétique sont fiables, même si l'identification phonémique en elle même peut être erronée.

Les résultats sont rassemblés sur la figure 3. Les patients atteints d'AMS présentent un comportement distinctif, avec des prononciations de voyelles significativement plus longues que la normale (p<0.001). Ce résultat confirme la pertinence de la mesure Dur_{voy} qui reflète l'allongement des voyelles associé à la dysarthrie ataxique, comportement décrit dans la littérature et basé sur l'analyse perceptive de la parole.



(a) Durée moyenne de voyelle Dur_{voy} (s) par population

(b) Débit de parole $Débit$ par population

FIGURE 3: Dur_{voy} et $Débit$ par population à partir de la reconnaissance automatique de phonème.

4.3 Débit de parole

L'allongement de phonème observé chez les patients atteints de l'AMS agit a fortiori sur leur débit de parole. La dysarthrie ataxique dont souffrent ces patients est caractérisée par un ralentissement du débit de la parole, contrairement à la dysarthrie hypokinétique associée à la MP caractérisée par des irrégularités (voire des accélérations) du débit.
Afin d'estimer automatiquement le débit de parole, nous proposons de calculer le nombre de voyelles prononcées par seconde, noté $Débit$; c'est un estimateur proche du nombre de syllabes par seconde et de la vitesse d'élocution, par ailleurs couramment utilisé (Rouas *et al.*, 2004).

$$Débit = \frac{\#\ de\ voyelles}{Durée\ totale\ de\ lecture - (Durée\ pause\ + Durée\ respiration)} \tag{3}$$

Le paramètre $Débit$ est extrait à partir des sorties de la reconnaissance automatique de phonèmes. Pour éviter tout biais, les pauses et respirations ne sont pas prises en compte pour le calcul.

Les résultats par type de population sont rassemblés sur la figure 3 (b). Comme pour l'étude de la durée des voyelles, des résultats prometteurs s'observent : les patients atteints d'AMS présentent un débit de parole significativement plus lent que les locuteurs témoins ou atteints de la MP ($p < 0.01$). La variance élevée de ce paramètre sur chaque population peut s'interpréter de deux manières différentes : soit les individus au sein d'une même population présentent des comportements très différents l'un de l'autre (plus que probable), soit chaque individu a un comportement très irrégulier au cours du temps et le paramètre, $Débit$, moyenne sur l'ensemble du texte lu, n'est pas significatif, compte tenu de l'instabilité éventuelle d'un débit "local" .

La suite naturelle de cette étude est d'examiner le comportement d'un débit "local" de parole, calculé sur une fenêtre glissante courte (de l'ordre de 2 secondes) au cours du temps.

5 Conclusion

Ce travail vise l'étude et la recherche de marqueurs issus d'un traitement automatique de la parole, afin d'aider un diagnostic différentiel entre les deux maladies, la MP et l'AMS. Les premiers résultats mettent en évidence trois indicateurs issus du pré traitement basé sur la simple reconnaissance automatique de phonèmes, à savoir la qualité de la reconnaissance phonétique, la durée des voyelles reconnues et le débit de parole.
Des différences significatives dans le calcul de ces paramètres sont en adéquation avec les observations effectuées par le corps médical sur les problèmes de prononciation dus à l'AMS, à savoir une qualité acoustique moindre, des voyelles plus longues et un débit de parole plus lent. Ces indicateurs ne conduisent pas à une différentiation entre patients atteints de la MP et témoins non atteints.
A noter également que le calcul de ces paramètres ne demande aucune connaissance a priori qu'une transcription phonétique du texte produit par les patients ; il ne nécessite pas de travail supplémentaire d'annotation ou de transcription sur le signal.

Les dernières expériences ouvrent de nouvelles pistes : une étude du suivi de ces paramètres au cours du temps, sur un texte de longueur raisonnable, comme celui utilisé dans ce protocole, devrait conduire à préciser le comportement des malades en termes de régularité ou non, et sans doute à différencier les patients atteints de la MP des témoins. D'autres corpus devront être étudiés et il est prévu d'augmenter le nombre de patients enregistrés pour consolider ces premières observations.

Remerciements Ce travail fait partie du projet VOICE4PD-MSA (ANR-16-CE19-0010) financé par l'ANR. Nous remercions le Pr. Wassilios Meissner, le Dr. Solange Milhé de Saint Victor, le Dr. Anne Pavy Le Traon, le Pr. Virginie Woisard et les équipes des CHU de Bordeaux et de Toulouse, pour le recueil des données du corpus. Les auteurs remercient également Régine André-Obrecht pour son travail d'analyse des résultats et de relecture ainsi que Khalid Daoudi, porteur du projet.

Références

BARBAGALLO G., SIERRA-PEÑA M., NEMMI F., TRAON A. P.-L., MEISSNER W. G., RASCOL O. & PÉRAN P. (2016). Multimodal mri assessment of nigro-striatal pathway in multiple system atrophy and parkinson disease. *Movement Disorders*, **31**(3), 325–334.

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1975). *Motor speech disorders*. Philadelphia : W. B. Saunders and Co.

DIBAZAR A. A., NARAYANAN S. & BERGER T. W. (2002). Feature analysis for automatic detection of pathological speech. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, volume 1, p. 182–183 : IEEE.

FREDOUILLE C., GHIO A., LAARIDH I., LALAIN M. & WOISARD V. (2019). Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. In *Intl Congress of Phonetic Sciences (ICPHs'19)*, p. 3051–3055, Melbourne, Australia.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of Interspeech'05*, p. 1149–1152, Lisbon, Portugal.

GILMAN S., WENNING G., LOW P. A., BROOKS D., MATHIAS C., TROJANOWSKI J., WOOD N. W., COLOSIMO C., DÜRR A., FOWLER C. *et al.* (2008). Second consensus statement on the diagnosis of multiple system atrophy. *Neurology*, **71**(9), 670–676.

HAREL B., CANNIZZARO M. & SNYDER P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease : A longitudinal case study. *Brain and cognition*, **56**(1), 24–29.

HUGHES A. J., DANIEL S. E., KILFORD L. & LEES A. J. (1992). Accuracy of clinical diagnosis of idiopathic parkinson's disease : a clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, **55**(3), 181–184.

J. HOLMES R., M. OATES J., J. PHYLAND D. & J. HUGHES A. (2000). Voice characteristics in the progression of parkinson's disease. *International Journal of Language & Communication Disorders*, **35**(3), 407–418.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2015). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing*, **6**(3), 9 :1–9 :24.

MA J. K.-Y., WHITEHILL T. L. & SO S. Y.-S. (2010). Intonation contrast in cantonese speakers with hypokinetic dysarthria associated with parkinson's disease. *Journal of Speech, Language, and Hearing Research*.

NIETHAMMER M. & EIDELBERG D. (2012). Metabolic brain networks in translational neurology : concepts and applications. *Annals of neurology*, **72**(5), 635–647.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on automatic speech recognition and understanding* : IEEE Signal Processing Society.

ROUAS J.-L., FARINAS J. & PELLEGRINO F. (2004). Evaluation automatique du débit de la parole sur des données multilingues spontanées. *XXVe Journees d'Etude sur la Parole (JEP 2004), Fes, Maroc*, p. 437–440.

RUSZ J., BONNET C., KLEMPÍŘ J., TYKALOVÁ T., BABOROVÁ E., NOVOTNỲ M., RULSEH A. & RŽIČKA E. (2015). Speech disorders reflect differing pathophysiology in parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of neurology*, **262**(4), 992–1001.

SCHRAG A., WENNING G. K., QUINN N. & BEN-SHLOMO Y. (2008). Survival in multiple system atrophy. *Movement Disorders*, **23**(2), 294–296.

SHAMA K., KRISHNA A. & CHOLAYYA N. U. (2006). Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on advances in signal processing*, **2007**, 1–9.

TISON F., DARTIGUES J., DUBES L., ZUBER M., ALPEROVITCH A. & HENRY P. (1994). Prevalence of parkinson's disease in the elderly : a population study in gironde, france. *Acta neurologica scandinavica*, **90**(2), 111–115.

TISON F., YEKHLEF F., CHRYSOSTOME V. & SOURGEN C. (2000). Prevalence of multiple system atrophy. *The Lancet*, **355**(9202), 495–496.

WHITEHILL T. L. (2010). Studies of chinese speakers with dysarthria : informing theoretical models. *Folia Phoniatrica et Logopaedica*, **62**(3), 92–96.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Variation stylistique en français québécois : l'effet de l'identité de l'interlocuteur

Mélanie Lancien[1, 2]
(1) Section SLI - Université de Lausanne, CH1015 Lausanne, Suisse
(2) Université du Québec à Chicoutimi, QC G7H 2B1 Chicoutimi, Canada
melanie.lancien@unil.ch

RÉSUMÉ

Les études portant sur l'effet de la situation de communication sur la variation vocalique, notamment celles de Bradlow (2003) ou Scarborough (2007, 2013) ont démontré une adaptation du degré d'hyper-hypo articulation à l'identité de l'interlocuteur, avec par exemple une plus forte hypoarticulation (Lindblom, 1990) lorsque l'on s'adresse à un ami que lorsque l'on s'adresse à étranger. Dans cette étude, nous adaptons le protocole Diapix (Baker et Hazan, 2011) de façon à explorer la variation vocalique dans la parole dirigée à un.e conjoint.e, un expérimentateur de la même communauté linguistique, une expérimentatrice d'une autre communauté, ou soi-même. L'analyse préliminaire des productions de deux couples montre d'ores et déjà une influence de l'identité de l'interlocuteur, avec des voyelles plus courtes et plus proches du centroïde du système lors des tâches en couple que lors des tâches avec les enquêteurs ou avec soi-même.

ABSTRACT

Stylistic variation in Quebec French: the effect of the interlocutor's identity.
Studies on the effect of the communication situation on vowel variation, in particular those of Bradlow (2003) or Scarborough (2007, 2013) have demonstrated an adaptation of the degree of hyper-hypo articulation to the identity of the interlocutor, with for example stronger hypoarticulation (Lindblom, 1990) when speaking to a friend than when addressing a foreigner. In this study, we adapt the Diapix task (Baker and Hazan, 2011) in order to explore vowels' variation in speech directed to a spouse, an experimenter from the same linguistic community, an experimenter from another community, or oneself. The preliminary analysis of the productions of two couples already shows an influence of the identity of the interlocutor, with vowels shorter and closer to the centroid of the system during the task in couple than during the tasks with the investigators or alone.

MOTS-CLÉS : Phonostylistique, Sociophonétique, Français québécois, Acoustique, Variation vocalique

KEYWORDS: Phonostylistics, Sociophonetics, Quebec french, Acoustics, Vowel variation

Volume 1 : Journées d'Études sur la Parole, pages 308–316.
HAL : hal-02798553.

1 Introduction

1.1 La variation stylistique

La parole est affectée par différentes sources de variation, dont des sources extralinguistiques telles que la situation de communication ou les émotions. Les études sociolinguistiques sur la variation (notamment diatopique et diastratique) menées par Labov (1972, 2002, entre autres) postulent qu'un locuteur varie son utilisation de la langue en fonction de trois critères : 1) Les relations entre le destinateur et le destinataire (relation filiale, amicale, professionnelle…), 2) Le contexte social (entretien formel, discussion privée…), 3) Le thème du discours (opinion politique, anecdote de vacances, …). À ces critères, Eskenazi (1993) ajoutera également le taux d'intelligibilité (endroit bruité, problèmes d'audition, …) caractérisant la situation d'interaction, et la couche sociale dont sont issus les interactants (différences de classe sociale entre les interlocuteurs).

De nombreux facteurs peuvent donc influencer la réalisation d'un segment (ici nous nous concentrerons sur les voyelles) en parole continue, a fortiori dans une situation de communication spontanée. Ces modifications sont souvent produites à des fins communicationnelles, comme le démontre le modèle « Hypo-Hyper » de Lindblom (1990). Les locuteurs d'une langue vont, selon ce modèle, varier la clarté de leur élocution en fonction des informations nécessaires à l'interlocuteur pour le bon déroulement de l'interaction. Ainsi lorsque l'interlocuteur a besoin d'un maximum d'informations acoustiques pour comprendre le message le locuteur va hyperarticuler, a contrario, si l'interlocuteur peut combler le manque d'informations acoustiques grâce à d'autres informations (en général liées au contexte de communication), le locuteur va réduire son effort et hypoarticuler.

L'une des premières recherches à avoir étudié la variation induite par la condition de production, du point de vue de la phonétique expérimentale est celle de Bernard Harmegnies et Dolors Poch-Olivé (1992), qui ont montré des différences dans l'espace acoustique occupé par les voyelles, qui est plus réduit en parole spontanée qu'en parole de laboratoire. Harmegnies & Poch-Olivé (1994) ont par la suite comparé, dans une étude de cas, 6 situations de communication différentes, qu'ils regrouperont en 3 grandes familles de styles homogènes sur la base de la dispersion/réduction de l'espace vocalique du locuteur. Ces 3 familles, pouvant être identifiées comme trois phonostyles, sont 1) la lecture, 2) la conversation spontanée, 3) la description/explication d'image. Chacun de ces styles implique des besoins différents pour la réussite de la tâche de communication et il apparaît que l'extension/réduction de l'espace vocalique s'adapte à ces besoins. Adda-Decker et Lamel (1999), ou encore Rouas et collègues (2010) ont également constaté, dans des comparaisons quantitatives entre la parole préparée (journalistique), la conversation spontanée, et la parole lue, une plus forte réduction spectrale et temporelle en parole préparée qu'en parole lue et une plus grande réduction en parole conversationnelle qu'en parole préparée.

La variation décrite ci-dessus est donc due à des facteurs extra ou paralinguistiques. Lindblom propose comme hypothèse qu'elle soit étroitement liée à l'interlocuteur et au contexte dans lequel se déroule l'interaction. À ce sujet, les travaux de Bradlow et collègues (2003) montrent que les caractéristiques de l'interlocuteur influencent effectivement le degré d'hyper-hypo articulation. Dans cette étude, les auteurs mettent en évidence, pour l'anglais, le fait que les locuteurs modifient les caractéristiques spectrales des voyelles qu'ils prononcent lorsqu'ils s'adressent à des enfants ayant des difficultés linguistiques : les différences articulatoires entre les phonèmes sont volontairement exagérées pour aider l'enfant à comprendre (hyperarticulation). Ce phénomène, que la littérature retiendra sous le nom de *clear speech*, sera étudié plus en détail par la suite, notamment par Scarborough et al. (2007, 2013), qui montrent des différences importantes de l'espace

acoustique sur un plan F1-F2, en fonction de la relation interpersonnelle (Friend), du fait qu'un mot soit énoncé en situation de communication réelle ou dans sa forme de citation (Real vs Citation) et d'éventuels problèmes auditifs de l'interlocuteur (HoH).

Dans la présente recherche nous nous plaçons dans la lignée de ces travaux, et proposons d'évaluer plus précisément l'impact de la proximité sociale (Berscheid et al., 1989) entre les locuteurs sur le degré d'hyper-hypo articulation de la parole. Pour ce faire, nous avons adapté le protocole Diapix (Baker et Hazan, 2011) au français québécois et avons demandé à 10 couples de venir en chambre sourde jouer au jeu des 12 différences seuls, avec leur conjoint.e., avec un expérimentateur issu de la même communauté linguistique, et avec une expérimentatrice européenne, pour finir par la lecture des mots cibles prévus par le protocole. Avant de présenter plus amplement nos hypothèses et notre méthode, une brève présentation du système vocalique du français québécois nous paraît incontournable.

1.2 Le système vocalique du français québécois

Le français québécois possède un système vocalique très différent de celui du français dit « standard ». Plusieurs auteurs, notamment Santerre (1976), Walker (1984), et Côté (2012), ont établi différents inventaires des voyelles du français québécois. Pour notre part nous nous appuierons sur le système présenté par Santerre (1976), nous nous focaliserons ici sur les phonèmes vocaliques monophtongues et oraux. Santerre propose un ensemble de 12 voyelles orales monophtongues : /i, y, u, e, ɛ, ɜ, ø, œ, o, ɔ, a, ɑ/ (/ɜ/ étant en fait un symbole choisi pour le /ɛ:/). Ces 12 voyelles subissent toutes de profondes modifications de leur forme de surface : relâchement, diphtongaison, dévoisement, apocope, mouvements sur l'axe de l'antéro-postériorité (pour les principales). Ces transformations sont dictées par des règles phonologiques très tôt mise en évidence, et sujettes à un (relatif) consensus, qui impliquent le type de syllabe (ouverte/ fermée), le segment suivant la voyelle (consonne allongeante/non allongeante), et l'accentuation.

C'est pour pouvoir rendre compte de ces différences de surface que nous mobilisons maintenant la notion de classe de voyelle (telle que définie par Yaeger, 1975). Nous prendrons en compte trois classes de voyelles : 1) **la classe (_R)** : la voyelle est située devant une consonne allongeante (/ʁ, v, z, ʒ/), dans cette situation les voyelles sont allongées et peuvent être diphtonguées. Par exemple, la classe (ɛR) représente la voyelle réalisée dans « père » /pɛʁ/ qui peut être réalisé [pɛːʁ] ou [paɛʁ]. 2) **la classe (_K)** : la voyelle est située devant une consonne non-allongeante. Dans cette situation la voyelle est relâchée si elle est haute (/i, u, y/), peut diphtonguer si c'est un /o, ɜ, ɑ, ø/, et se déplace sur l'axe de l'antéro-postériorité dans quelques rares cas. Par exemple, la classe (iK) représente la voyelle de « pipe » /pip/ qui est généralement réalisé [pɪp]. 3) **la classe (_#)** : la voyelle est en position finale de syllabe ouverte. Dans cette position elle ne change généralement pas de timbre (à l'exception du /a/ qui se postériorise, et du /ɛ/ qui s'ouvre). Par exemple, la classe (i#) représente la voyelle de « vie » /vi/ généralement prononcé [vi].

Pour des raisons de place, nous n'entrerons pas dans les détails et ne nous attarderons pas sur les exceptions (comme par exemple le fait que /e/ dans la classe (e#) peut diphtonguer, ou que /a/ dans la classe (aR) devient /ɑ/). Pour plus de précisions et un aperçu exhaustif de la complexité du système nous invitons le lecteur à consulter Walker (1984) Dumas (1987) ou encore Cedergren and Simoneau (1985) en plus des références citées précédemment.

2 Méthodologie

2.1 Protocole expérimental

Pour cette étude nous avons choisi d'adapter le protocole Diapix (Baker & Hazan, 2011) au français québécois. Ce protocole comprend 4 jeux de 3 paires d'images ayant pour thèmes la plage, la rue, la ferme. Entre chaque paire d'images, 12 différences sont dissimulées. Les participants doivent collaborer pour trouver les différences entre leurs images sans se les montrer. Ce protocole permet d'induire de la parole spontanée en interaction tout en fixant un champ lexical. Sur les quatre jeux d'images, nous en avons sélectionné un pour chaque type d'interaction que nous souhaitions tester.

Ainsi nous avons récolté la parole de 20 locuteurs du français québécois : 10 femmes et 10 hommes, de 20 à 51 ans, originaires de la région du Saguenay–Lac-Saint-Jean dans la province du Québec. Ces locuteurs ont tous été placés en interaction avec : 1) leur conjoint (distance sociale la plus faible, noté « EnCouple » ci-après), 2) un expérimentateur de la même origine régionale (distance sociale plus forte, noté « EnqSag » ci-après), 3) une expérimentatrice française (distance sociale la plus forte dans notre expérience, noté « EnqEur » ci-après), ainsi que dans une tâche de *self-directed speech* (noté « Solo » ci-après). Après avoir joué toutes leurs parties de détection des différences les participants étaient invités à relire un ensemble de 241 mots mobilisés par le protocole. Ces mots étaient lus une fois en isolé, et une fois dans une phrase porteuse (« il a dit X deux fois »), cette tâche est notée « Lecture » ci-après.

Les données décrites ont été enregistrées en chambre sourde grâce à un Tascam HD-P2 et un micro serre-tête Shure SM-10A (dynamique, cardioïde). Les enregistrements ont été faits en stéréo de façon à ce que chaque locuteur ait son propre canal lors des étapes du protocole réalisées en interaction. Au total nous avons, au moment de l'écriture de cet article, récolté les productions de 20 locuteurs (10 couples), cependant nous ne présentons ici que les résultats obtenus sur les deux premiers couples enregistrés (2 hommes et 2 femmes âgées de 25 à 38 ans).

2.2 Traitement des données

Les données enregistrées ont été transcrites orthographiquement sous Praat (Boersma, 2002), puis alignées grâce à la version québécoise de l'outil d'alignement SPPAS (Lancien et al., à paraître). Suite à cela, un script Praat a permis d'extraire les moyennes de F1, F2, et F3, leurs valeurs aux cinq cinquièmes de chaque phone, ainsi que la durée de chaque voyelle. Les valeurs de formants extraites ont été filtrées de façon à éliminer celles qui n'étaient pas réalistes (Gendrot & Adda-Decker, 2005), ainsi que les phones précédés ou suivis d'une voyelle ou d'une semi-consonne. Grâce à ces mesures nous avons par la suite calculé la distance des phones au centroïde du système, ainsi que la distance au centroïde de la catégorie (voir Huet & Harmegnies, 2000). Ces mesures sont communément utilisées depuis les années 1990 pour décrire le degré d'hypo ou hyperarticulation des voyelles (voir par exemple Huet & Harmegnies, 2000 ; Harmegnies & Poch-Olive, 1994 ; ou Audibert et al., 2015).

2.3 Analyse des données

Pour analyser les données recueillies nous avons construit un modèle linéaire mixte (Bates et al., 2015) comprenant la classe vocalique, l'âge des participants, le sexe des participants, et la condition de production comme effets fixes, ainsi que le mot et le locuteur comme effet aléatoire. La variable dépendante était à tour de rôle : la durée, la distance du phone au centre du système, et la distance du phone au centre de sa catégorie. La durée (en secondes) n'étant pas normalement distribuée (dû aux limites physiologiques de la parole), nous avons eu recours un à logarithme pour analyser les durées vocaliques. Les mesures de formants n'ont, elles, pas été transformées. Les modèles mixtes ont été complétés par des mesures de taille d'effet (Nakagawa et Schielzeth, 2013) et des tests posthocs (Tukey HSD).[1]

3 Résultats

3.1 Données recueillies

Pour les 4 locuteurs dont nous présentons l'analyse des productions, nous avons pris en compte 27 084 occurrences des 12 voyelles présentées par Santerre (1976) ainsi que du schwa. La distribution des données dans les classes de voyelles et dans les conditions de productions est détaillée en TABLE 1 (pour des raisons de place nous n'avons pas détaillé les 13 catégories de voyelles * 3 classes pour la colonne « classe »).

Condition de production / Classe	**EnCouple**	**EnqEur**	**EnqSag**	**Lecture**	**Solo**	**Total général**
(_#)	1561	1899	1835	1937	1035	8267
(_K)	2464	3403	3250	2778	1870	13765
(_R)	978	1289	1131	1013	641	5052
Total général	**5003**	**6591**	**6216**	**5728**	**3546**	**27084**

TABLE 2 : Nombre d'occurrences de voyelles recueillies pour les deux premiers couples en fonction de la classe de voyelle et de la condition d'interaction

3.2 Résultats généraux

Pour la durée vocalique, les premiers résultats ne montrent pas d'effet de l'âge ou du sexe. En revanche un effet de la classe vocalique et de la condition de production sont constatés, ainsi qu'une interaction entre ces deux facteurs ($p<0.001$). Les mêmes résultats sont constatés pour la distance au centroïde du système. En revanche, pour la distance au centroïde de la catégorie, on observe un effet significatif de l'âge, du sexe, et de la classe vocalique ($p=<0.001$), mais pas de l'interlocuteur. Pour les 3 mesures, on remarque également un effet du locuteur et du mot ($p<0.001$).

[1] Nos résultats ne portant que sur les productions de quatre locuteurs.trices, nous n'avons, à ce stade, pas pris en compte les interactions entre facteurs (notamment entre les facteurs sociaux). Les caractéristiques sociales des enquêteurs.trices seront également prises en compte par la suite.

3.3 L'effet de la condition de production et de l'interlocuteur

En ce qui concerne la durée des phones, les résultats des tests posthocs montrent des différences significatives entre toutes les conditions de production (p=<0.002), sauf pour la paire EnqEur et EnqSag et la paire Lecture et Solo. On remarque donc que la distance sociale influence la durée puisque la parole produite lors du jeu en couple montre les durées vocaliques les plus courtes, les durées vocaliques lors des échanges avec les enquêteurs étant légèrement plus longues, les productions en *self-directed speech* encore un peu plus longues, et finalement la lecture montre les durées vocaliques les plus longes. Ces résultats sont résumés par la FIGURE 1 qui présente les moyennes et intervalles de confiance marginales prédites par le modèle.

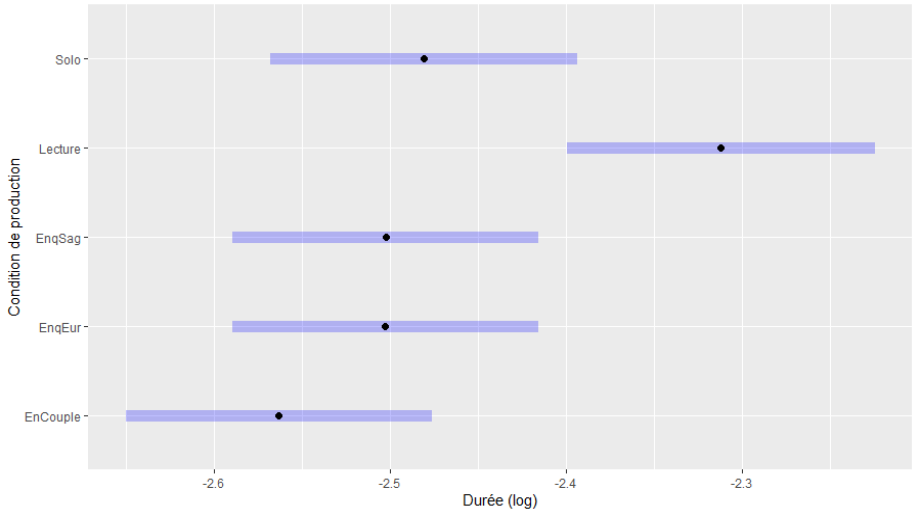


FIGURE 1 : Graphique présentant les résultats des tests posthoc sur les conditions de production pour la durée. Le point central représente la moyenne marginale prédite par le modèle et les barres horizontales les intervalles de confiance.

Pour la distance au centroïde du système, les tests posthoc montrent des différences significatives entre toutes les conditions de production (p=<0.05). On constate un effet de la condition de production avec des segments plus centralisés dans les conditions d'interaction qu'en Lecture ou en Solo, mais on constate également un effet de la distance sociale, avec des phones plus centralisés lors du jeu en couple (EnCouple) que lors du jeu avec l'enquêteur local (EnqSag), et des phones plus centralisés lors du jeu avec l'enquêteur local que lors du jeu avec l'enquêtrice française (EnqEur). De nouveau, la FIGURE 2 montre les moyennes et intervalles de confiance marginales prédites par le modèle.

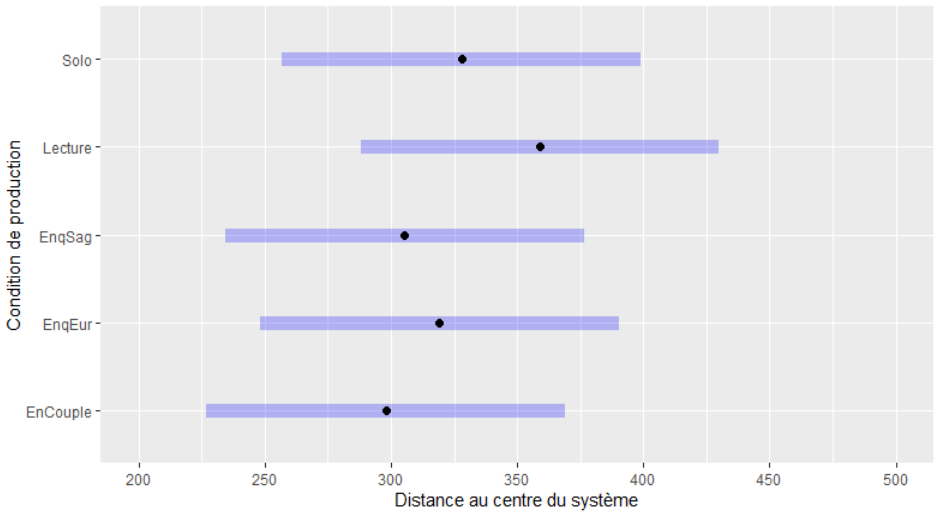


FIGURE 2 : Graphique présentant les résultats des tests posthoc sur les conditions de production pour la distance au centre du système. Le point central représente la moyenne marginale prédite par le modèle et les barres horizontales les intervalles de confiance.

Pour les mesures de durée et de distance au centroïde du système on remarque une tendance dans le positionnement des styles sur le continuum hyper-hypo : le jeu en couple montre plus d'hypoarticulation que le jeu avec un enquêteur originaire de la même communauté, qui montre lui-même plus d'hypoarticulation que le jeu avec un enquêteur européen, qui montre lui-même plus d'hypoarticulation que le jeu seul, qui montre lui-même plus d'hypoarticulation que la lecture. En somme : EnCouple < EnqSag =< EnqEur < Solo < Lecture. En revanche, les mesures de distance aux centroïdes des catégories ne montrent aucune variation significative liée à l'identité de l'interlocuteur, ou même à la situation de production de façon plus générale.

En somme nous constatons, en plus de l'effet (bien connu de la littérature) du style (lecture / parole spontanée / parole spontanée en interaction), un effet de l'interlocuteur. En effet, toutes conditions égales par ailleurs, nos locuteurs hypoarticulent significativement plus lorsqu'ils jouent en couple, que lorsqu'ils jouent avec un enquêteur inconnu. L'origine de l'enquêteur inconnu ne semble cependant pas avoir d'influence sur la durée des phones.

4 Conclusions et discussion

Grâce au protocole Diapix, nous avons réussi à induire des interactions entre un individu et trois types d'interlocuteurs : son/sa conjoint.e, un expérimentateur inconnu venant de la même région, et une expérimentatrice inconnue originaire d'une variété de français différente (ici le français de France). Nous avons également pu induire de la parole dirigée à soi-même, et avons complété par une lecture des mots ciblés par le protocole. Ces données doivent, à terme, nous permettre d'établir un profil des trois styles de parole (ou conditions de production) précédemment cités, notamment en termes de positionnement sur le continuum de l'hyper-hypo articulation.

L'analyse préliminaire des données recueillies pour 4 locuteurs met en évidence un effet de la situation de communication sur la durée des voyelles et sur leur distance au centre du système : on remarque que la lecture montre les segments les plus hyperarticulés, et dans la condition que nous avons appelée Solo (proche de la condition de description d'image chez Harmegnies & Poch-Olivé (1994)) on constate une différence de degré d'hypoarticulation à la fois avec la lecture et les trois

conditions de parole en interaction. Les résultats constatés dans la condition Solo correspondent aux résultats rapportés par Harmegnies & Poch-Olivé (1994) dans leur condition de description. Enfin, si l'on s'intéresse aux trois conditions de production impliquant une interaction (EnCouple, EnqSag, EnqEur), on remarque que l'identité de l'interlocuteur a également un effet sur la variation vocalique: les voyelles sont plus hypoarticulées lors de la partie en couple que lors de la partie avec les enquêteurs, et sur le plan de la distance du phone au centroïde du système la partie jouée avec l'enquêteur de la même origine régionale montre des voyelles plus centralisées que ce qui est observé dans la partie jouée avec l'enquêtrice européenne. Nos choix d'interlocuteur étant basés sur la distance sociale entre l'interlocuteur et le participant enregistré, nous interprétons ces différences comme une adaptation du degré d'hyper-hypo articulation à la proximité sociale entre les individus en interaction[2].

Les résultats généraux présentés ici ne montrent cependant pas les variations liées à la classe vocalique, or cette variation est fondamentale puisque la littérature nous apprend que certaines voyelles varient beaucoup plus que d'autres (les (εR) diphtonguent par exemple fréquemment – Paradis (1985) – alors que les (εK) sont moins impactés par la variation linguistique induite par les classes). Nous avons donc d'ores et déjà commencé à reproduire nos analyses classe par classe de façon à pouvoir mettre en évidence les classes vocaliques pertinentes pour l'étude de la variation extralinguistique et notamment phonostylistique visée ici. De plus, le calcul de la distance au centroïde des catégories pourrait se révéler plus pertinent une fois calculé en fonction des classes vocaliques propres au FQ et non pas des catégories vocaliques de façon générale. L'ajout d'autres types de mesures, notamment de distance F1-F2 et F2-F3, et écarts type de F1 et F2, tel qu'utilisées par Rvachew et al. (2006) pour l'étude des voyelles du français québécois, est également en cours.

Références

ADDA-DECKER M. & LAMEL L. (1999). Pronunciation variants across system configuration, language and speaking style. In *Speech Communication*, vol. 29, no 2-4, p. 83-98. DOI : 10.1016/S0167-6393(99)00032-1

AUDIBERT N., FOUGERON C., GENDROT C., ADDA-DECKER M. (2015). Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation. *18th International Congress of Phonetic Sciences* (ICPhS'15), Aug 2015, Glasgow, United Kingdom. pp.5.

BAKER R. & HAZAN V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, volume 43(3), p. 761-770. Springer. DOI : 10.3758/s13428-011-0075-y

BATES D., MAECHLER M., BOLKER B., WALKER S., CHRISTENSEN R., SINGMANN H., & BOLKER M. (2015). Package 'lme4'. Convergence, 12(1), 2.

BERSCHEID E., SNYDER M., & OMOTO A.M. (1989). The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of personality and Social Psychology*, volume 57(5), p. 792.

BOERSMA P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, vol. 5.

BRADLOW A., KRAUS N. & HAYES, E. (2003). Speaking clearly for children with learning disabilities. *Journal of Speech, Language, and Hearing Research,* volume 46, p. 80-97. DOI: 10.1044/1092-4388(2003/007)

[2] Les tâches et images ayant été réalisées et distribuées le même ordre, un effet de l'ordre est possible, cependant cet effet aurait dû mener à plus d'hypoarticulation dans la dernière phase de jeu (avec l'expérimentatrice européenne), puisqu'une forte fréquence des mots engendre une plus grande hypoarticulation de ces mots (Pluymaekers et al., 2005, par exemple), or les mots utilisés pour décrire les images avaient été énormément répétés avant cette dernière tâche.

CEDERGREN H. & SIMONEAU L. (1985). La chute des voyelles hautes en français de Montréal : «As-tu entendu la belle syncope?». In Lemieux M., Cedergren H., et Coll. réd. *Les tendances dynamiques du français parlé à Montréal*. Montréal : Office de la langue française, vol. 1, p. 57-144

CÔTÉ, M-H. (2012). Laurentian French (Quebec): extra vowels, missing schwas and surprising liaison consonants. In R. Gess, C. Lyche & T. Meisenburg (éds), *Phonological variation in French: Illustrations from three continents*. Amsterdam : John Benjamins, 235-274.

DUMAS D. (1987). *Nos façons de parler*. Presses de l'Université du Québec.

ESKENAZI, M. (1993). Trends in speaking styles research. In *Third European Conference on Speech Communication and Technology*.

GENDROT C. & ADDA-DECKER M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *Ninth European Conference on Speech Communication and Technology*.

HARMEGNIES B. & POCH-OLIVE D. (1994). Formants frequencies variability in French vowels under the effect of various speaking styles. Le *Journal de Physique IV*, 4(C5):C5–509.

HUET K. & HARMEGNIES B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *Actes des JEP*'2000, p. 225-228.

LABOV W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

LABOV W. (2002). Driving forces in linguistic change. In International Conference on Korean Linguistics. Seoul National University. Seoul, South Korea.

LINDBLOM B. (1990) Explaining Phonetic Variation: A Sketch of the H&H Theory. In: Hardcastle W.J., Marchal A. (eds) *Speech Production and Speech Modelling*., vol 55. Springer, Dordrecht. DOI: 10.1007/978-94-009-2037-8_16

NAKAGAWA S. & SCHIELZETH H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4(2)*, p. 133-142.

PARADIS, C. (1985). *An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi and Jonquiere (Quebec)*. Doctoral dissertation, University of Pennsylvania.

PLUYMAEKERS M., ERNESTUS M., & BAAYEN R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561-2569.

POCH-OLIVE D. & HARMEGNIES B. (1992). Variations structurelles des systèmes vocaliques en français et espagnol sous l'effet du style de parole. *Le Journal de Physique IV*, 2(C1) :C1–283.

SANTERRE, L. (1976). Voyelles et consonnes du français québécois populaire. In *Identité culturelle et francophonie dans les Amériques*, volume 1, p. 21-36, PUL Québec.

SCARBOROUGH R., DMITRIEVA O., HALL-LEW L., ZHAO Y., & BRENIER J. (2007). An acoustic study of real and imagined foreigner-directed speech. *Journal of the Acoustical Society of America*, volume 121(5), p. 3044-3044. DOI: 10.1121/1.4781735

SCARBOROUGH R. & ZELLOU G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, vol. 134(5), p. 3793-3807. DOI: 10.1121/1.4824120

ROUAS J-L., BEPPU M. & ADDA-DECKER M. (2010). Comparison of spectral properties of read, prepared and casual speech in French. In *LREC* proceedings.

RVACHEW S., MATTOCK K., POLKA L., & MÉNARD L. (2006). Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French. The Journal of the Acoustical Society of America, 120(4), 2250-2259. DOI: 10.1121/1.2266460

WALKER, D. (1984). The pronunciation of Canadian French. University of Ottawa press. DOI : 10.2307/327346

YAEGER, M. (1975). Speaking style: Some phonetic realizations and their significance. *Pennsylvania working paperson linguistic change and variation*, volume. I, No. 1, US Regional Survey.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

De la possibilité d'un relâchement des voyelles hautes dans les troncations finissant par /v, z, ʒ, ʁ/ en français québécois

Mélanie Lancien[1, 2]
(1) Section SLI - Université de Lausanne, CH1015 Lausanne, Suisse
(2) Université du Québec à Chicoutimi, QC G7H 2B1 Chicoutimi, Canada
melanie.lancien@unil.ch

RÉSUMÉ

Le français québécois possède trois voyelles hautes tendues (/i, y, u/), et trois relâchées ([ɪ, ʏ, ʊ]), les relâchées étant décrites comme des allophones des tendues produits en syllabe fermée par une consonne non allongeante (Walker, 1984 ; Dumas, 1987 ; entre autres). Cependant Côté (2012) pose l'hypothèse que ce relâchement pourrait survenir dans des troncations finissant par une consonne allongeante (Troncation(_R)). Dans cette étude pilote, nous explorons cette hypothèse. A travers de courts textes (lus par deux locuteurs natifs) nous avons induit ces voyelles en Troncation(_R) ainsi que dans des positions formant des paire minimales (PaireMinimale(_R) / semi-minimales (finissant par une consonne non allongeante - PaireMinimale(_K)) avec les Troncation(_R). Les premières analyses temporelles (ANOVAs) montrent des /i, y, u/ plus courts en Troncation(_R) qu'en PaireMinimale(_R), et une analyse qualitative permet d'observer des variations spectrales entre Troncation(_R), PaireMinimale(_R) et PaireMinimale(_K), notamment à travers les moyennes de F1 et F2 (Hz).

ABSTRACT

On the possibility of high vowels' laxing in truncations ended by /v, z, ʒ, ʁ/ in Quebec French.
Quebec French has three high tense vowels (/i, y, u/), and three high lax vowels ([ɪ, ʏ, ʊ]), the laxed ones being considered as allophones of the tensed ones that arise when the vowel is in a syllable closed by a non-lengthening consonant (Walker, 1984; Dumas, 1987; among others). However, Côté (2012) hypothesizes that this laxing could occur in truncations ending with a lengthening consonant (Truncation(_R)). In this pilot study, we explore this hypothesis. We used short texts (read by two native speakers) to induce high vowels in Truncation(_R) as well as in positions forming minimal (PaireMinimale(_R) / semi-minimal pairs (ending with a non-lengthening consonant - PaireMinimale(_K)) with Truncation(_R). Our first temporal analysis (ANOVAs) show shorter vowels in Truncation(_R) than in PairMin(_R), and a more qualitative analysis makes it possible to observe spectral variations between Truncation(_R), PaireMinimale(_R) and PaireMinimale(_K), in particular through mean F1 and F2 (Hz).

MOTS-CLÉS : Français québécois, voyelles, relâchement.
KEYWORDS : Quebec French, vowels, laxing.

Volume 1 : Journées d'Études sur la Parole, pages 317–325.
HAL : hal-02798554.

1 Introduction

1.1 Le système vocalique du français québécois

Le français québécois (FQ) possède un système vocalique plus diversifié que celui du français dit « standard ». Plusieurs auteurs, notamment Santerre (1976), Walker (1984), et Côté (2012), ont établi différents inventaires des voyelles du français québécois. Ces inventaires vont de 16 voyelles phonologiques (12 voyelles orales /i, y, u, e, ɛ, ɜ, ø, œ, o, ɔ, a, ɑ/ et quatre nasales /ɔ̃, ɑ̃, ɛ̃, œ̃/) pour Santerre (1976), jusqu'à 23 voyelles contrastives pour Côté (2012). Ces différents systèmes sont synthétisés en FIGURE 1. Malgré des désaccords dans la composition des inventaires, tous ces auteurs s'accordent sur la vaste variation allophonique qui touche ces phonèmes et change leurs caractéristiques spectrales et temporelles drastiquement. En effet, ces phonèmes sont soumis à une forte variation contextuelle : ils se renforcent (allongement, diphtongaison), s'affaiblissent (relâchement, dévoisement, syncope), ou se déplacent sur l'axe de l'antéro-postériorité en fonction du type de syllabe dans laquelle ils se trouvent (fermée/ouverte), du phone fermant la syllabe (consonne allongeante ou non), ainsi que de leur position prosodique (accentuée ou non accentuée).

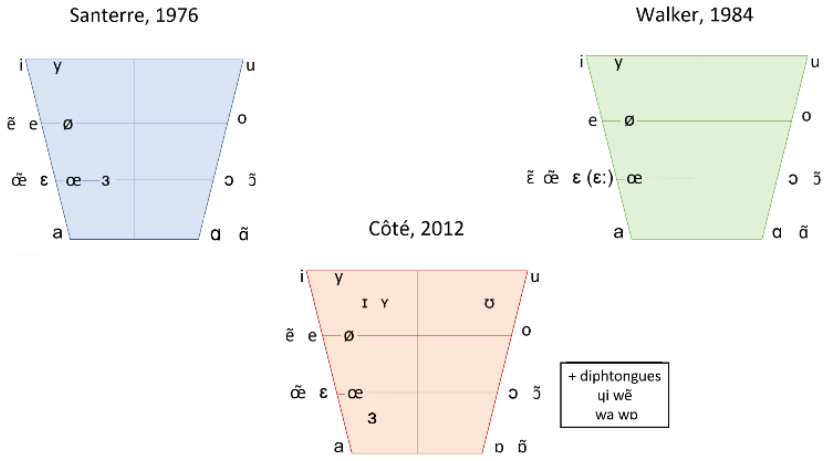


FIGURE 1 : Systèmes vocaliques du français québécois proposés par Santerre (1976) – en bleu –, Walker (1984) – en vert –, et Côté (2012) – en orange –.

Dans les systèmes présentés en FIGURE 1, on note plusieurs différences, une des plus flagrantes étant l'ajout de quatre diphtongues phonologiques par Côté (2012), cependant nous ne nous attarderons que sur l'une d'entre elles : l'ajout d'une série de voyelles orales hautes relâchées (/ɪ, ʏ, ʊ/) dans le système proposé par Côté (2012). Ces trois phonèmes supplémentaires sont au cœur de la problématique que nous comptons explorer dans cet article.

Dans les travaux de Santerre (1976) et Walker (1984), les voyelles hautes sont au nombre de trois : /i, y, u/, et ces voyelles se relâchent lorsqu'elles sont accentuées et suivies d'une consonne non allongeante (voir la sous-section ci-après pour plus de détails). Cependant les travaux de Côté (2012) militent pour deux séries de voyelles hautes : trois tendues /i, y, u/ et trois relâchées /ɪ, ʏ, ʊ/. L'auteure fait appel à deux arguments principaux : celui de l'acquisition, et celui que nous appellerons « des troncations ». L'argument de l'acquisition du langage se base sur le fait que l'on trouve certaines paires minimales impliquant des tendues et des relâchées, ces paires, par exemple « cool » /ku:l/ et « coule » /kʊl/, impliquant des mots d'emprunt à l'anglais (sur lesquels nous revenons également dans

la sous-section ci-dessous). L'argument de l'auteur repose sur le fait que de telles oppositions phonologiques, chez les enfants, viennent d'une réanalyse lors de l'acquisition du langage, qui implique deux séries de voyelles hautes : des tendues et des relâchées. Le second argument, mobilisant des troncations, repose sur le fait qu'intuitivement une opposition entre un mot contentant une voyelle tendue (ex : « muses » /myz/) et la version tronquée d'un mot qui contenait une voyelle tendue mais dont la troncation contient une voyelle relâchée (ex : « muz » /mʏz/ pour « musique » /myzik/) serait envisageable. La troncation « muz » /mʏz/ pour musique contiendrait donc une voyelle relâchée et serait en opposition avec le mot « muse » qui lui contient une voyelle tendue, formant ainsi une paire minimale y/ʏ. Dans cet article, nous proposons d'analyser acoustiquement des troncations du type mentionné de façon à apporter à cette hypothèse une dimension plus expérimentale.

1.2 Voyelles hautes et transformations

Dans cet article nous nous intéressons donc aux trois voyelles hautes /i, y, u/. Ce trio est soumis à 6 types de transformations : le relâchement, l'allongement, la diphtongaison, le dévoisement, la syncope, ou l'ouverture. Ici nous nous focalisons principalement sur le phénomène de relâchement (voir Dumas, 1981 ; Walker, 1984 ; Dumas, 1987 ; entre autres).

Le relâchement des voyelles hautes est lié au phénomène d'allongement : les voyelles hautes non allongées sont relâchées en syllabe finale fermée (= syllabe accentuée) et dans certains contextes prétoniques par assimilation régressive (Walker, 1984 ; Yaeger et al., 1977 ; Dumas, 1987 ; Dumas et Boulanger, 1982, Brent, 1971). Concernant les emprunts à l'anglais, McLaughlin, (1986) remarque que pour les emprunts de la première vague, tels que « toune » (de l'anglais « thune ») le relâchement (ou ouverture selon sa terminologie) a également lieu, en revanche pour les emprunts plus récents ce relâchement peut ne pas être d'actualité.

D'un point de vue acoustique : les voyelles tendues /i, y, u/ sont plus courtes et plus périphériques que leur contrepartie relâchée (Gendron, 1966, Paradis, 1985, Martin, 2002, Poliquin, 2006, Arnaud et al., 2011). Les variantes relâchées /ɪ, ʏ, ʊ/ ont également un F1 plus élevé que celui des tendues, et sont légèrement plus centralisées avec un F2 plus bas pour /ɪ/ et /ʏ/ et plus élevé /ʊ/ (Sigouin & Arnaud, 2015 ; MacKenzie & Sankoff, 2010 – les mesures relevées par MacKenzie & Sankoff, 2010 sont résumées en TABLE 1). Dans ces descriptions formantiques, les auteurs n'ont cependant considéré qu'un point de mesure par occurrence (ou une moyenne de plusieurs points) ne permettant pas d'étudier l'évolution en cours de production. Les travaux plus récents, réalisés par Arnaud et al. (2011) sur des logatomes nous apprennent que dans un espace F1xF2 les variantes relâchées se centralisent en cours d'émission alors que les tendues se déplacent en périphérie.

	[iC]	*/i/*	*[yC]*	*/y/*	*[uC]*	*/u/*
F1 (Hz)	458	381	441	378	457	367
F2 (Hz)	1994	2343	1739	1939	1270	971
N	153	168	159	145	160	153

TABLE 2 : Résumé des valeurs moyennes de F1 (Hz) et F2 (Hz) relevées par MacKenzie & Sankoff (2010, p94) pour les voyelles tendues (/i, y, u/) et relâchées ([iC, yC, uC]) du français québécois.

Dans cet article nous revenons sur l'hypothèse de Côté (2012:243): « I could not find actual exemples [...] but a form such as [mʏz] for *musique* is quite conceivable [...] in opposition to *muse* 'muse' [my:z]. ». Notre but sera de créer des situations dans lesquelles les types de formes mentionnées par Côté (2012) (e.g. « muz- » pour « musique » en opposition à « une muse ») sont utilisées par des locuteurs du français québécois. Par la suite nous analyserons acoustiquement les voyelles produites de façon à mettre en évidence de potentielles différences temporelles ou spectrales qui pourraient alimenter l'hypothèse de Côté (2012) et montrer un relâchement des voyelles dans les troncations finissant par une consonne allongeante.

2 Méthodologie

2.1 Protocole expérimental

Pour cette étude nous avons choisi d'induire différents types de troncations via de courts textes à lire. En tout 72 textes ont été lus par un locuteur montréalais (40ans) et un locuteur saguenéen (27 ans). Les types de textes utilisés sont présentés et résumés en TABLE 2. Les données ont été enregistrées en chambre sourde grâce à un Tascam HD-P2 et un micro serre-tête Sure SM-10A (dynamique, cardioïde).

Exemple de texte	Type de forme	Nombre
Michel arrive au travail. Il ne voit personne dans les bureaux. […] il se dit : A matin c'est vide au bur-	Troncations finissant par /v, z, Z, R/ (Troncation(_R))	19 textes
Le « bijam » est un instrument de musique moldave. Lucie prend des cours de bijam. Elle arrive en retard et dit au professeur : J'm'excuse, j'suis en retard pour le bij-	= ci-dessus mais avec des non-mots (Troncation(_R))	15 textes
Max et sa famille vont à l'église. Son fils lui demande quelle est la robe bizarre que porte le curé. Il lui répond : Ici les curés portent la bure	Mots entiers formant des paires minimales avec Troncation(_R) (noté PaireMinimale(_R))	18 textes
Mathilde va chez le fleuriste avec sa mère. Elles veulent des tulipes pour le salon. Mathilde demande à sa mère : J'aimerais des roses avec les tul-	Troncations se terminant par une consonne non allongeante (Troncation(_K))	10 textes
Martin et Lise essaient de coudre une robe. Martin dit à Lise : J'aimerais ça qu'on mette du tulle	Mots entiers formant des paires minimales avec les Troncation(_K) (PaireMinimale(_K))	10 textes

TABLE 2 : Résumé des formes proposées par le protocole expérimental.

En plus de la lecture du texte, chaque locuteur a dû répéter les mots cibles hors contexte, nous permettant ainsi d'avoir une répétition en isolation et une répétition en contexte. Dans tous les textes, la troncation cible était le dernier mot d'une phrase de 7-8 syllabes, de façon à lui assurer une position accentuée et donc à garantir la possibilité d'action de la règle de relâchement. Nous avons donc pu comparer les voyelles /i, y, u/ dans 6 contextes phonologiques différents, et pour la position (_R) dans des mots et des non mots. Pour exemplifier : nous avons pu comparer les /y/ de "muz-" (Troncation(_R)), "musique" (MotEntier(_R)), "muses" (PaireMinimale(_R)), "buch-" (Troncation(_K)), "bucher" (MotEntier(_K)), "buches" (PaireMinimale(_K)).

2.2 Données recueillies

En FQ, les voyelles hautes sont également sujettes à des apocopes et des dévoisements, ainsi, toutes les voyelles prévues n'ont pas nécessairement été produites, et toutes les voyelles produites n'ont pas pu être conservées pour l'analyse (dû à l'absence de F1 pour les voyelles soufflées). Au total, 84 /i/, 90 /u/ et 79 /y/ ont été examinés pour le locuteur montréalais, et 82 /i/, 81 /u/, et 71 /y/ pour le locuteur saguenéen. La TABLE 3 ci-dessous récapitule le nombre de phones conservés pour l'analyse dans chaque position prosodique, pour chaque classe de mot, pour le locuteur saguenéen et pour le locuteur montréalais.

Position Prosodique	*Type*						***Total***
	MotEntier(_K)	MotEntier(_R)	PaireMin(_K)	PaireMin(_R)	Troncation(_K)	Troncation(_R)	
Isolation	9 | 10	32 |34	11 | 11	15 | 17	10 | 9	26 | 34	103 | 115
Phrase	12 | 13	48 | 50	12 | 11	18 | 23	10 | 10	31 | 31	131 | 138
Total	21 | 23	80 | 84	23 | 22	33 | 40	20 | 19	57 | 65	234 | 253

TABLE 3 : Résumé des données conservées pour l'analyse (à gauche de la barre verticale le nombre pour le locuteur saguenéen et à droite celui pour le locuteur montréalais)

2.3 Traitement des données

Les données enregistrées ont été transcrites orthographiquement sous Praat, puis alignées grâce à la version québécoise de l'outil d'alignement SPPAS (Lancien et al., à paraitre). La segmentation des voyelles cibles a été corrigée manuellement lorsque nécessaire (en prenant pour repère les débuts et fins des formants supérieurs). Suite à cela, un script Praat a permis d'extraire les moyennes de F1, F2, et F3, leurs valeurs aux cinq cinquièmes de phone, ainsi que la durée de chaque voyelle ciblée. Ces mesures nous ont permis de calculer les mesures de compacité F1-F2, compacité F2-F3, distance entre le début et la fin de la voyelle dans un espace F1*F2 et dans un espace F1*F2*F3 ($\sqrt{(f1\_80\% - f1\_20\%)^2 + (f2\_80\% - f2\_20\%)^2}$; et $\sqrt{(f1\_80\% - f1\_20\%)^2 + (f2\_80\% - f2\_20\%)^2 + (f3\_80\% - f3\_20\%)^2}$).

2.4 Analyse des données

Pour analyser les données recueillies nous avons choisi de les diviser en plusieurs sous-groupes en fonction du locuteur, de la catégorie vocalique (/i, y, u/) et de la position prosodique (à savoir : en position isolée / dans une phrase), nous laissant ainsi avec 12 sous-groupes. Nous avons réalisé une ANOVA pour chacun de ces sous-groupes et pour chaque indice acoustique cité ci-dessus. L'indice acoustique était la variable dépendante, et la variable indépendante fixe était la classe de mot (Troncation(_R), MotEntier(_R), PaireMinimale(_R), Troncation(_K), MotEntier(_K), PaireMinimale(_K)).

3 Résultats

Dans un premier temps nous nous focaliserons sur les résultats concernant les différences entre les voyelles en Troncation(_R) – où la voyelle devrait être relâchée selon l'hypothèse que nous testons – et en PaireMinimale(_R) – où la voyelle devrait être tendue –, puis nous mettrons en relief ces résultats par des comparaisons plus qualitatives entre les trois classes Troncation(_R), PaireMinimale(_R), et PaireMinimale(_K).

Pour les trois voyelles hautes chez nos deux locuteurs, les ANOVAs montrent uniquement des différences significatives de longueur entre les classes testées, et notamment entre les voyelles en PaireMinimale(_R) et en Troncation(_R), indiquant que les voyelles dans la troncation se terminant par une consonne allongeante sont plus courtes que leur homologue en mot monosyllabique se terminant par une consonne allongeante[1]. Cependant, l'absence de significativité pour les mesures spectrales pourrait être liée au peu de données recueillies et à la forte variation formantique liée aux segments précédant et suivant la voyelle, nous avons donc choisi de rapporter les différences observées qualitativement dans les mesures formantiques effectuées.

3.1 Le locuteur saguenéen

Dans les productions du locuteur saguenéen, on ne trouve que très peu de différences entre les valeurs de F1 et F2 en Troncation(_R) et en PaireMinimale(_R), indiquant que la qualité des deux voyelles est très proche. Cependant si l'on examine les différences de F1 et F2 entre les voyelles /i/ et /y/ en classes Troncation(_R), PaireMinimale(_R), et PaireMinimale(_K), on remarque que la classe Troncation(_R) se situe presque systématiquement entre les deux autres. Bien que les ANOVAs n'aient montré aucune différence significative entre ces trois groupes pour les deux premiers formants, on peut donc tout de même noter une tendance des voyelles en Troncation(_R) à se situer à mi-chemin entre les voyelles relâchées et les voyelles tendues (ce positionnement n'est pas observé pour les formants des /u/, notamment à cause d'une plage de variation très forte des formants dans toutes les classes). Ces observations sont congruentes avec les indices de relâchement mis en évidence par la littérature, à savoir un F1 plus élevé et un F2 plus bas pour les relâchées que pour les tendues. Ces variations sont visibles en FIGURE 3. Ce pattern se retrouve également dans les mesures de

[1] Pour des raisons d'espace nous n'avons pas pû insérer les tableaux présentant les résultats statistiques de toutes nos ANOVAs

distance F1-F2 et F2-F3, ainsi que dans les trajectoires des formants dans un espace F1*F2 et dans un espace F1*F2*F3 (dans une moindre mesure, surtout pour les /y/).

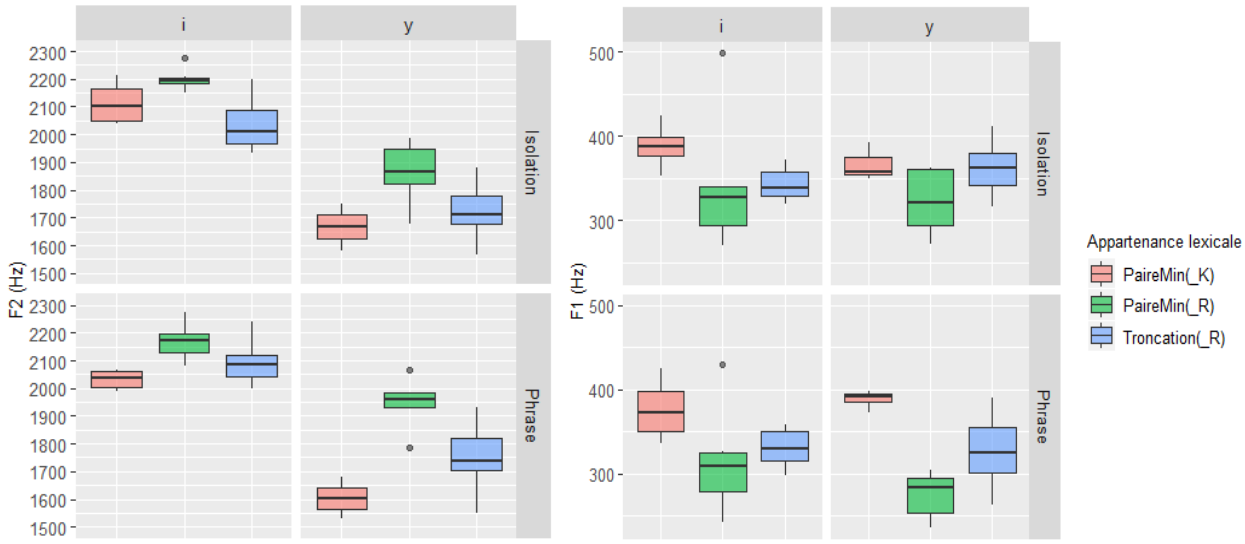


FIGURE 3 : Boîtes à moustaches représentant la dispersion des variations de F1 (à droite) et F2 (à gauche) en Hz pour les voyelles /i/ et /y/ dans les classes Troncation(_R) (en bleu), PaireMinimale(_R) (en vert), et PaireMinimale(_K) (en saumon) pour les données du locuteur saguenéen.

3.2 Le locuteur Montréalais

Pour le locuteur montréalais, nous observons le même type de résultats que précédemment. Bien qu'aucune différence ne soit significative, on peut noter une tendance des voyelles en Troncation(_R) à se situer à mi-chemin entre les voyelles relâchées (PaireMinimale(_K)) et les voyelles tendues (PaireMinimale(_R)), avec des F1 plus hauts et des F2 plus bas en Troncation(_R) qu'en PaireMinimale(_R). Ces variations sont visibles en FIGURE 4. Ce pattern se retrouve également dans les mesures de distance F1-F2 et F2-F3, mais pas dans les trajectoires des formants dans un espace F1*F2 et dans un espace F1*F2*F3.

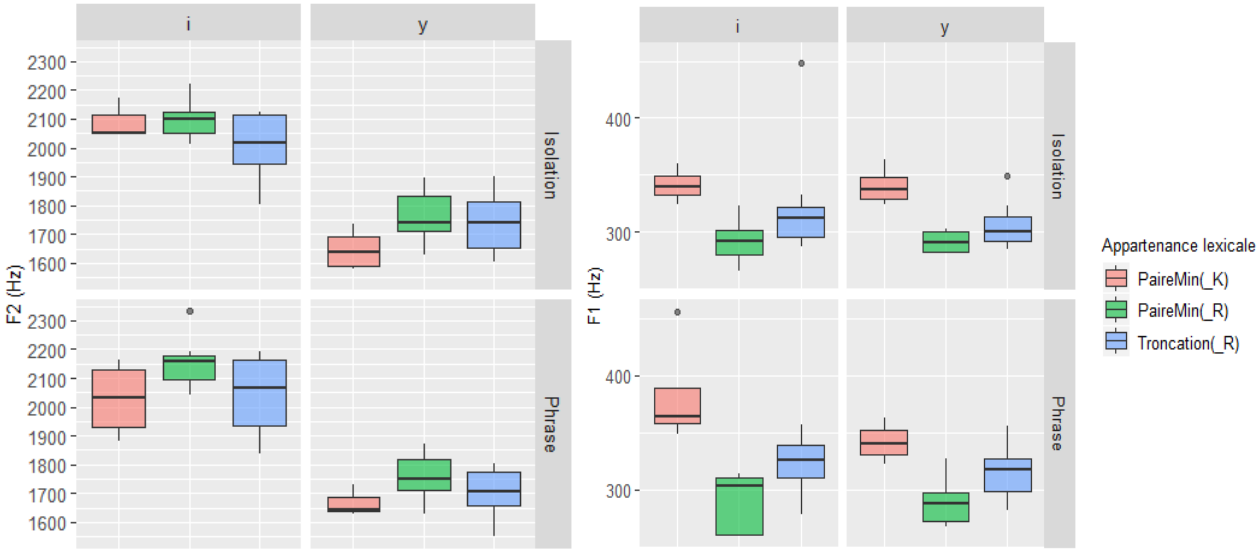


FIGURE 4 : Boîtes à moustaches représentant la dispersion des variations de F1 (à droite) et F2 (à gauche) en Hz pour les voyelles /i/ et /y/ dans les classes Troncation(_R) (en bleu), PaireMinimale(_R) (en vert), et PaireMinimale(_K) (en saumon) pour les données du locuteur montréalais.

4 Discussion et conclusions

Dans cette expérience nous avons tenté d'induire des formes tronquées de mots ayant pour noyau une voyelle haute directement suivie par une consonne allongeante dans le but de tester l'hypothèse avancée par Côté (2012). Cette hypothèse, qui s'oppose à l'interprétation classique du relâchement faite par la phonologie (à savoir : les voyelles hautes tendues se relâchent lorsqu'elles sont en syllabe fermée par une consonne non allongeante et accentuées), pose la possibilité d'un relâchement des voyelles hautes dans les formes tronquées susmentionnées. Ce relâchement « irrégulier » est mobilisé par l'auteur comme un argument en faveur de la classification des voyelles hautes relâchées comme des phonèmes du français québécois et non plus seulement comme des allophones des voyelles hautes tendues.

Nous avons induit les formes ciblées par la lecture de petits textes mettant en scène l'utilisation de ces troncations. Ces textes, lus par un locuteur montréalais et un locuteur saguenéen, ont ensuite permis l'analyse des propriétés acoustiques des voyelles hautes, notamment par la comparaison de troncations finissant par /ʁ, v, z, ʒ/ à des mots complets monosyllabiques formant des paires minimales avec ces troncations et finissant soit par une consonne allongeante (dans ces cas, la voyelle est supposée être tendue) ou par une consonne non allongeante (dans ces cas, la voyelle est supposée être relâchée).

Si les résultats des ANOVAs ont montré des voyelles significativement plus courtes dans les Troncation(_R) que dans les PaireMinimale(_R) – le raccourcissement des voyelles étant caractéristique du relâchement selon Martin (2002) mais pas selon Arnaud (2011) – on ne remarque cependant rien de significatif sur le plan spectral (ce qui peut aisément s'expliquer par le peu d'occurrences analysées et la forte variabilité de nos données). Il reste tout de même intéressant de noter que, d'un point de vue qualitatif, l'observation des moyennes et de la variance de F1 et F2 (Hz) ainsi que des mesures de distance entre F1 et F2, F2 et F3, et des mesures dynamiques en 2D et 3D a montré un positionnement général des voyelles prononcées en Troncation(_R) entre celles prononcées en PaireMinimale(_R) et PaireMinimale(_K), avec des indices caractéristiques d'un possible relâchement (tels qu'un F1 plus haut et un F2 plus bas). En somme, si le /y/ de « muz- » n'est pas aussi relâché que celui de « mute » (/mʏt/), il n'en reste pas moins qu'il n'est pas pour autant aussi tendu que celui de « muse » (/myz/).

À ce stade, il est impossible d'établir une conclusion stable sur la qualité des voyelles hautes dans les troncations finissant par une consonne allongeante. Les différences observées ici montrent cependant qu'il serait intéressant d'analyser ce phénomène à plus grande échelle : plus de locuteurs et plus d'occurrences nous permettraient d'obtenir un meilleur profilage spectral des voyelles dans les formes induites par notre protocole. Il nous paraît également impératif de croiser les analyses acoustiques avec des analyses perceptives. Les prochaines étapes de ces travaux seront donc de prolonger la collecte des données, ainsi que de construire un test AXB ayant pour stimuli les données récoltées pour notre analyse acoustique. Ainsi nous pourrons répertorier les voyelles perçues comme relâchées/tendues dans les formes tronquées (mais aussi dans les formes en (_K), puisque le

relâchement n'est pas systématique) et pourrons observer plus finement les caractéristiques spectrales et temporelles de ces voyelles.

Références

ARNAUD V., SIGOUIN C., & ROY J. P. (2011). Acoustic Description of Quebec French High Vowels: First Results. In the proceedings of *ICPhS* (p. 244-247).

BRENT E. (1971). *Canadian French: a synthesis* (Doctoral dissertation, Cornell University).

CEDERGREN H. & SIMONEAU L. (1985). La chute des voyelles hautes en français de Montréal : « As-tu entendu la belle syncope ? ». In Lemieux M., Cedergren H., et Coll. réd. *Les tendances dynamiques du français parlé à Montréal*. Montréal : Office de la langue française, vol. 1, p. 57-144

CÔTÉ, M-H. (2012). Laurentian French (Quebec): extra vowels, missing schwas and surprising liaison consonants. In R. Gess, C. Lyche & T. Meisenburg (éds), *Phonological variation in French: Illustrations from three continents*. Amsterdam: John Benjamins, 235-274. DOI: 10.1075/silv.11.13cot

DUMAS D. (1981). Structure de la diphtongaison québécoise. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, vol. 26(1), p. 1-61. DOI : 10.1017/S0008413100023513

DUMAS D. & BOULANGER A. (1982). Les Matériaux d'origine des voyelles fermées du français Québécois. *Revue québécoise de linguistique*, vol. 11(2), p. 49-72. DOI : 10.7202/602487ar

DUMAS D. (1987). *Nos façons de parler*. Presses de l'Université du Québec. ISBN 978-2-7605-0445-5

GENDRON J. D. (1966). Contribution à l'étude du français rural parlé au Canada. *Travaux de Linguistique et de Littérature*, vol. 4, p. 173-189.

MARTIN P. (2002). Le système vocalique du français du Québec. De l'acoustique à la phonologie. *La linguistique*, vol. 38(2), p. 71-88. DOI : 10.3917/ling.382.0071

MACKENZIE L. & SANKOFF G. (2010). A quantitative analysis of diphthongization in Montreal French. *University of Pennsylvania Working Papers in Linguistics,* vol. *15*(2).

MCLAUGHLIN A. (1986). Une (autre) analyse de la distribution des variantes des voyelles hautes en français montréalais in Etudes de phonologie historique du français québécois. *Revue québécoise de linguistique théorique et appliquée*, vol. 5(4), p. 21-60.

PARADIS C. (1985). *An Acoustic Study of Variation and Change in the Vowel System of Chicoutimi and Jonquiere* (Quebec) (Doctoral dissertation, Graduate School of Arts and Sciences, University of Pennsylvania).

POLIQUIN, G. C. (2006). *Canadian French vowel harmony* (Doctoral dissertation).

SANTERRE, L. (1976). Voyelles et consonnes du français québécois populaire. In *Identité culturelle et francophonie dans les Amériques*, volume 1, p. 21-36, PUL Québec.

SIGOUIN C. & ARNAUD V. (2015). Quebec French close vowels in lengthening contexts: tense, lax or diphthongised? An acoustic study. In the proceedings of *ICPhS.*

WALKER, D. (1984). The pronunciation of Canadian French. University of Ottawa press. DOI : 10.2307/327346

YAEGER M., CEDERGREN H., & SANKOFF D. (1977). Harmonie et conditionnement consonantique dans le système vocalique du français parlé à Montréal. *Phonologie et Société,* ed. by H. Walter. Paris : Didier. DOI :10.2307/413342

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Paramètres acoustiques et phonétiques dans la parole parkinsonienne avant et après traitement LSVT LOUD®

Maëlle Le Cerf[1], Emmanuel Ferragne[2]
(1) UFR Santé, Université de Franche-Comté, 25000 Besançon
(2) Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Univ. Sorbonne Nouvelle
maellelecerf@orange.fr, emmanuel.ferragne@u-paris.fr

RÉSUMÉ

Objet : Notre recherche examine l'effet du Lee Silverman Voice Treatment (LSVT LOUD®) sur l'aire et la position de l'espace vocalique, la fréquence fondamentale (f0), les paramètres de qualité de voix, le débit de parole, le temps maximum phonatoire (TMP) et le ressenti de handicap vocal chez des patients francophones atteints de la maladie de Parkinson. **Méthode** : Un même protocole a été proposé en prétest et post-test à 12 patients parkinsoniens. **Résultats** : En post-test, nous observons une descente significative de l'espace vocalique, une différence de f0 entre la parole lue et la parole spontanée, une amélioration significative des paramètres de qualité de voix (*jitter*, *shimmer*, *HNR*) et du ressenti de handicap vocal. Le débit de parole des patients est maintenu, le TMP subit un effet de l'exercice.

ABSTRACT

Acoustic and phonetic parameters in parkinsonian speech before and after LSVT LOUD®
Purpose: Our research examines the effect of the Lee Silverman Voice Treatment (LSVT LOUD®) on the area and position of the vowel space, fundamental frequency (f0), voice quality parameters, speech rate, maximum phonation time (MPT) and patient-perceived vocal handicap in French-speaking patients with Parkinson's disease. **Method**: The same protocol was proposed in pre-test and post-test to 12 Parkinsonian patients. **Results**: In post-test, we observed a significant lowering of the vowel space, a difference of f0 between read speech and spontaneous speech, a significant improvement in voice quality parameters (*jitter*, *shimmer*, *HNR*) and perceived vocal handicap. The patients' speech rate is maintained, the MPT shows an effect of exercise.

MOTS-CLÉS : Maladie de Parkinson, dysarthrie, LSVT LOUD®, analyse acoustique
KEYWORDS: Parkinson's disease, dysarthria, LSVT LOUD®, acoustic analysis

Volume 1 : Journées d'Études sur la Parole, pages 326–334.
HAL : hal-02798555.

1 Introduction

La maladie de Parkinson (MP) est la deuxième maladie neuroévolutive la plus fréquente en France. Affectant environ 1% de la population de plus de 65 ans, elle est la cause majeure de handicap chez les personnes âgées (Inserm, 2015). L'évolution de cette maladie chronique est lente ; le but pour les personnes atteintes est de conserver leur autonomie le plus longtemps possible. Le parcours de soins du patient parkinsonien est multiple et se déploie autour d'un axe principal : l'éducation thérapeutique. La prise en charge orthophonique recommandée par la Haute Autorité de Santé (HAS) est un entraînement intensif grâce au traitement LSVT LOUD® (HAS, 2016).

La dysarthrie est un trouble de l'exécution motrice de la parole, dont l'origine est une lésion du système nerveux central ou périphérique. Dans la MP, elle est qualifiée d'hypokinétique et couvre un tableau clinique large où la respiration, la phonation, l'articulation, les résonances et la prosodie peuvent être touchées. Le schéma hyporespiratoire a pour conséquences une réduction du temps maximum phonatoire (TMP), et une diminution significative du nombre de syllabes produites par expiration. Le débit de parole des patients parkinsoniens a tendance à être plus rapide, en lien avec un contrôle articulatoire dégradé en fin d'énoncés (Liu *et al*., 2019). Plusieurs paramètres de la voix sont modifiés. La hauteur est impactée, due à une rigidité du muscle crico-thyroïdien, et à une réduction de la pression sous-glottique (Robert & Spezza, 2005). L'empan de la fréquence fondamentale est diminué, ce qui conduit à qualifier subjectivement la voix de « monotone » (Liu *et al*., 2019). Les études ne trouvent pas de consensus quant aux changements de la hauteur de la voix dans la MP (Ghio *et al*., 2014). L'hypophonie parkinsonienne est due à une mauvaise coordination du vibrateur laryngé, et une diminution des volumes d'air expirés. Les mesures de l'instabilité de la hauteur (*jitter*) et de l'intensité (*shimmer*) augmentent dans la parole parkinsonienne (Jiménez-Jiménez *et al*., 1997). La détérioration de la fonction laryngée donne lieu à un timbre dit « éraillé » ou « soufflé », typiquement caractérisé par l'ajout de bruit dans le signal acoustique, qu'il est possible de quantifier en examinant la diminution du *harmonic-to-noise ratio – HNR* (Yücetürk *et al*., 2002). L'altération des voyelles dans la MP est caractérisée par un déplacement de l'espace vocalique vers le haut, le formant F1 étant réduit du fait de l'aperture diminuée des voyelles (Audibert & Fougeron, 2012). L'ensemble de ces troubles engendre une perte d'intelligibilité, exacerbant l'impact psychosocial de la dysarthrie (Atkinson-Clement *et al*., 2019). L'évaluation du handicap vocal grâce au Voice Handicap Index (VHI) permet de connaître l'impact des troubles sur la qualité de vie des patients, et de juger de l'efficacité des pratiques thérapeutiques.

La dopathérapie et la stimulation cérébrale profonde apportent des effets bénéfiques sur les troubles moteurs, mais les effets sont parfois insatisfaisants sur la parole (Brabenec *et al*., 2017). La rééducation orthophonique LSVT LOUD® apparaît être une alternative à ces traitements. Elle cible le calibrage de l'intensité vocale adaptée aux situations de parole (Ramig *et al*., 2001). Bien que les échantillons de patients des études menées soient trop faibles pour soutenir ou réfuter un type de rééducation orthophonique par rapport à un autre (Herd *et al.*, 2012), la HAS considère ce protocole « *comme la méthode de référence* » (HAS, 2016, p. 54). L'essai clinique randomisé de Ramig *et al.* (2018) constitue à ce jour la meilleure preuve d'efficacité du traitement. Seize séances d'une heure,

en 4 semaines, permettent aux patients d'utiliser une parole forte (90dB) en entraînement, pour parler à intensité normale (60dB) en spontané. L'apprentissage et la répétition des exercices ont des effets neuroplastiques prouvés (Liotti *et al*., 2003 ; Narayana *et al*., 2010).

Plusieurs études ont été réalisées pour évaluer l'efficacité de cette pratique. D'une part, l'intensité est améliorée de façon significative à long terme (+24 mois) sur des voyelles prolongées (Wight & Miller, 2015). Cette amélioration n'est pas significative pour les tâches de lecture et de monologue. Les paramètres acoustiques se rapprochent de la parole non pathologique, à travers un élargissement du triangle vocalique (Sapir *et al*., 2007). L'évolution de ces paramètres permet une amélioration de l'intelligibilité de phrases (Cannito *et al*., 2012). D'autre part, des effets annexes sont relevés : réduction des troubles de la déglutition et réduction de l'hypomimie (El Sharkawi *et al*., 2002 ; Spielman *et al*., 2003). Enfin, le ressenti de handicap vocal est amélioré, jusqu'à 12 mois post-traitement (Wight & Miller, 2015). L'étude de la parole dite « pathologique » est à ce jour un sujet de recherche commun entre phonéticiens et cliniciens. La dysarthrie hypokinétique fait l'objet de nombreux travaux pour établir de façon objective l'impact des troubles respiratoires, vocaux, prosodiques et articulatoires engendrés par la MP, sur la qualité de la voix et la parole des patients. L'effet du traitement LSVT LOUD® sur certaines variables manque toutefois d'études. Nous proposons donc d'étudier des variables acoustiques et phonétiques, et d'analyser l'effet que ce traitement a sur des patients francophones atteints de la MP. Ainsi, d'après les résultats issus des précédentes recherches, et à condition d'un suivi assidu du protocole par les patients, une prise en charge LSVT LOUD® devrait avoir un effet bénéfique sur les altérations vocales et articulatoires.

2 Expérience

12 patients atteints de la MP ont participé à notre étude : 3 femmes et 9 hommes, âgés de 49 à 80 ans. Les patients étaient admis au Centre de Rééducation Fonctionnelle de Quingey (Franche-Comté, France) pour suivre un programme de rééducation LSVT LOUD® dispensé par les deux orthophonistes de l'établissement. Tous les patients ont été enregistrés à 2 reprises avec le même protocole : quelques heures/jours avant le début de leur rééducation, puis 5 semaines plus tard, le jour de leur sortie du centre. Les enregistrements se sont déroulés de janvier à décembre 2019.

Les patients ont été invités à converser pendant 2 minutes, à lire un extrait de 39 mots issu du *Petit Prince* (Saint-Exupéry), contenant les dix voyelles orales du français, et à produire 3 /a/ tenus (le patient n°2 n'a pas accompli cette tâche). Enfin, les scores au questionnaire VHI ont été recueillis. Les données ont été enregistrées à des heures identiques entre le prétest et le post-test, en début de journée, dans un même bureau. Le signal audio a été acquis au moyen d'un microphone USB Audio-Technica AT2020 placé à environ 30 centimètres de la bouche des patients, relié à un ordinateur portable fonctionnant sur batterie. Les signaux étaient enregistrés via le logiciel ROCme! (Ferragne *et al*., 2012). Les patients étaient assis sur une chaise, face à l'écran de l'ordinateur posé sur une table.

Pour l'échantillon conversationnel, les tours de parole du patient ont été segmentés manuellement sur le logiciel Praat. Cet enregistrement nous a permis d'analyser le débit de parole moyen en nombre de syllabes par seconde, et les valeurs moyennes de la f0 en demi-tons par rapport à 1 Hz, et son écart-type. Nous avons mesuré ces mêmes valeurs de fréquences grâce à l'extrait de lecture de texte. De plus, nous avons segmenté et étiqueté manuellement les 52 voyelles du texte avec Praat, en se basant sur la présence des bandes formantiques dans le spectrogramme. L'estimation des valeurs du milieu temporel des formants a été mesurée de façon semi-automatique, après ajustement manuel des valeurs de fréquence estimées aux pics d'énergie visibles sur le spectrogramme[1]. Ainsi, nous avons représenté les enveloppes convexes des triangles vocaliques et estimé l'aire du triangle vocalique. Chaque tenue vocalique a été segmentée manuellement. La valeur moyenne du TMP a été mesurée, ainsi que les valeurs de *jitter*, *shimmer* et *HNR* à l'aide d'un script.

3 Résultats

Nous anticipions un déplacement de l'espace vocalique vers les fréquences élevées dans sa représentation conventionnelle avec axes de F1 et F2 inversés, du fait de l'augmentation du formant F1, grâce au traitement LSVT LOUD®. La FIGURE *1* illustre cette augmentation de F1. L'analyse des valeurs du F1 moyen montre un effet bénéfique significatif (t = -7,61 ; df = 11 ; p < 0,01) du traitement, entraînant la descente de l'espace vocalique.

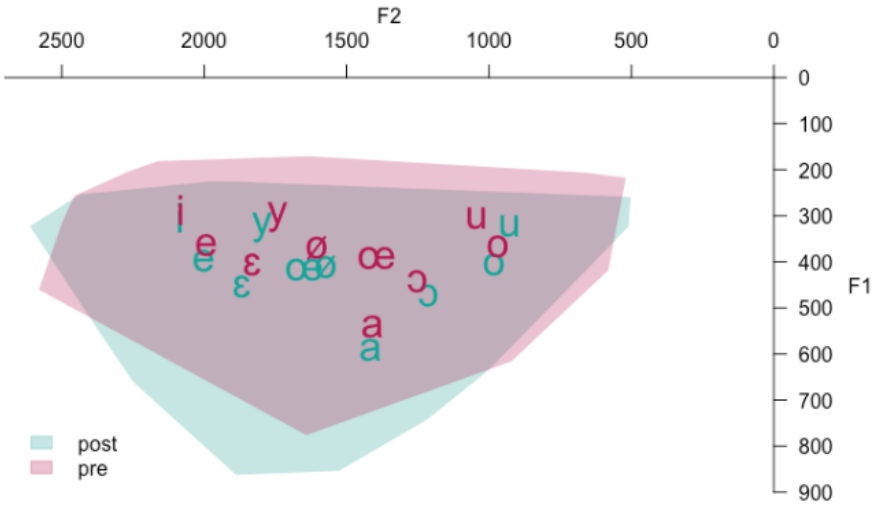


FIGURE 1 : Superposition des enveloppes convexes (Hz), selon chaque test

Nous anticipions un agrandissement de l'espace vocalique par augmentation de l'aire des enveloppes convexes. La comparaison des valeurs des aires entre pré- et post-test ne renvoie pas de résultats significatifs (t = -0,50 ; df = 11 ; p > 0,05).

1 Scripts disponibles à cette adresse : https://github.com/emmanuelferragne/CminR-Praatik .

Nous anticipions une augmentation du f0 moyen dans les deux modalités de parole (lecture et parole spontanée), suite au traitement LSVT LOUD®. Les résultats de notre étude calculés par un modèle linéaire mixte ayant comme facteurs fixes la modalité (lecture *vs* spontané) et le test (pré *vs* post), et le participant comme facteur aléatoire, montrent une augmentation significative ($p < 0,01$) de f0 en tâche de lecture, mais pas de différence significative f0 moyen en tâche de parole spontanée après le traitement. L'écart-type à la moyenne de f0 subit un effet de la parole : l'écart-type est significativement plus élevé pour la tâche de lecture ($p < 0,05$).

Concernant la qualité de voix, nous anticipions une diminution du *jitter* et du *shimmer*, et une augmentation du *HNR* suite au traitement LSVT LOUD®. Les résultats relevés par t-test appariés montrent un effet significatif bénéfique du traitement pour chacun des paramètres : *jitter* ($t = -3,44$; $df = 10$; $p < 0,01$), *shimmer* ($t = -5,81$; $df = 10$; $p < 0,01$), *HNR* ($t = 4,54$; $df = 10$; $p < 0,01$). La FIGURE *2* illustre les valeurs normalisées de ces paramètres de la qualité de voix.

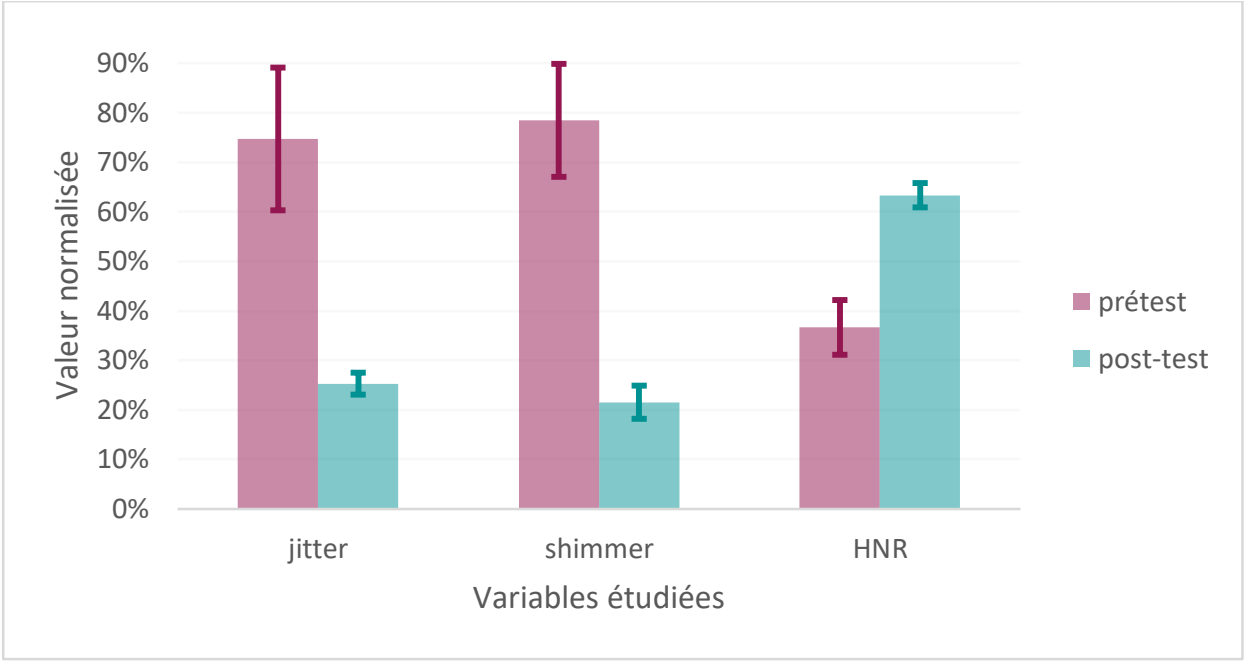


FIGURE 2 : Valeurs normalisées des paramètres de qualité de voix, selon le test

Pour étudier l'effet du traitement sur le ressenti de handicap vocal des patients, nous anticipions que le score obtenu au VHI diminuerait en post-test. Les résultats obtenus par t-test apparié évoquent un effet positif significatif du traitement LSVT LOUD® du score global ($t = -3,08$; $df = 11$; $p < 0,05$) et de chacun des sous-domaines étudiés : émotionnel ($t = -2,63$; $df = 11$; $p < 0,05$), physique ($t = -2,65$; $df = 11$; $p < 0,05$) et fonctionnel ($t = -3,71$; $df = 11$; $p < 0,01$). Ce ressenti étant propre à chaque individu, nous proposons une étude individuelle par comparaison des scores globaux : d'après Jacobson *et al.* (1997), une baisse de 18 points au score global signe une diminution bénéfique significative du ressenti de handicap, représentée en FIGURE *3*. Chaque barre fléchée indique la diminution, ou l'augmentation (n°7 et 11), du score global au VHI pour chaque patient.

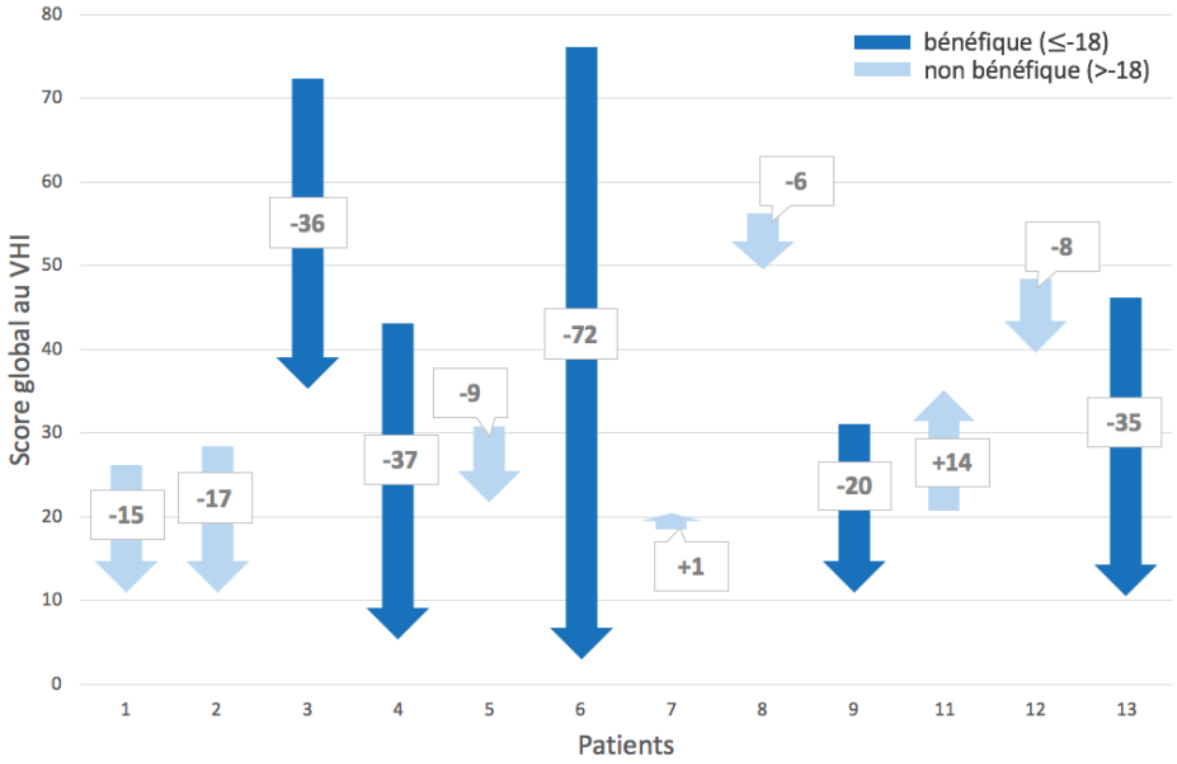


FIGURE 3 : Évolution du score global entre le post-test et le prétest, pour chaque patient

Enfin, notre étude ne met pas en évidence des effets significatifs pour l'analyse du débit (d'après un t-test, $t = 0{,}91$; $df = 454{,}34$; $p > 0{,}05$) et du TMP (d'après un t-test apparié, $t = -1{,}00$, $df = 10$, $p > 0{,}05$) suite au traitement LSVT LOUD® chez les patients étudiés. Le débit propre à chaque locuteur est globalement maintenu. Le TMP subit un effet de l'exercice : les tenues vocaliques demandées par la LSVT LOUD® imposent trois paramètres indissociables : intensité, durée et qualité. Le travail de la variation de hauteur se pratique sur des /a/ tenus, aigus et graves mais de durées limitées : 5 à 10 secondes. Lors de l'évaluation réalisée en post-test, la majorité des patients a demandé une reformulation de consigne pour le TMP : par exemple « aigu ou grave ? », ou encore « je le fais fort ? ». Ainsi, ce stimulus étant utilisé pour deux consignes différentes en rééducation, il est probable que certains patients ont hésité à faire durer leurs /a/. Cet effet de l'exercice n'est cependant pas retrouvé dans la littérature, qui fait état d'une augmentation du TMP grâce au traitement LSVT LOUD® (Ramig *et al.*, 1995).

4 Discussion

L'évaluation du handicap vocal ressenti par le patient permet, en clinique, d'adapter au mieux la prise en charge. Notre étude rapporte un effet significatif du traitement LSVT LOUD® sur le score global obtenu au VHI. Toutefois, lorsque nous comparons chaque patient à lui-même, seuls 5 patients sur les 12 ont un ressenti significativement amélioré par rapport au prétest. Ainsi, ces

résultats nous amènent à nous questionner sur la plainte vocale non systématique des patients parkinsoniens. D'une part, une évaluation de la plainte liée à la parole peut sembler plus adaptée pour certains patients. La version française du Speech Handicap Index (SHI) existe et peut être un outil fiable et sensible pour compléter l'évaluation de la plainte des patients (Degroote *et al*., 2012). D'autre part, la plainte vocale (ou de parole) peut émerger au cours de la rééducation, et contribuer à la prise de conscience des troubles.

Les paramètres objectifs de qualité de voix (*jitter*, *shimmer*, *HNR*) sont altérés dans la parole parkinsonienne (Jiménez-Jiménez *et al*., 1997 ; Yücetürk *et al*., 2002). Les résultats de notre étude montrent que ces mêmes paramètres sont significativement améliorés grâce au traitement LSVT LOUD®. Ce constat nous amène à conclure qu'une meilleure stabilité de la fréquence et de l'amplitude de vibration des cordes vocales participe à retrouver le caractère naturel de la voix, objectif premier de la prise en charge orthophonique des dysarthries.

Le groupe de patients parkinsoniens d'Audibert & Fougeron (2012) se distingue des autres sujets dysarthriques de l'étude par une réduction de F1, les voyelles étant produites de façon plus fermée. Les résultats de notre étude montrent que le traitement LSVT LOUD® permet une augmentation de F1 rapprochant ainsi la production des voyelles de celle de la parole normale. Ce constat nous amène à penser que l'entraînement permet une augmentation de la mobilité des articulateurs : l'aperture buccale s'en trouve modifiée et les voyelles sont produites de façon plus ouverte. Sapir *et al*. (2007) objectivent une augmentation de l'aire de l'espace vocalique grâce au traitement LSVT LOUD®. Les résultats de notre étude ne rapportent pas d'augmentation significative de l'aire de l'espace vocalique. Celle-ci est notable pour certains patients, mais n'est pas généralisable à l'ensemble de la population étudiée, ce qui nous amène à conclure sur le caractère nécessairement individuel de cette mesure. Il est à noter que les mesures effectuées diffèrent entre les 2 études : Sapir *et al*. (2007) utilisent seulement les voyelles /a/, /i/ et /u/ ; notre étude prend en compte l'ensemble des voyelles du système vocalique français au sein d'un texte lu, subissant alors l'effet de coarticulation évident.

D'après Ramig *et al*. (1995), le f0 moyen n'augmente pas de façon significative grâce au traitement LSVT LOUD®. Les résultats de notre étude n'apportent pas de résultats significatifs de l'évolution de f0 moyen en parole spontanée entre les tests. Sapir *et al*. (2007) évoquent une différence des résultats pour une même variable, selon la tâche proposée : le f0 moyen augmente pour la tâche de lecture, mais pas pour la tâche de parole spontanée. Notre étude retrouve cette dissociation significative entre les tâches de parole. En effet, la lecture est une tâche différente de la parole spontanée qui amène, chez tous les types de locuteurs, un comportement de performance vocale qui fait augmenter f0 (Ghio *et al*., 2014). De plus, il existe un décalage entre la perception interne qu'a le patient parkinsonien de son mouvement, et la réelle production motrice. Ainsi, l'ajustement explicite de la hauteur à la situation de lecture est plus aisé qu'un ajustement automatique dégradé par la maladie. L'étude de Ramig *et al*. (1995) évoque une augmentation significative de l'écart-type de f0 dans la tâche de monologue. Nous retrouvons cette augmentation de l'écart-type de f0, seulement sur la tâche de lecture. Ce constat peut être le résultat de la demande implicite de « mettre le ton » lors d'une lecture à voix-haute, et du passage au style direct proposé dans notre stimulus.

5 Conclusion

Les résultats présentés dans cette étude contribuent à élargir les connaissances dans le domaine de l'étude phonétique de la parole pathologique et notamment, de l'effet qu'a la LSVT LOUD® sur la parole de patients parkinsoniens francophones. Même si les différences interindividuelles sont patentes (effets de l'âge et du sexe), la position du triangle vocalique et les paramètres de qualité de voix tendent à se rapprocher de la parole normale. L'éducation thérapeutique des patients les amène à mieux se connaître, à avoir confiance en leur voix et en leur parole pour diminuer le retrait social dont souffrent les malades de Parkinson. Pour les études futures, il sera intéressant de dissocier les types de discours en tâche de lecture, d'objectiver les effets du traitement à plus long terme (+1 an, + 2 ans), et de comparer ces données à celles d'un groupe contrôle de sujets sains appariés. La plus-value de notre recherche réside donc, d'une part, dans l'étude de l'effet de la LSVT LOUD® spécifiquement sur le français et, d'autre part, dans la proposition d'un stimulus conversationnel, permettant alors une analyse plus écologique de l'utilisation de la parole par les patients.

Références

ATKINSON-CLEMENT, C., LETANNEUX, A., BAILLE, G., CUARTERO, M.-C., VERON-DELOR, L., ROBIEUX, C., BERTHELOT, M., ROBERT, D., AZULAY, J.-P., DEFEBVRE, L., FERREIRA, J., EUSEBIO, A., MOREAU, C., & PINTO, S. (2019). Psychosocial Impact of Dysarthria: The Patient-Reported Outcome as Part of the Clinical Management. *Neurodegenerative Diseases*, 1-10.

AUDIBERT, N., & FOUGERON, C. (2012). Distorsions de l'espace vocalique : Quelles mesures ? Application à la dysarthrie. *XXIXème Journées d'Études de la Parole,* Grenoble.

BRABENEC, L., MEKYSKA, J., GALAZ, Z., & REKTOROVA, I. (2017). Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission*, *124*(3), 303-334.

CANNITO, M. P., SUITER, D. M., BEVERLY, D., CHORNA, L., WOLF, T., & PFEIFFER, R. M. (2012). Sentence Intelligibility Before and After Voice Treatment in Speakers with Idiopathic Parkinson's Disease. *Journal of Voice*, *26(2)*, 214-219.

DEGROOTE, G., SIMON, J., BOREL, S., & CREVIER-BUCHMAN, L. (2012). The French version of Speech Handicap Index: Validation and comparison with the Voice Handicap Index. *Folia Phoniatrica et Logopaedica*, 20-25.

EL SHARKAWI, AE., RAMIG, LO., LOGEMANN, JA., PAULOSKI, BR., RADEMAKER, A., SMITH, C., PAWLAS, A., BAUM, S., & WERNER, C. (2002). Swallowing and voice effects of Lee Silverman Voice Treatment (LSVT®): A pilot study. *Journal of Neurology, Neurosurgery & Psychiatry*, *72*(1), 31-36.

FERRAGNE, E., FLAVIER, S., & FRESSARD, C. (2012). ROCme! : Logiciel pour l'enregistrement et la gestion de corpus oraux. *XXIXème Journées d'Études de la Parole*, Grenoble, 19-20.

GHIO, A., ROBERT, D., GRIGOLI, C., MAS, M., DE LOOZE, C., MERCIER, C., & VIALLET, F. (2014). Les anomalies de la fréquence fondamentale chez le locuteur parkinsonien : Contraste entre les effets respectifs de l'hypodopaminergie due à la maladie de Parkinson et de l'apport thérapeutique par L-Dopa. *Revue de Laryngologie Otologie Rhinologie*, *135*(2), 63-70.

HAS. (2016). *Guide parcours de soins pour la maladie de Parkinson*. https://www.has-sante.fr/portail/upload/docs/application/pdf/2012-04/guide_parcours_de_soins_parkinson.pdf
HERD, C. P., TOMLINSON, C. L., DEANE, K. H., BRADY, M. C., SMITH, C. H., SACKLEY, C. M., & CLARKE, C. E. (2012). Comparison of speech and language therapy techniques for speech problems in Parkinson's disease. *Cochrane Database of Systematic Reviews*, *Issue 8*.
HOLMES, RJ., OATES, JM., PHYLAND, DJ., & HUGHES, AJ. (2000). Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders*, *35*(3), 407-418.
Inserm. (2015). *La maladie de Parkinson*. Inserm - La science pour la santé. https://www.inserm.fr/information-en-sante/dossiers-information/parkinson-maladie
JIMENEZ-JIMENEZ, F. J., GAMBOA, J., NIETO, A., GUERRERO, J., ORTI-PAREJA, M., MOLINA, J. A., GARCIA-ALBEA, E., & COBETA, I. (1997). Acoustic voice analysis in untreated patients with Parkinson's disease. *Parkinsonism & Related Disorders*, *3*(2), 111-116.
LIOTTI, M., RAMIG, L. O., VOGEL, D., NEW, P., COOK, C. I., INGHAM, R. J., INGHAM, J. C., & FOX, P. T. (2003). Hypophonia in Parkinson's disease: Neural correlates of voice treatment revealed by PET. *Neurology*, *60*(3), 432-440.
LIU, L., JIAN, M., & GU, W. (2019). Prosodic Characteristics of Mandarin Declarative and Interrogative Utterances in Parkinson's Disease. *Interspeech 2019*, 3870-3874.
NARAYANA, S., FOX, P. T., ZHANG, W., FRANKLIN, C., ROBIN, D. A., VOGEL, D., & RAMIG, L. O. (2010). Neural correlates of efficacy of voice therapy in Parkinson's disease identified by performance–correlation analysis. *Human Brain Mapping*, *31*(2), 222-236.
RAMIG, LO., COUNTRYMAN, S., THOMPSON, L., & HORII, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson Disease. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1232-1251.
RAMIG, LO., SAPIR, S., FOX, C., & COUNTRYMAN, S. (2001). Changes in vocal loudness following intensive voice treatment (LSVT®) in individuals with Parkinson's disease: A comparison with untreated patients and normal age-matched controls. *Movement Disorders*, *16*(1), 79-83.
RAMIG, LO., HALPERN, A., SPIELMAN, J., FOX, C., & FREEMAN, K. (2018). Speech treatment in Parkinson's disease: Randomized controlled trial (RCT). *Movement Disorders*, *33*(11), 1777-1791.
ROBERT, D., & SPEZZA, M. (2005). La dysphonie parkinsonienne. In C. Özsancak & P. Auzou (Éd.), *Les troubles de la parole et de la déglutition dans la maladie de Parkinson* (p. 131-143). Solal.
SAPIR, S., SPIELMAN, JL., RAMIG, LO., STORY, BH., & FOX, C. (2007). Effects of Intensive Voice Treatment (the Lee Silverman Voice Treatment [LSVT]) on Vowel Articulation in Dysarthric Individuals with Idiopathic Parkinson Disease: Acoustic and Perceptual Findings. *Journal of Speech, Language, and Hearing Research*, *50*(4), 899-912.
SPIELMAN, J. L., BOROD, J. C., & RAMIG, L. O. (2003). The Effects of Intensive Voice Treatment on Facial Expressiveness in Parkinson Disease. *Cognitive and Behavioral Neurology*, *16*(3), 177-188.
WIGHT, S., & MILLER, N. (2015). Lee Silverman Voice Treatment for people with Parkinson's: Audit of outcomes in a routine clinic. *International Journal of Language & Communication Disorders*, *50*(2), 215-225.
YÜCETÜRK, A., YILMAZ, H., EĞRILMEZ, M., & KARACA, S. (2002). Voice analysis and videolaryngostroboscopy in patients with Parkinson's disease. *European Archives of Oto-Rhino-Laryngology*, *259*(6), 290-293.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude comparative de corrélats prosodiques de marqueurs discursifs français et anglais selon leur fonction pragmatique

Lou Lee[1,2], Denis Jouvet[2] Katarina Bartkova[1], Yvon Keromnes[1], Mathilde Dargnat[1]
(1) Université de Lorraine, CNRS, ATILF, F-54000 Nancy, France
(2) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{lou.lee;katarina.bartkova;yvon.keromnes;mathilde.dargnat}
@univ-lorraine.fr, denis.jouvet@loria.fr

RÉSUMÉ

Ce papier présente une étude des caractéristiques prosodiques de marqueurs discursifs en fonction de leur sens pragmatique. L'étude est menée sur trois marqueurs discursifs français (*alors*, *bon*, *donc*) et trois marqueurs anglais (*now*, *so*, *well*) afin de comparer leurs caractéristiques prosodiques dans ces deux langues. Plusieurs paramètres prosodiques ont été calculés sur les marqueurs discursifs, et analysés selon les fonctions pragmatiques de ceux-ci. L'analyse a été effectuée sur plusieurs centaines d'occurrences de marqueurs discursifs extraits de corpus oraux français et anglais. Les résultats montrent que certaines fonctions pragmatiques des marqueurs discursifs amènent leurs propres caractéristiques prosodiques au niveau des pauses et des mouvements de la fréquence fondamentale. On observe également que les fonctions pragmatiques similaires partagent fréquemment des caractéristiques prosodiques similaires à travers les deux langues.

ABSTRACT

Comparative study on prosodic correlates of discourse markers in French and English according to their pragmatic function

The goal of this study is to investigate the prosodic characteristics of discourse markers according to their pragmatic meaning. The paper focusses on three French discourse markers (*alors*, *bon*, *donc*) and three English markers (*now*, *so*, *well*) in order to compare their prosodic characteristics in these two languages. Several prosodic features were calculated for the discourse markers under consideration and analysed according to the pragmatic functions. Hundreds of occurrences of discourse markers were extracted from French and English speech corpora and analysed. Results show that some pragmatic functions of discourse markers bring about their own prosodic behaviour regarding pauses and movements of the fundamental frequency. Moreover, similar pragmatic functions frequently share similar prosodic characteristics.

MOTS-CLÉS : prosodie, pragmatique, marqueurs discursifs, patterns F0
KEYWORDS: prosody, pragmatics, discourse markers, F0 patterns

Volume 1 : Journées d'Études sur la Parole, pages 335–343.
HAL : hal-02798556.

1 Introduction

Les marqueurs discursifs (MD) ont été de plus en plus étudiés pendant ces dernières décennies (voir notamment Hansen, 1998 ; Schiffrin, 1987 ; Aijmer, 2002). Les marqueurs discursifs fonctionnent au niveau discursif ou situationnel et donc amènent un sens différent selon le contexte ou la situation du discours. Leur fonctionnement peut varier fortement, et pour étudier cela il faut prêter attention à leur rôle pragmatique, au-delà de leur signification littérale, ou sémantique.

La prosodie est fréquemment utilisée dans les études sur le discours ou sur la parole émotionnelle car elle permet une interprétation au-delà du sens sémantique original. Or, la prosodie des marqueurs discursifs (MD) est relativement peu étudiée, en particulier pour la langue française. La prosodie est une information très utile pour détecter les sens pragmatiques et discursifs de ces marqueurs.

Quelques travaux récents ont étudié la corrélation entre les paramètres prosodiques des marqueurs discursifs et leur usage comme MD ou non-MD ou encore leur sens pragmatique en contexte (Horne et al., 2009 ; Wichmann, 2005 ; Cabarrão et al., 2015). Selon ces études, l'information prosodique sert en effet à correctement prédire le statut du mot, c'est-à-dire, s'il est employé comme MD ou non-MD. Non seulement le statut du mot est prédictible, mais aussi le sens pragmatique plus spécifique peut aussi être marqué par la prosodie.

Dans cette étude, nous présentons une analyse prosodique sur trois marqueurs discursifs français (*alors*, *bon*, *donc*) et trois marqueurs anglais (*now*, *so*, *well*) selon leurs différentes fonctions pragmatiques. Comme nous l'avons observé dans nos études précédentes sur les marqueurs discursifs français (Bartkova et al., 2016 ; Jouvet et al., 2017 ; Lee et al., 2019), les réalisations prosodiques des marqueurs discursifs dépendent de leurs fonctions pragmatiques. La présente étude emmène nos recherches plus loin, en considérant une autre langue, l'anglais.

Nous allons d'abord présenter les données de parole utilisées et l'annotation des fonctions pragmatiques, et ensuite l'analyse prosodique : analyse de la présence de pauses avant ou après les marqueurs discursifs, et mouvements de la fréquence fondamentale (F0).

2 Corpus de parole et annotations pragmatiques

2.1 Présentation des corpus français et anglais

Les occurrences des mots étudiés, fréquemment utilisés comme marqueurs discursifs, ont été extraites aléatoirement de corpus oraux français et anglais. Les corpus français correspondent à quelques centaines d'heures de parole, et sont composés d'enregistrements de parole « préparée » provenant du corpus de l'évaluation ESTER2 (Galliano et al., 2009). Pour l'anglais, le corpus TED-LIUM 3 (Hernandez et al., 2018) a été utilisé ; il correspond à 452 heures de parole, et est composé d'enregistrement de conférences TED. Les transcriptions manuelles de ces corpus ont été alignées automatiquement avec le signal de parole pour fournir la position des mots, et pour permettre l'extraction des données de parole à étudier.

Les marqueurs discursifs étudiés sont les suivants : *alors*, *bon*, *donc* pour le français, et *now*, *so*, *well* pour l'anglais. Le choix des mots à étudier a été effectué en fonction de leur fréquence d'occurrence dans nos corpus. Le but étant d'avoir des données correspondant à un usage réel, et en quantité suffisante pour que les analyses soient pertinentes et fiables.

Pour les mots français, 200 occurrences ont été aléatoirement extraites du corpus ESTER, pour chaque mot étudié, avec un contexte large (15 mots précédant et 15 mots suivant le mot cible) pour que son sens dans le contexte soit compréhensible. Ces données ont été écoutées, et manuellement annotées, d'abord avec les étiquettes MD et non-MD, puis avec les étiquettes des fonctions pragmatiques choisies pour chaque marqueur discursif. Pour la plupart des mots, environ 70 % des occurrences ont été identifiées comme MD. Le même processus a été appliqué sur les données anglaises, mais seulement 200 occurrences ont été extraites du corpus TED-LIUM, pour chaque mot. De plus, afin que l'analyse prosodique soit la plus pertinente possible, la segmentation phonétique du mot concerné (ainsi que celle du mot précédent et du mot suivant) a été manuellement vérifiée et corrigée, si nécessaire.

2.2 Fonctions pragmatiques des marqueurs discursifs

Le poids sémantique du mot lors de son usage MD est significativement plus léger en comparaison de celui d'un usage non-MD (Maschler & Schiffrin, 2015 ; Hansen, 1998). La présence d'un marqueur discursif amène au discours un sens pragmatique, et non pas un sens littéral ou sémantique. Plusieurs sens pragmatiques peuvent être attribués à un même marqueur discursif, en fonction de son contexte d'utilisation.

Pour chaque mot étudié, les étiquettes des fonctions pragmatiques ont été choisies d'abord en fonction de la littérature (Beeching, 2007 ; Degand & Fagard, 2011 ; Denturck, 2008 ; Lefeuvre, 2011), puis elles ont été ajustées afin de couvrir toutes les fonctions pragmatiques rencontrées dans nos données de parole. Pour les données étudiées, chaque marqueur discursif est utilisé avec trois à sept fonctions pragmatiques différentes, et ces fonctions peuvent être partagées entre différents marqueurs discursifs (cf. Lee et al., 2019, pour plus d'informations sur le processus d'annotation et l'accord entre annotateurs).

Le tableau 1 présente le nombre d'occurrences correspondant à un usage comme MD, et la fréquence de chacune des fonctions pragmatiques pour chacun des marqueurs discursifs. À noter que la somme des pourcentages n'est pas égale à 100 %, car certains usages peu fréquents, ou des usages complexes (correspondant à la combinaison de plusieurs marqueurs discursifs, comme « …*alors bon*… ») ne sont pas mentionnés dans le tableau.

TABLE 1. Fonctions pragmatiques des marqueurs discursifs.

	alors	***bon***	***donc***	***now***	***so***	***well***
Nb. occurrences	*116*	*106*	*131*	*76*	*109*	*108*
Introduction	45 %	20 %	17 %	33 %	13 %	13 %
Reprise	5 %		22 %	13 %	27 %	9 %
Conclusion	9 %		40 %		39 %	
Parenthèse	16 %	26 %	19 %	54 %	21 %	24 %
Hésitation	9 %	8 %				
Reformulation	2 %	9 %				5 %
Confirmation		17 %				
Clôture		8 %				
Citation						23 %
Auto-réponse						26 %

Dans la suite de cette étude, nous nous concentrons sur les quatre fonctions pragmatiques les plus fréquentes de nos données, et présentes dans les deux langues : **Introduction** (introduire un sujet ou démarrer la parole) ; **Reprise** (reprendre la parole après une rupture de parole) ; **Conclusion** (débuter, i.e. introduire une conclusion) ; et **Parenthèse** (ajouter des informations supplémentaires ou faire un détour de parole). L'exemple (1) ci-dessous correspond à la fonction Introduction et l'exemple (2) à la fonction Parenthèse du même marqueur *so*.

(1) {need bees and they're disappearing and it's a big problem what can we do here **so** what i do is honeybee research i got my phd studying honeybee health} – fichier "NoahWilsonRich_2012X", position [465.395-475.045]

(2) {swearing and so on and then we <unk> it ended with kiss my ass and **so** basically he thought he was dealing with something smart and of course you know we} – fichier "SergeyBrin_2004", position [1122.340-1135.530]

Nous focalisons l'analyse prosodique des marqueurs discursifs sur l'étude de la présence de pauses avant et/ou après le marqueur discursif, et sur le mouvement de la fréquence fondamentale (F0) par rapport au mot précédent et au mot suivant. L'analyse est menée sur les données provenant des corpus ESTER pour le français et TED-LIUM pour l'anglais qui correspondent tous les deux à de la parole préparée, et permettent donc d'effectuer des comparaisons sur le même type de parole.

2.3 Présence de pause avant et/ou après le marqueur discursif

La présence de pauses avant ou après le marqueur, et leur position (avant ou après le mot) dans le contexte immédiat ont été analysées. Pour chaque fonction pragmatique considérée, le tableau 2 précise le nombre d'occurrences des marqueurs discursifs ainsi que la distribution des pauses autour de ces marqueurs. Quatre cas sont possibles pour les pauses : pas de pause, ni avant, ni après le mot ; pause uniquement avant le mot ; pause uniquement après le mot ; et pause avant et après le mot. Pour ne pas alourdir le tableau, nous ne présentons pas les fréquences d'une pause uniquement après le mot qui est un cas très peu fréquent.

TABLE 2. Distribution des pauses autour des marqueurs discursifs selon leur fonction pragmatique. (Les cases barrées indiquent que la fonction pragmatique n'a pas été observée pour ce mot. Les pourcentages ne sont pas indiqués lorsqu'il y a moins de 10 occurrences.)

Fonction pragmatique	**Position de pauses**	**FR - ESTER**			**EN - TEDLIUM**		
		alors	***bon***	***donc***	***now***	***so***	***well***
Non-MD	*Nb. occurrences*	*57*	*72*	*18*	*98*	*54*	*67*
	Ni avant ni après	46%	75%	67%	59%	78%	70%
	Pause avant	37%	6%	6%	8%	15%	3%
	Avant & après	9%	4%	17%	2%	--	--
Introduction	*Nb. occurrences*	*52*	*21*	*21*	*25*	*14*	*14*
	Ni avant ni après	13%	29%	81%	12%	14%	29%
	Pause avant	69%	67%	10%	52%	50%	50%
	Avant & après	13%	5%	5%	32%	29%	7%
Reprise	*Nb. occurrences*	*6*		*29*	*10*	*28*	*10*
	Ni avant ni après	--		48%	0%	4%	0%
	Pause avant	--		38%	70%	61%	60%
	Avant & après	--		14%	30%	32%	40%
Conclusion	*Nb. occurrences*	*10*	*0*	*52*		*41*	
	Ni avant ni après	40%	--	52%		24%	
	Pause avant	50%	--	38%		61%	
	Avant & après	10%	--	--		12%	
Parenthèse	*Nb. occurrences*	*19*	*28*	*25*	*41*	*23*	*26*
	Ni avant ni après	58%	39%	92%	20%	43%	15%
	Pause avant	37%	32%	4%	54%	48%	62%
	Avant & après	5%	21%	--	24%	4%	15%

Pour les occurrences de non-MD, la plupart des cas correspondent à « pas de pause, ni avant ni après » (de 46% à 78 %) alors que celles de MD montrent plus d'occurrences avec des pauses dans le contexte immédiat. Par exemple, pour la fonction pragmatique Introduction, on note une proportion élevée de présence d'une pause avant le mot, tant en français qu'en anglais (avec une exception pour le marqueur *donc*). Pour les fonctions pragmatiques Reprise et Conclusion, en comparaison de la fonction pragmatique Introduction, la présence d'une pause avant le mot est assez similaire pour les marqueurs discursifs français, et un peu plus fréquente pour les marqueurs discursifs anglais. La présence de pause avant est un peu moins fréquente dans le cas de la fonction Parenthèse, surtout pour le français. Cela est sans doute dû à la caractéristique assez informelle de cette fonction correspondant à une interruption du flux de la parole par le locuteur lui-même pour y ajouter quelque chose.

2.4 Mouvement de la fréquence fondamentale (F0)

Les valeurs de F0 ont été calculées afin de permettre l'étude du mouvement de F0 entre le marqueur discursif et son contexte immédiat. En raison des systèmes d'accentuation différents

pour ces deux langues, les positions considérées sont les suivantes. Pour le français, on considère les valeurs de F0 sur la dernière voyelle du mot précédent (w-1), sur la dernière voyelle du marqueur discursif, et sur la dernière voyelle du mot suivant (w+1). Pour l'anglais, on considère les valeurs de F0 sur la voyelle accentuée du mot précédent (w-1), sur la voyelle du marqueur discursif (à noter que les marqueurs discursifs anglais étudiés sont tous monosyllabiques), et sur la voyelle accentuée du mot suivant (w+1).

Les valeurs calculées de F0, exprimées en semi-tons (ST), ont été mesurées par rapport à la valeur médiane de F0 de chaque locuteur (De Looze & Hirst, 2010). Cela permet une meilleure interprétation et compréhension des mouvements de F0 dans l'étendue vocale théorique qui, pour les différents locuteurs, va de la valeur médiane de F0 moins 6 ST (1/2 octave en dessous), jusqu'à la valeur médiane plus 6 ST, voire plus 12 ST (c'est-à-dire de 1/2 à 1 octave au-dessus).

La quantification vectorielle a été utilisée pour effectuer un clustering des mouvements de F0, et extraire des patterns représentatifs correspondant aux centroïdes de chaque cluster. Les patterns représentatifs ont ensuite été interprétés en termes de direction du mouvement de F0 (plateau, montant, descendant) et de niveau de F0 (haut, médian, bas). Après quelques essais empiriques, le choix de 7 clusters est apparu comme un bon compromis, permettant d'éviter des patterns quasi identiques, des écarts-types trop grands caractérisant des clusters non homogènes, ou des clusters trop petits.

Pattern de F0 dans un contexte sans pause

Sept clusters ont été créés en considérant les valeurs de F0 sur le mot précédent (w-1), sur le marqueur discursif (MD), et sur le mot suivant (w+1). Pour classer les directions des mouvements de F0 en trois catégories (plateau, montant, descendant), la différence entre deux valeurs de F0 a été considérée comme suffisamment significative quand elle atteignait au moins 3 ST (Mertens & d'Alessandro, 1995).

TABLE 3. Valeurs de F0 pour les centroïdes (moyenne ± écart-type), direction du mouvement de F0, et niveau de F0 pour chaque cluster.

Cluster	F0 (w-1)	F0 (MD)	F0 (w+1)	Mouvement de F0	Niveau F0
Cluster 6 (92 items)	-0.64 ±1.29	-0.15 ±1.32	-0.21 ±1.12	plateau – plateau	médian bas
Cluster 5 (84 items)	1.62 ±1.42	0.79 ±1.40	1.81 ±1.78	plateau – plateau	médian haut
Cluster 2 (56 items)	-4.43 ±1.49	-3.05 ±2.25	-0.46 ±2.78	plateau – montant	bas
Cluster 0 (66 items)	-0.64 ±3.65	3.11 ±2.54	7.52 ±2.62	montant – montant	haut
Cluster 3 (26 items)	0.66 ±2.98	6.51 ±2.50	-0.23 ±1.66	montant – descendant	haut
Cluster 1 (60 items)	7.02 ±2.41	0.23 ±3.29	1.22 ±2.14	descendant – plateau	médian
Cluster 4 (50 items)	1.44 ±2.20	-1.16 ±1.85	-4.49 ±1.42	descendant–descendant	médian

TABLE 4. Nombre d'occurrences de chaque cluster selon les fonctions pragmatiques, et pourcentages d'items correspondant en français et en anglais. Les cas moins fréquents ont été grisés.

	Cluster 6			**Cluster 5**			**Cluster 2**			**Cluster 0**			**Cluster 3**			**Cluster 1**			**Cluster 4**		
	nb	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**
Parenthèse	11	16%	18%	15	24%	18%	11	12%	23%	7	2%	27%	3	4%	5%	2	24%	5%	8	16%	5%
Conclusion	16	28%	58%	3	6%	8%	6	16%	8%	7	16%	17%	1	3%	0%	6	16%	8%	5	16%	0%
Introduction	6	17%	13%	9	27%	13%	6	17%	13%	6	7%	50%	5	13%	13%	4	13%	0%	2	7%	0%

Le tableau 4 indique les patterns prosodiques observés pour 3 fonctions pragmatiques : par exemple, pour la fonction pragmatique Parenthèse, le pattern correspondant au cluster 6 (i.e., pattern « plateau-plateau » de niveau médian bas, d'après le tableau 3) est observé 11 fois ; et cela correspond à 16 % des occurrences de la fonction pragmatique Parenthèse sur les données françaises, et à 18 % des occurrences de cette fonction pragmatique sur les données anglaises. Ce tableau montre qu'une grande proportion de ces trois fonctions pragmatiques est réalisée sous forme « plateau-plateau » de niveau médian (clusters 6 et 5) ; cela correspond à un pourcentage allant de 26 % des occurrences (pour 'Introduction' sur les données anglaises) à 66 % des occurrences (pour 'Conclusion' sur les données anglaises), tant en français qu'en anglais. On trouve ensuite la forme « plateau-montant » (i.e., cluster 2) pour 8 % à 23 % des occurrences, tant en français qu'en anglais. Les autres patterns sont moins fréquents, et l'usage est plus spécifique à l'une ou l'autre des langues ; excepté la fonction pragmatique Conclusion pour laquelle on observe aussi un usage significatif du pattern « montant-montant » (cluster 0) dans les deux langues.

Pattern de F0 dans un contexte de pause avant

Le même traitement a été mené pour les marqueurs discursifs précédés par une pause. Sept clusters ont été créés, et le tableau 5 indique les valeurs des centroïdes, ainsi que les mouvements de F0, et niveaux de F0 correspondant.

TABLE 5. Valeurs des centroïdes (moyenne ± écart-type), direction du mouvement de F0, et niveau de F0 pour chaque cluster

Cluster	F0 (MD)	F0 (w+1)	Mouvement de F0	Niveau F0
Cluster 6 (138 items)	-0.18 ±1.08	0.02 ±1.06	plateau	médian
Cluster 4 (40 items)	-3.40 ±1.65	-2.53 ±1.79	plateau	bas
Cluster 0 (50 items)	7.26 ±2.03	9.84 ±1.97	montant léger	haut
Cluster 1 (140 items)	3.32 ±1.61	5.13 ±1.42	montant léger	haut médian
Cluster 2 (74 items)	-1.42 ±2.05	3.67 ±1.12	montant	médian
Cluster 3 (68 items)	-0.05 ±2.61	9.34 ±1.59	montant escarpé	médian
Cluster 5 (62 items)	3.97 ±2.28	-0.04 ±1.56	descendant	haut

Les tableaux 5 et 6 sont équivalents aux tableaux 3 et 4, mais fournissent les informations prosodiques dans le cas où une pause précède le marqueur discursif (et sans pause après). Les cas les plus fréquents correspondent aux clusters 1 (i.e., « montant léger » et niveau F0 haut), et 6 (i.e., « plateau » et niveau F0 médian). On observe que pour chaque langue une proportion significative des occurrences (18 % à 41 %) suit chacun de ces patterns. Les autres patterns de mouvements de F0 sont beaucoup moins utilisés.

TABLE 6. Nombre d'occurrences de chaque cluster selon les fonctions pragmatiques, et pourcentages d'items correspondant en français et en anglais. Les cas les moins fréquents ont été grisés

	Cluster 6			**Cluster 4**			**Cluster 0**			**Cluster 1**			**Cluster 2**			**Cluster 3**			**Cluster 5**		
	Nb	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**	**nb**	**Fr**	**En**
Introduction	17	19%	25%	4	6%	4%	9	6%	21%	20	27%	21%	11	19%	4%	16	17%	25%	3	6%	0%
Parenthèse	16	41%	18%	3	6%	4%	7	0%	14%	18	29%	27%	7	6%	12%	9	6%	16%	6	12%	8%
Conclusion	16	40%	25%	6	16%	8%	3	0%	13%	13	24%	29%	3	4%	8%	2	0%	8%	6	16%	8%

Il est possible de supposer que le type de parole des corpus analysés, ici, parole semi-préparée, joue un rôle dans le choix des formes des patterns prosodiques utilisés par les locuteurs. Aussi, nous prolongerons l'analyse en comparant les patterns utilisés en parole préparée à ceux utilisés en parole spontanée.

3 Conclusion

Le but de l'étude était l'analyse des caractéristiques prosodiques des marqueurs discursifs selon leur fonction pragmatique dans deux langues, le français et l'anglais. Les résultats montrent que certaines fonctions pragmatiques des marqueurs discursifs amènent leurs propres caractéristiques prosodiques au niveau de la présence de pauses avant ou après le mot, et au niveau des mouvements de F0. Notamment, les fonctions pragmatiques similaires partagent des caractéristiques prosodiques similaires dans les deux langues, quant à la distribution des pauses dans le contexte immédiat : les trois fonctions pragmatiques Introduction, Reprise et Conclusion sont le plus fréquemment précédées par une pause avant alors que la fonction Parenthèse est plus fréquemment prononcée sans aucune pause ni avant, ni après. Cependant, en ce qui concerne les mouvements de F0 autour des marqueurs discursifs, la plupart de ces patterns se situaient au niveau médian de F0 du locuteur ayant le plus souvent une forme de plateau, ce qui peut être expliqué par la caractéristique pragmatique même des marqueurs dont la fonction est de signaler une introduction de quelque chose de plus important que le marqueur lui-même.

Remerciements

Ce travail a été partiellement financé par le CPER LCHN (Contrat Plan Etat Région "Langues, Connaissances et Humanités Numériques").

Références

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus* (Vol. 10). John Benjamins Publishing.

Bartkova, K., Bastien, A., & Dargnat, M. (2016). How to be a Discourse Particle?.

Beeching, K. (2007). La co-variation des marqueurs discursifs *bon*, *c'est-à-dire*, *enfin*, *hein*, *quand même*, *quoi* et *si vous voulez* : une question d'identité ?. *Langue française*, (2), 78-93.

Cabarrão, V., Moniz, H., Ferreira, J., Batista, F., Trancoso, I., Mata, A. I., & Curto, S. (2015). Prosodic classification of discourse markers. In *International Congress of Phonetic Sciences (ICPhS 2015)*. International Phonetic Association.

De Looze, C., & Hirst, D. (2010). L'échelle OME (Octave-MEdiane) : une échelle naturelle pour la mélodie de la parole. *Proceedings of the XXVIIIème Journées d'Étude Sur La Parole (JEP 2010), Mons, Belgium*.

Degand, L., & Fagard, B. (2011). *Alors* between discourse and grammar: the role of syntactic position. *Functions of language*, *18*(1), 29-56.

Denturck, E. (2008). Étude des marqueurs discursifs. *L'exemple de* quoi. *Faculteit Taal-en Letterkunde, Sectie*, *2*, 2007-2008.

Galliano, S., Gravier, G., & Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Hansen, M. B. M. (1998). *The function of discourse particles: A study with special reference to spoken standard French* (Vol. 53). John Benjamins Publishing.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., & Estève, Y. (2018). TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer* (pp. 198-208). Springer, Cham.

Horne, M., Hansson, P., Bruce, G., Frid, J., & Filipsson, M. (2009). Discourse markers and the segmentation of spontaneous speech–The case of Swedish men '*but/and/so*'. *Lund Working Papers in Linguistics*, *47*, 123-139.

Jouvet, D., Bartkova, K., Dargnat, M., & Lee, L. (2017). Analysis and automatic classification of some discourse particles on a large set of French spoken corpora. In *International Conference on Statistical Language and Speech Processing* (pp. 32-43). Springer, Cham.

Lee, L., Bartkova, K., Jouvet, D., Dargnat, M., & Keromnes, Y. (2019). Can prosody meet pragmatics? Case of discourse particles in French.

Lefeuvre, F. (2011). *Bon* et *quoi* à l'oral : marqueurs d'ouverture et de fermeture d'unités syntaxiques à l'oral. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (64-65), 223-240.

Maschler, Y., & Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, *1*, 189-221.

Mertens, P., & d'Alessandro, C. (1995). Pitch contour stylization using a tonal perception model. In *Proc. 13th International Congress of Phonetic Sciences* (Vol. 4, pp. 228-231).

Schiffrin, D. (1987). *Discourse markers* (No. 5). Cambridge University Press.

Wichmann, A. (2005). Prosody and discourse: a diachronic approach. In *Actes de la conférence Interfaces Discours Prosodie (IDP), Aix en Provence* (pp. 1-11).

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Phénomènes de proéminence dans les subordonnées en conversation spontanée

Manon Lelandais[1]
(1) EA 4398 SeSyLIA, 5 rue de l'Ecole de Médecine, 75006 Paris, France
manon.lelandais@sorbonne-nouvelle.fr

RÉSUMÉ

À partir d'un corpus vidéo de conversation spontanée en anglais britannique, cette étude a pour but de déterminer si deux différents types syntaxiques de constructions subordonnées expriment la même absence de proéminence, dans le cadre d'une analyse multimodale. En syntaxe, les subordonnées sont décrites comme des structures dépendantes qui précisent l'information de premier plan dans le discours. Alors que de nombreux travaux étudient leurs caractéristiques verbales, peu d'études s'attachent à décrire l'articulation entre les différentes modalités communicatives lors de leur production, et à fournir une vision plus nuancée de leur apport informationnel. Nous prenons en compte un ensemble de variables appartenant à plusieurs modalités, considérées comme des facteurs de proéminence. Notre étude montre que les subordonnées ne fournissent pas les mêmes types de proéminence en fonction de leur type syntaxique, et que leur création s'appuie majoritairement sur des indices de nature intonative et gestuelle plutôt que syntaxique.

ABSTRACT

Prominence phenomena in subordinate constructions in conversational speech
Based on a video recording of conversational British English, this paper tests within the framework of a multimodal analysis whether two different syntactic subordinate constructions express the same absence of prominence in terms of informational content. Subordinate constructions have been described in syntax as dependent structures elaborating on primary elements of discourse. Although their verbal characteristics have been deeply analyzed, few studies have focused on the articulation of the different communicative modalities in their production or provided a qualified picture of their informational input. Our study includes a range of variables in different modalities, regarded as cues for prominence. We show that subordinate constructions express different types of prominence depending on their syntactic type, and that the expression of prominence mainly relies on intonational and gestural cues rather than on syntactic cues.

MOTS-CLÉS : proéminence, subordination, structure informationnelle, multimodalité.
KEYWORDS: prominence, subordination, informational structure, multimodality.

1 Introduction

Cet article porte sur la subordination en conversation spontanée, et s'intéresse plus précisément aux séquences discursives contenant une construction subordonnée. Dans la littérature syntaxique, les

Volume 1 : Journées d'Études sur la Parole, pages 344–352.
HAL : hal-02798557.

subordonnées sont souvent décrites comme des structures qui modifient ou spécifient des éléments primaires, associées à un autre contenu propositionnel dans la structure hôte (Huddleston & Pullum, 2006). Les subordonnées fournissent une caractérisation sémantique plus poussée ou des détails informationnels à propos du référent ou de l'état de faits qu'elles modifient (*ibid.*). Cet article se concentre sur les deux types syntaxiques de propositions subordonnées finies les plus fréquentes dans notre corpus conversationnel (décrit dans la section 3.2) : les circonstancielles temporelles et les relatives déterminatives.

Les circonstancielles temporelles modifient un syntagme verbal ou une proposition entière. Elles restreignent le cadre temporel dans lequel les éléments référentiels doivent être traités (Langacker, 2008). Dans *she was an airforce pilot when she was young*, la subordonnée spécifie les circonstances dans lesquelles *she was an airforce pilot* est valide en tant qu'énoncé. Au contraire, les relatives déterminatives modifient une expression nominale, en affinant l'identification du référent de cette expression (*ibid.*). Dans *the Spanish girls that were there on our second class* la relative déterminative *that were there on our second class* augmente la pertinence de *the Spanish girls*, en créant une sous-catégorie pour ce référent.

Si les subordonnées sont définies comme véhiculant de l'information d'arrière-plan dans le discours (Longacre, 1996), la littérature montre très peu de consensus quant à leur poids communicatif. Alors que certains travaux les décrivent comme des phénomènes d'ancrage à la fois grammatical et discursif (Fleischman, 1985), d'autres remettent en cause la correspondance systématique entre subordination syntaxique, sémantique, et discursive (Langacker, 2008). Cette étude a donc pour but de déterminer si deux différents types syntaxiques de subordonnées montrent la même proéminence. Notre hypothèse principale se base sur la capacité de ces constructions à montrer différents types de proéminence en fonction de leur type syntaxique. Le terme de proéminence est ici utilisé en tant que synonyme de focalisation, et est défini comme un phénomène lié à la structure informationnelle du discours (Longacre, 1996). Il désigne du point de vue de la production linguistique un effort communicatif de la part du locuteur, dont le résultat est une configuration syntaxique, prosodique, et/ou gestuelle particulière dans laquelle un élément de surface est censé être perçu comme démarqué des autres.

En conversation spontanée, les ressources linguistiques de la parole communiquent avec des modalités kinétiques, comme la direction du regard, les mouvements de sourcils et de tête, ainsi que les gestes manuels. Ces modalités ne fonctionnent pas indépendamment les unes des autres, bien que l'une d'elles puisse être plus proéminente à un moment donné (Norris, 2004). Comparé au grand nombre d'études sur la subordination du point de vue de la syntaxe et de la prosodie, la contribution de la gestualité est souvent négligée. Or, le développement d'outils et de procédés analytiques pouvant être mis en relation facilite la description des subordonnées en tant que phénomènes multimodaux.

2 Cadre théorique

Dans la grammaire catégorielle traditionnelle[1], les deux subordonnées sous étude sont utiles sémantiquement sans pour autant représenter des éléments constitutifs (Lehmann, 1988). Cependant, cette classification a été décrite comme imprécise pour analyser la parole spontanée,

[1] Cet article ne prend pas en compte la subordination discursive, ni les indices de proéminence au niveau du discours, de façon à ne pas multiplier les paramètres. Ils sont néanmoins intégrés dans un travail de plus grande ampleur. De même, l'état de l'art concerne uniquement l'anglais.

particulièrement quant à la nature des éléments introducteurs (Haiman & Thompson, 1984). Une hiérarchie de relations est suggérée pour évaluer le poids syntactico-sémantique des constituants. Les propositions contiennent un noyau essentiel (contenant un évènement, processus, ou état et ses compléments), ainsi qu'une périphérie optionnelle (correspondant à la localisation ou l'environnement; *ibid.*). Alors que la transitivité et la dynamicité dans les composants nucléaires sont associées aux informations de premier plan (Longacre, 1996), la nominalité et les phénomènes d'identification sont associés à un poids sémantique inférieur (*ibid.*).

La subordination prosodique est essentiellement réalisée par l'intonation (Bolinger, 1984). Au long d'un paragraphe vocal, la hauteur intonative (*i.e.* F0, ou fréquence fondamentale) décline naturellement de manière progressive. Un ton abaissé par rapport à un ton haut précédent correspond à la relation neutre entre deux groupes prosodiques (*ibid*). Une F0 haute sur la syllabe accentuée d'un item lexical véhicule de l'information nouvelle dans le discours (Baumann & Grice, 2006). De plus, alors que l'emphase peut être créée à l'aide d'un contour montant-descendant sur la syllabe nucléaire, des contours plats ou descendants-montants sont utilisés pour véhiculer de l'information d'arrière-plan (Ward & Hirschberg, 1984). Les subordonnées sont typiquement moins modulées (*i.e.* montrant moins de mouvement intonatif) que leur co-texte (Baumann & Grice, 2006). L'emphase peut également être réalisée à l'aide d'un allongement syllabique, alors que la compression des syllabes signale de l'information d'arrière-plan (Wells, 2006).

En ce qui concerne la gestualité, cette étude inclut des mouvements kinétiques co-verbaux considérés comme partie des énoncés, selon la définition de Kendon (2004). Nous intégrons la direction du regard, les mouvements de tête, ainsi que les gestes manuels. Représenter des référents par des gestes manuels est un processus cumulatif, souvent réalisé à travers une série de plusieurs unités gestuelles (*ibid.*). La création de la cohérence discursive est donc assurée par des traits gestuels (Hoetjes *et al.*, 2015), par exemple par le biais de répétitions formelles. Lorsqu'inscrits dans une continuité, les gestes manuels encodant un même référent sont plus schématiques, alors que ceux qui véhiculent de l'information nouvelle sont plus précis et plus clairs (Kita *et al.*, 1998). Les gestes abstraits qui organisent le discours sont donc traditionnellement associés à l'arrière-plan du discours et à la subordination (Cassell & McNeill, 1990), alors que les gestes représentationnels comme les iconiques sont associés au premier plan (*ibid.).* De même, la direction du regard du locuteur quitte généralement le co-locuteur pour l'élaboration du discours une fois le tour de parole pris et sécurisé (Holler *et al.*, 2014). Un changement de direction du regard vers le co-locuteur en plein tour est souvent lié à la focalisation, fonctionnant comme un appel au co-locuteur (*ibid.*). Un changement de direction du regard vers un objet peut également fonctionner en tant que geste déictique (*ibid.*). Les battements de tête isolés (*i.e.* de brefs mouvements de menton vers le bas) peuvent aussi mettre en avant des entités du discours en participant à la focalisation (Granström & House, 2005).

3 Corpus et méthodologie

3.1 Hypothèses

En fonction du cadre théorique défini par la littérature, une liste spécifique d'indices syntaxiques, vocaux, et gestuels de proéminence a été prise en compte. Il est attendu que le nombre et la nature de ces indices de proéminence (variable dépendante) varient en fonction du type syntaxique de subordination (variable indépendante; 2 modalités : circonstancielles, déterminatives). Ces indices sont précisément, pour la modalité syntaxique, la transitivité (Longacre, 1996) et la présence

d'aspect perfectif et d'auxiliaires modaux (*ibid.*). Les indices de proéminence de la modalité vocale sont l'allongement syllabique, les contours montants-descendants (Ward & Hirschberg, 1984), la modulation intonative (Baumann & Grice, 2006), et l'utilisation de la plage intonative haute avec des contours hauts (*ibid.*). Enfin, nous recensons dans la modalité gestuelle la présence de gestes manuels représentationnels (Cassell & McNeill, 1990), les changements de direction du regard vers le co-locuteur (Holler *et al.*, 2014), les battements manuels et de tête (Granström & House, 2005).

3.2 Annotation du corpus

Afin de vérifier ces hypothèses sur de l'anglais conversationnel, nous avons utilisé le corpus ENVID[2], dont les annotations déjà réalisées sont décrites dans Lelandais & Ferré (2016). Ce corpus a d'abord été transcrit orthographiquement sous Praat (Boersma & Weeninck, 2013). Les subordonnées sous étude ont été localisées et codées sur une piste séparée. Toutes les annotations faites dans Praat ont ensuite été exportées dans ELAN (Sloetjes & Wittenburg, 2008), un outil d'annotation vidéo facilitant la mise en commun et le traitement des données.

Un total de 110 constructions ont été annotées dans le corpus, classifiées selon leur type syntaxique sur une piste dans Praat (55 restrictives relatives, 55 circonstancielles; accord entre deux codeurs experts = 100%). La sélection a porté sur les subordonnées entourées par un co-texte gauche et droit du même locuteur, autre qu'une pause silencieuse cédant le tour de parole. Une seconde piste d'annotation délimite leur environnement : le groupe intonatif précédent est étiqueté L (*left co-text*), et le groupe intonatif suivant est étiqueté R (*right co-text*).

Le corpus a été segmenté en groupes intonatifs selon la British School of Intonation (Wells, 2006). L'algorithme Momel-Intsint (Hirst, 2007; Bigi, 2012) a également été utilisé pour l'annotation automatique des points cibles de F0 dans le signal. La nature de chaque contour nucléaire a été codée manuellement autour de chaque séquence sous étude (descendant; montant; montant-descendant; descendant-montant; plat; accord entre deux codeurs experts = 81.9%). Le registre intonatif a été annoté en fonction de la gamme intonative de chaque locuteur (haut; moyen; bas) à la fois sur les segments entiers, et sur les syllabes initiales et finales de chaque segment. L'allongement syllabique est également annoté sur chaque syllabe finale des séquences sous étude, défini selon 2 paliers d'allongement.

Les gestes ont été codés manuellement sous ELAN. Chaque unité gestuelle commence à l'ébauche du geste et finit au retour à la position de repos (Kendon, 2004). Dans le cas de deux gestes consécutifs, la première unité gestuelle finit à un changement de trajectoire/configuration des articulateurs. Les mouvements de tête ont été annotés en tant que *nods* (acquiescement de haut en bas), *shakes* (mouvement d'un côté de l'autre sur un axe horizontal), *tilts* (inclinaison de la tête d'un côté), *beats* (bref mouvement vers le bas) et *jerks* (mouvement du menton vers le haut; accord entre deux codeurs experts = 81.3%). Sur des pistes séparées, la direction du regard a été annotée soit vers le co-participant, soit ailleurs (accord entre codeurs = 100%). Les gestes manuels ont été annotés selon leur lien avec la parole et leur relation aux affiliés lexicaux, selon la typologie de McNeill (2005), qui distingue entre les emblèmes, les iconiques, les métaphoriques, les pointages, les battements, et les adaptateurs (accord entre codeurs = 72.1%). La fonction des gestes est annotée sur une autre piste, et différencie les gestes représentationnels des gestes organisationnels (*ibid.*; accord entre codeurs = 84.9%). Ces deux fonctions n'étant pas contradictoires, un seul geste peut être annoté comme remplissant les deux à la fois.

[2] Ce corpus contient 5 dialogues entre des locuteurs britanniques natifs (10). Les participants sont amis et ont reçu pour seule consigne de parler aussi librement que possible.

3.3 Analyses

Nous utilisons une série de modèles linéaires mixtes généralisés (*GLMMs* estimés par la méthode du maximum de vraisemblance) réalisée avec le logiciel R 3. 4. 0 (R Core Team, 2017), et le *package* lme4 (Bates *et al.*, 2017), afin d'expliquer l'effet de chaque type syntaxique sur les indices de proéminence. Du fait de la variation entre les locuteurs et les dialogues dans la production des subordonnées, nous avons systématiquement inclus Locuteur et Dialogue en tant que facteurs aléatoires dans le modèle.

4 Résultats et discussion

4.1 Circonstancielles

Aucun effet du type syntaxique de subordonnée n'a été trouvé sur les indices syntaxiques de proéminence, *i.e.* la présence de transitivité et la présence d'aspect perfectif, en ce qui concerne les circonstancielles. En revanche, des effets ont été trouvés sur les indices vocaux et gestuels. Le premier de ces effets porte sur la présence de contours hauts et montants ($\beta = 1.63$, SE = .58, $p < .05$). Ces contours font écho à l'important rôle d'organisation textuelle accompli par ce type syntaxique. L'exemple (1) associé à la Figure 1 illustrent l'association entre les circonstancielles et ces contours. Le symbole # représente une pause silencieuse, et (h) représente une reprise de souffle audible.

(1) Alex L (h) you know
SC **when you're not allowed to laugh #**
R (h) and then there's like a massive silence #

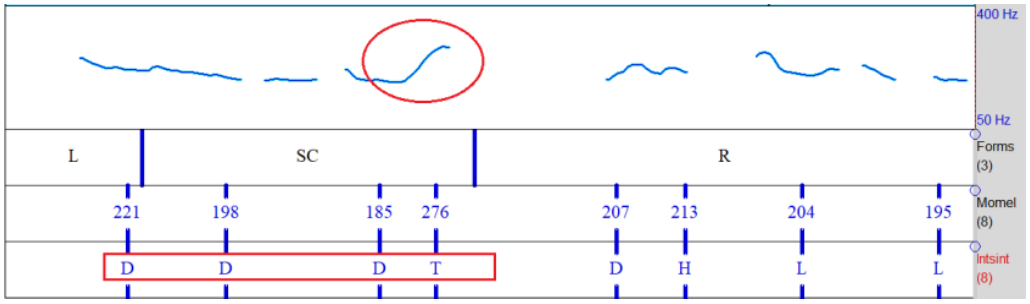


FIGURE 1: Contour intonatif de l'exemple (1) dans Praat. La seconde piste d'annotation montre les valeurs F0 sur chaque syllabe cible tandis que la dernière représente les valeurs Intsint (D pour abaissement - *downstep*, T pour gamme haute - *top of speaker's range*)

Alors que la subordonnée (SC) et son co-texte gauche (L) partagent un même groupe intonatif, la syllabe finale de SC est plus haute que l'initiale (276 Hz vs. 221 Hz). La syllabe initiale du co-texte droit (R) est quant à elle abaissée (207 Hz, valeur Intsint "D"). La circonstancielle représente un repère, et comprend les groupes intonatifs suivants.

Par ailleurs, un effet du type syntaxique a également été trouvé sur l'augmentation des gestes représentationnels par rapport au co-texte. Cet effet est illustré par la Figure 2, associée à l'exemple (2). L'activité gestuelle est représentée entre crochets.

(2) Rhianna L i tried [(a) driving once] in her car
SC when we were on a # [(b) little road in the countryside] #
R [(c) and hem (swallows) she said turn left]

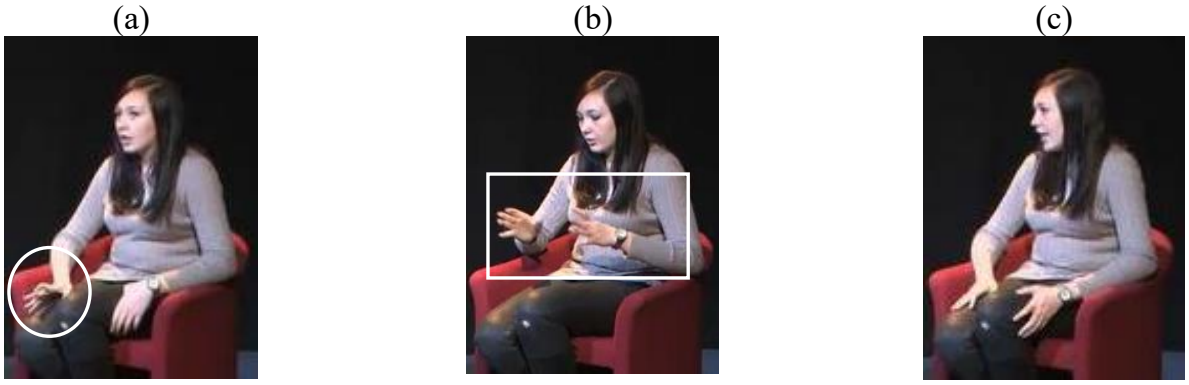


FIGURE 2*:* Réalisation gestuelle de la séquence (2), où (a), (b), et (c) représentent différents moments de sa production. Un geste représentationnel est réalisé sur (b)

Dans l'exemple (2), Rhianna produit un geste métaphorique en périphérie basse (a), insistant sur le caractère exceptionnel de la situation qu'elle décrit. Elle produit ensuite un large iconique (b) en co-occurrence avec la circonstancielle, ses deux mains dessinant des lignes parallèles devant elle, en représentation de la route (*little road*). Le regard de Rhianna sur son geste iconique possède une valeur déictique. Rhianna produit également un battement de tête en co-occurrence avec son geste manuel. Le groupe intonatif qui suit est accompagné d'un retour à la position de repos (c). Comme dans cet exemple (2), un effet du type syntaxique est également trouvé sur la production de battements de tête dans le cas des circonstancielles (ß = 0.98, SE = .05, $p < 0.05$). Ces battements de tête font écho, à la manière des tons montants, au rôle d'organisation textuelle tenu par les circonstancielles.

4.2 Relatives déterminatives

Dans le cas des relatives déterminatives, le type syntaxique a un effet sur quatre indices de proéminence. Le premier de ces effets porte sur la présence de marques aspectuelles ou de modaux (ß = -2.02, SE = .8, $p < 0.02$), comme il est visible dans l'exemple (3).

(3) Michelle L (h) and so she # disowned everything
SC that # she could associate my nana #
R for example her accent # and #

L'emploi de la forme désactualisée de *can* dans l'exemple (3) évoque une possibilité révolue. Le prédicat apporte une information nouvelle à la co-locutrice et permet d'établir un focus plus étroit que celui de L. Cette relative permet d'apporter des traits sémantiques nécessaires à l'identification du référent tout en introduisant de la scalarité.

Au niveau vocal, les relatives montrent davantage d'allongement syllabique sur leur syllabe finale que sur celle des circonstancielles (ß = 0.38, SE = .08, $p < 0.0001$) et que sur celles du co-texte (durée phonémique moyenne de la syllabe finale de SC de 0.095 secondes vs. 0.069 secondes sur la finale de L), bien que cet indice soit également un indice de frontière. Un effet du type syntaxique a également été trouvé sur les contours montants-descendants dans le cas des déterminatives ($p > .05$).

Moins de ces contours sont produits dans les circonstancielles (ß = -1.47, SE = .57, $p < .05$). L'exemple (4) associé à la Figure 3 montre ces caractéristiques.

(4) Zoe L (laughs) but how do we explain to people
SC **that obviously use computers a lot more than me #**
R it's not very good

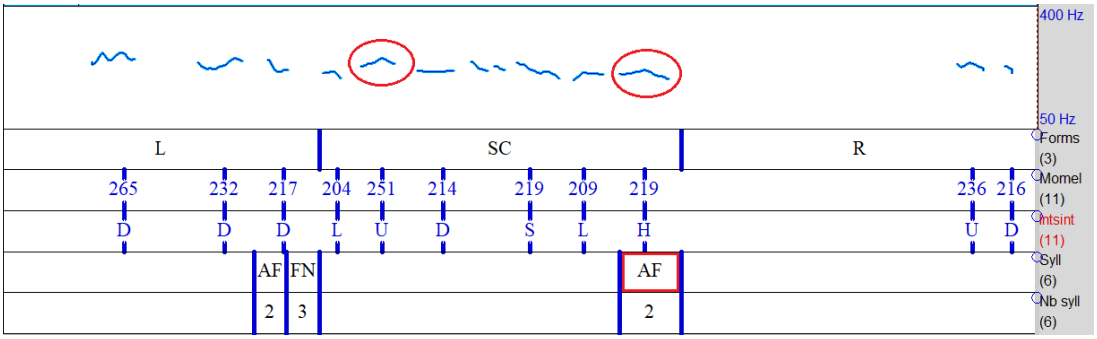


FIGURE 3: Contour intonatif de l'exemple (4) sous Praat, dans lequel la relative montre un allongement syllabique important (AF = syllabe accentuée finale composée de 2 phonèmes)

Dans l'exemple (4), Zoe exprime son angoisse à l'idée d'utiliser du matériel informatique qu'elle ne connaît pas. Elle utilise deux contours montants-descendants sur les syllabes accentuées de *obviously* et *me*, qui soulignent dans un usage contrastif l'écart de connaissances entre elle et ses étudiants. La syllable accentuée finale (AF) de la relative, *me,* est également significativement allongée par rapport à la syllable accentuée finale de L (AF), la première de *people*. Elles comportent pourtant le même nombre de phonèmes (2).

Enfin, un effet a également été trouvé sur la production de battements manuels (ß = 1.29, SE = .32, $p = .0001$) dans le cas des relatives déterminatives. Moins de battements manuels sont également produits dans le co-texte (L : ß = -1.12, SE = .4, $p < .01$; R : ß = -1.22, SE = 5, $p = .01$). L'exemple (5) ainsi que la Figure 4 illustrent cette association.

(5) Tom L like lead to the [(a) cells]
SC **that [(b) deve] lop [(c) cancer]**
R [(d) but # and again #
why not

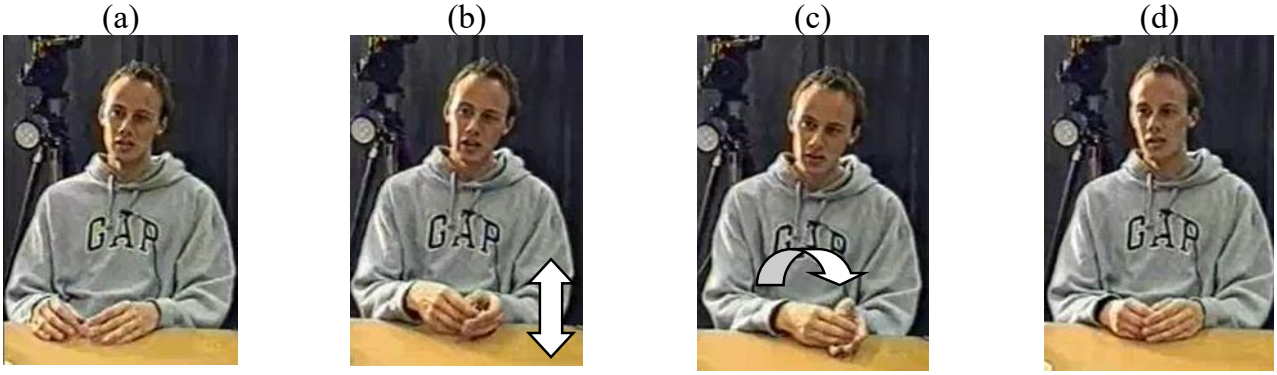


FIGURE 4: Deux battements manuels successifs produits en co-occurrence avec la relative déterminative dans l'exemple (5)

La relative déterminative se démarque du reste de la séquence par ses deux battements manuels successifs (b) et (c). L'antécédent de la relative (the *cells*) est redéfini avec l'attribution d'une propriété. Alors que le premier battement est réalisé sur un axe vertical au centre de l'espace gestuel du locuteur, le second est placé plus loin sur le côté gauche du locuteur. Leur succession en co-occurrence avec le verbe de procès et son complément d'objet direct met en avant le prédicat et le processus que ce dernier décrit, tout en indiquant pragmatiquement le contenu informationnel le plus pertinent de la séquence. Tom maintient son regard vers le co-locuteur tout au long de la séquence. La configuration du dernier battement est tenue jusqu'à la fin de la production de la relative, à la suite de laquelle Tom retourne à une position de repos (d).

5 Conclusion

Notre analyse montre que les différents types syntaxiques de constructions subordonnées peuvent être différenciés selon leurs indices de proéminence. Les circonstancielles sont caractérisées par des contours intonatifs hauts et montants, une augmentation des gestes représentationnels par rapport à leur co-texte, ainsi que des battements de tête. Ces indices de proéminence montrent majoritairement le rôle des circonstancielles en tant que repères dans l'organisation du discours. Les relatives déterminatives comptent d'avantage d'indices de proéminence que les circonstancielles. Elles sont associées à des marques aspectuelles et de modaux, un allongement syllabique significatif, des contours intonatifs montants-descendants, et des battements manuels. Prosodie et gestualité exploitent donc des dimensions communes, à des intensités et sur des temporalités différentes le long d'une séquence contenant une subordonnée.

Cette étude se concentre sur les indices de proéminence mobilisés par les locuteurs lors de la production des subordonnées. Nous travaillons à la construction d'un test de perception de la proéminence dans les subordonnées, afin de vérifier si différents types de subordonnées sont également différenciables, et si tous les indices mentionnés ont le même poids dans la perception d'une proéminence.

Remerciements

Je remercie deux relecteurs anonymes pour leurs conseils constructifs sur une version précédente de cet article.

Références

BATES, D., MAECHLER, M., BOLKER, B. & WALKER, S. (2017). Linear mixed-effects models using eigen and s4. Consulté le 29 juin 2017 sur http://cran.r-project.org/

BAUMANN, S. & GRICE, M. (2006). The intonation of accessibility. *Journal of Pragmatics*, 38, 1636–1647.

BIGI, B. (2012). SPPAS: a tool for the phonetic segmentation of speech. In *Proceedings of LREC 2012*, Istanbul, p. 1748–1755.

BOERSMA, P. & WEENINCK, D. (2013). Praat: Doing Phonetics by Computer. Consulté le 30 janvier 2013 sur http://www.fon.hum.uva.nl/praat/

BOLINGER, D. (1984). Intonational signals of subordination. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, p. 401–423, Berkeley, USA: eLanguage.

CASSELL, J. & MCNEILL, D. (1990). Gesture and ground. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, p. 57–68, Berkeley, USA: eLanguage.

FLEISCHMAN, S. (1985). Discourse functions of tense-aspect oppositions in narrative: Toward a theory of grounding. *Linguistics*, 23, 851–882.

GRANSTRÖM, B. & HOUSE, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46(3), 473–484.

HAIMAN, J. & THOMPSON, S. (1984). "Subordination" in Universal Grammar. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, p. 510–523, Berkeley, USA: eLanguage.

HIRST, D. (2007). A Praat Plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the 16th International Conference of Phonetic Sciences*, p. 1233–1236, Sarrebruck, Allemagne: Université de Sarrebruck.

HOETJES, M., KOOLEN, R., GOUDBEEK, M., KRAHMER, E. & SWERTS, M. (2015). Reduction in gesture during the production of repeated reference. *Journal of Memory and Language*, 79, 1–17.

HOLLER, J., SCHUBOTZ, L., KELLY, S., HAGOORT, P., SCHUETZE, M. & ÖZYÜREK, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, 133(3), 692–697.

HUDDLESTON, R. & PULLUM, G. (2006). Coordination and subordination. In B. AARTS & A. MCMAHON, Éds, *The Handbook of English Linguistics*, p. 198–219. Oxford: Blackwell.

KENDON, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

KITA, S., VAN GIJN, I. & VAN DER HULST, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Proceedings of the Gesture and Sign Language in Human-Computer Interaction International Gesture Workshop*, p. 23–35, Bielefeld, Allemagne.

LANGACKER, R. (2008). *Cognitive Grammar: a basic introduction*. Oxford: Oxford University Press.

LEHMANN, C. (1988). Towards a typology of clause linkage. In J. HAIMAN & S. THOMPSON, Éds, *Clause Combining in Grammar and Discourse*, p. 181–225, Amsterdam: Academic Press.

LELANDAIS, M. & FERRÉ, G. (2016). Prosodic boundaries in subordinate syntactic constructions. In *Proceedings of Speech Prosody 2016*, p. 183–187, Boston, USA: ISCA.

LONGACRE, R. (1996). *The Grammar of Discourse*. New York: Springer.

MCNEILL, D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

NORRIS, S. (2004). *Analyzing Multimodal Interaction: a methodological framework*. Londres: Routledge.

R CORE TEAM. (2017). A language and environment for statistical computing. Consulté le 29 juin 2017, sur http://www.r-project.org/

SLOETJES, H. & WITTENBURG, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of LREC 2008*, Marrakech, Maroc. http://www.lat-mpi.eu/tools/elan/

WARD, G. & HIRSCHBERG, J. (1985). Implicating uncertainty: the pragmatics of fall-rise intonation. *Language*, 61(4), 747–776.

WELLS, J. (2006). *English Intonation: an introduction*. Cambridge: Cambridge University Press.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Une base de données de phrases en français pour l'étude du rôle conjoint des incertitudes sémantique et acoustique dans la perception de la parole.

Loriane Leprieur[1] Olivier Crouzet[1,2] Étienne Gaudrain[2,3]
(1) Laboratoire de Linguistique de Nantes - LLING / UMR6310 Université de Nantes - CNRS
chemin de la Censive et du Tertre, 44312 Nantes Cedex, France
(2) ENT Department - University Medical Center Groningen, Rijksuniversität Groningen, Pays-Bas
(3) Centre de recherche en Neurosciences de Lyon - CNRS UMR 5292- CNRL Inserm U1028
CH Le Vinatier - Bâtiment 452, 95 bd Pinel, 69675 Bron Cedex
loriane.leprieur@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr,
etienne.gaudrain@cnrs.fr

RÉSUMÉ

Les effets de contexte dans la perception de la parole reposent aussi bien sur des sources acoustiques que sémantiques. Le contexte acoustique fournit des informations essentielles pour l'adaptation au locuteur et aux variations dialectales. En parallèle, le contexte sémantique contribue à prédire un ensemble de mots éligibles pour une interprétation licite des énoncés. Afin d'étudier plus précisément les interactions entre ces effets de contexte, nous avons créé une base de données de phrases courtes conçues pour observer ces phénomènes dans des protocoles expérimentaux. Cette base de données est constituée de 28 triplets de phrases porteuses terminées par des cibles de paires minimales de mots CV ou CVC, autour de voyelles acoustiquement proches associées à 4 contrastes vocaliques. Afin d'évaluer la validité des 3 catégories de contexte sémantique considérées, des mesures de similarité sémantique et de fréquence lexicale ont été réalisées à partir de différents corpus de langue française.

ABSTRACT

A dataset of french sentences to study the joint roles of semantic and acoustic uncertainty in speech perception.

Context effects in speech perception rely on both acoustic and semantic sources of information. On one hand, acoustic context provides information concerning speaker-specific and dialectal variation. On the other hand, semantic contextual information contributes to the selection of appropriate lexical candidates. In order to investigate the interaction between these sources of contextual information, a dataset of short sentences has been conceived that are dedicated to studying these phenomena in future research. The final database is organised around 28 carrier sentence triplets that end with minimal CV or CVC word-pairs. Target word-pairs are articulated around 4 french vowel constrasts. In order to assess the validity of the 3 corresponding semantic categories, measures of semantic similarity and lexical frequency have been computed from various language databases.

MOTS-CLÉS : perception de la parole, base de données, effets de contexte, plongements de mots, incertitude.

KEYWORDS: speech perception, database, context effects, word embeddings, entropy.

Volume 1 : Journées d'Études sur la Parole, pages 353–361.
HAL : hal-02798558.

1 Introduction

La perception du signal sonore est soumise à différentes sources d'incertitude qui trouvent leur origine dans la grande variabilité des réalisations phonétiques (Joos, 1948; Peterson, 1961; Nearey, 1989; Meunier, 2005). Les effets de contexte constituent deux sources d'information *extrinsèque* qui interviennent dans la catégorisation perceptive ((Ladefoged & Broadbent, 1957; Connine & Clifton, 1987) et interagissent avec l'analyse *intrinsèque* d'un segment.

Les sources d'incertitude acoustiques sont liés aux caractéristiques sonores des énoncés : des paramètres tels que la fréquence des formants, la fréquence fondamentale, la durée ou l'intensité, varient pour un segment donné (variation *intrinsèque*) mais s'expriment aussi par des « tendances » globales à l'intérieur d'une fenêtre plus large que le segment (variation *extrinsèque*). Ces différentes sources de variation interviennent dans la catégorisation perceptive d'un segment.

1.1 Effets de contexte acoustique

Les effets du contexte acoustique ont été initialement mis en évidence dans les travaux princeps de Ladefoged & Broadbent (1957). Les participants devaient répéter un mot cible (/bɪt/, /bɛt/, /bat/ ou /bʌt/) précédé d'une phrase de consigne *« Please say what this word is »*. Cette phrase était altérée synthétiquement sur les valeurs de fréquences de F_1 et F_2 pour créer six contextes acoustiques distincts correspondant à des manipulations phonétiques du contexte : (1) F_1 et F_2 originaux, (2) F_1 bas, (3) F_1 haut, (4) F_2 bas, (5) F_2 haut, (6) F_1 bas et F_2 haut. Les réponses perceptives des participants étaient toujours réalisées sur une cible acoustique non-modidiée. Néanmoins, les auteurs montrent que les réponses perceptives sont systématiquement influencées par le contexte acoustique de la phrase : /bɪt/ est perçu /bɪt/ à 87.5 % dans le contexte d'origine mais comme /bɛt/ à 90 % quand F_1 est abaissé. Inversement /bɛt/ est perçu /bɛt/ par 77 % à 95 % des participants dans les contextes où F_1 est intact ou abaissé, mais comme /bɪt/ à 97 % quand F_1 est rehaussé. /bat/ est moins bien reconnu dans son contexte d'origine (à 58 % contre 42 % pour /bɛt/) mais une hausse de F_1 fait basculer la perception vers /bɛt/ à 80 %. Enfin, /bʌt/ est reconnu /bʌt/ à 82 % dans son contexte d'origine, mais une baisse de F_2 fait diminuer la performance d'identification à 38 % au profit de /bat/ (60 %). Ces changements de catégorisation perceptive sont interprétés comme des effets de compensation : si la cible ne change pas, les variations de son environnement modifient sa perception.

Ces effets ont également été confirmés plus récemment par Sjerps (Sjerps *et al.*, 2013). Les expériences présentées reproduisent les effets de compensation observés tout en donnant des pistes de réponse sur le niveau auquel ces variations sont traitées dans la catégorisation des phonèmes. Les auteurs utilisent pour cela un stimulus au format [Vpapu] où la voyelle cible (V) est un des dix intervalles d'un continuum entre [i] et [ɛ]. Le mot porteur [papu] est également modifié pour avoir deux contextes $F_1 + 200$ Hz et $F_1 - 200$ Hz sur les deux voyelles. En tâche de catégorisation les participants doivent d'abord identifier comme [i] ou [ɛ] les dix intervalles cibles dans les deux contextes porteurs. Les résultats montrent que le basculement de la perception de la voyelle de [i] vers [ɛ] sur les paliers médians est affecté par le contexte : si les paliers 5 et 6 sont reconnus comme [ɛ] dans environ 40 % des essais lorsque F_1 est abaissé, le taux de reconnaissance de [ɛ] augmente à 65 % lorsque la fréquence de F_1 est élevée. Cette expérience suggère aussi que ces effets de compensation prennent place y compris en présence d'une quantité temporelle d'information très limitée. La seconde expérience est une tâche de discrimination en 4I-oddity : un objet déviant parmi un ensemble de 4. Les participants devaient trouver la voyelle déviante dans un ensemble exclusivement [ipapu] ou [ɛpapu] sur les dix variantes

du continuum de cibles et dans les deux contextes porteurs. Les participants ont montré davantage de difficultés à reconnaître [ɛpapu] dans un contexte de F_1 bas (environ 80 % indépendamment de l'écart entre le déviant et les standards), et inversement pour la reconnaissance de [i] en contexte de F_1 élevé (environ 60 % pour un palier déviant proche du standard, environ 80 % lorsque le palier déviant est éloigné). Ce type de tâche demande une concentration portée davantage sur les indices acoustiques du signal, et non sur ses propriétés phonologiques. Ces résultats soutiennent l'idée que les effets de compensation prennent place à une étape très précoce des processus de catégorisation.

1.2 Effets de contexte sémantique

La catégorisation phonétique est aussi influencée par des informations provenant du contexte sémantique. Ces phénomènes ont été initialement mis en évidence par (Connine & Clifton, 1987) avec un matériel constitué de paires minimales de mots CVC où la variation portait sur le caractère voisé / non-voisé de la consonne initiale (p. ex. : ang. *dime* / *time*). Ces paires avaient subi un traitement synthétique pour créer un continuum de dix paliers, dont 5 allant vers le voisement et 5 allant vers le dévoisement de la consonne. Ces cibles étaient précédées de deux phrases introductrices possibles, chacune tendant vers le sens d'un seul mot de la paire. Les participants identifiaient le mot cible (ce qui donnait une indication sur la reconnaissance de la consonne initiale variable) et devaient dire si la phrase finale formée faisait alors sens ou non. Les résultats ont montré que les cibles phonétiques sont perçues correctement indépendamment du contexte sur les paliers extrêmes du continuum, mais que la perception varie selon le contexte sur les seuils proches de la frontière catégorielle entre les deux consonnes : le contexte correspondant au mot voisé exerce une attraction de la variable vers la consonne voisée, et inversement. L'expérience montre également un temps de réaction plus lent lorsque la cible n'est pas sémantiquement liée à la phrase. Selon (Connine & Clifton, 1987), le contexte sémantique apporterait une information destinée à faciliter la perception mais il aurait un rôle tardif : l'ambiguité générée lors de la phase perceptive initiale serait résolue à un niveau post-perceptif (décisionnel) par l'information contextuelle.

Ces deux sources d'informations –acoustique intrinsèque d'une part, liée au contexte sémantique d'autre part– exercent donc des influences combinées sur la catégorisation du signal. Gaskell & Marslen-Wilson (2001) ont cherché à mieux comprendre ces mécanismes à partir de deux approches concernant l'impact potentiel du contexte sémantique dans les modèles de reconnaissance des mots. Selon certains modèles, l'information sémantique peut bloquer très précocément l'émergence d'un candidat lexical alors que d'autres modèles attribuent un rôle primordial aux informations ascendantes et n'intègrent les informations sémantiques que très tardivement dans le processus d'identification. Cette opposition correspond à des divergences observées dans la littérature. Par exemple, Tabossi (1988), utilisant du matériel fondé sur la distinction sémantique entre homophones (p. ex. : *bank* désigne en anglais à la fois une institution financière et le bord d'une rivière), observe des effets du contexte sémantique. À l'inverse, Connine *et al.* (1994), qui avaient eu recours à des versions ambiguës intermédiaires entre deux mots phonologiquement distincts (p. ex : ang. *dip / tip*), concluaient au caractère tardif du contexte sémantique.

Parmi les hypothèses expliquant les divergences des résultats, Gaskell & Marslen-Wilson (2001) explorent la possibilité qu'il n'existe pas de mécanisme d'analyse sémantique résolvant l'ambiguité entre deux mots phonologiquement distincts alors que le contexte pourrait par contre résoudre l'ambiguité lexicale entre homophones. Pour ce faire, ils étudient le comportement de mots subissant une altération liée à la coarticulation (p. ex. : ang. *ru**n** / ru**m** – **d**oes / **p**icks*) dans une tâche d'amorçage

multimodal de répétition : les participants doivent réaliser une tâche de décision lexicale sur une présentation visuelle du mot cible (*rum* ou *run*) juste après la présentation auditive du mot en contexte.

Les auteurs observent que en situation où le contexte phonétique influence l'interprétation vers une cible coronale (« rum does », réalisé comme un équivalent phonétique de « run does ») la présence du contexte sémantique facilite l'accès vers l'interprétation de la cible coronale (« run ») sans bloquer la cible non-coronale (« rum »). Les deux représentations sont accessibles. Ces expériences suggèrent donc que l'information acoustique ascendante demeure prioritaire. Le contexte sémantique n'apporte une information que lorsque les indices phonétiques et phonologiques sont insuffisants pour résoudre l'ambiguité entre deux représentations actives.

La base de données que nous présentons ici a pour objectif d'étudier plus spécifiquement les interactions entre les effets du contexte sémantique (Connine & Clifton, 1987; Gaskell & Marslen-Wilson, 2001) et ceux de l'adaptation au contexte acoustique (Ladefoged & Broadbent, 1957; Sjerps *et al.*, 2013) en proposant un ensemble de phrases contrôlées du point de vue de paramètres objectifs fondés sur des méthodes de plongements de mots (Mikolov *et al.*, 2013).

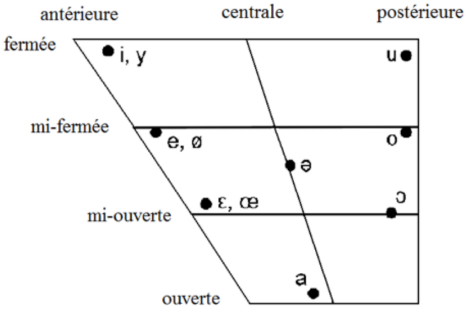
2 Méthode

La construction de la base de données s'est faite en deux étapes :

1. Une première phase de pré-sélection de mots-cible ;
2. Une seconde phase de construction de phrases à partir de mesures de similarité lexicale entre mots (Mikolov *et al.*, 2013; Fauconnier, 2015).

Les relations de similarité ont ensuite été évaluées pour les phrases générées et celles qui ne remplissaient pas les critères de relation établis entre les 3 catégories de lien sémantique ont été supprimées de la base de données. Des mesures de fréquence des mots-cible sélectionnés ont été réalisées afin de vérifier l'absence de déséquilibre entre les deux groupes.

2.1 Sélection des mots-cible



Voyelle	F_1	F_2	F_3
a	788	1503	2737
ɛ	660	2080	2954
a	788	1503	2737
ɔ	634	1180	2690
y	305	2046	2535
ø	469	1605	2581
o	461	855	2756
u	311	804	2485

FIGURE 1 – Trapèze vocalique du français, avec les valeurs des fréquences des trois premiers formants caractéristiques des voyelles cibles pour chaque paire vocalique de l'ensemble de données. Ces valeurs de fréquence des formants sont extraites de Calliope (Coll.) (1989).

Les mots-cible sont des mots français correspondant à des paires distinctives de type CV ou CVC selon les paires sélectionnées (p. ex. : « gare » /gaʁ/ et « guerre » /gɛʁ/) et qui se distinguent par leur voyelle. À partir d'une recherche de mots monosyllabiques du français dans la base de données BRULEX (Content *et al.*, 1990), nous avons sélectionné des paires de mots associées à 4 couples de voyelles du français. Ces couples de voyelles correspondent à des catégories dont les valeurs de fréquence des formants relevées sont proches dans un espace vocalique $F_1 \sim F_2 \sim F_3$ (Calliope (Coll.), 1989). Les mots cibles sont tous des noms communs. Pour une paire donnée, ils sont de même genre ou ont la même forme au pluriel afin que le déterminant soit identique dans les différents contextes sémantiques. La base de données constituée porte ainsi sur les paires [a / ɛ], [a / ɔ], [y / ø] et [o / u]. Nous avons initialement identifié 46 paires de mots qui pouvaient potentiellement servir de base à la construction des phrases.

2.2 Construction des phrases

Pour chacune des 46 paires de mots identifiées, nous avons procédé à la conception de phrases porteuses (Sujet-Verbe) correspondant à 3 catégories sémantiques :

Contexte 1 Phrase porteuse dont la signification est associée au mot 1 de la paire (exemple : « balle », /bal/) mais n'est pas liée au mot 2 (exemple : « belle », /bɛl/) ;

Contexte 2 Phrase porteuse dont la signification est associée au mot 2 de la paire (exemple : « belle », /bɛl/) mais n'est pas liée au mot 1 (exemple : « balle », /bal/) ;

Contexte 0 Phrase porteuse dont la signification n'est associée à aucun des deux mots de la paire ;

Des exemples de ces 3 contextes sont donnés dans le tableau 1.

TABLE 1 – Exemple de combinaison entre une paire de mots et les phrases forgées correspondant aux 3 contextes sémantiques possibles. Le contexte 1 favorise le mot 1, le contexte 2 favorise le mot 2, le contexte 0 ne favorise aucun des deux mots.

Mot de la paire	Contexte	Phrase finale
(1) balle	(1) Le joueur a dévié la	Le joueur a dévié la balle.
	(2) Le prince a charmé la	Le prince a charmé la balle.
	(0) La salade a raccourci la	La salade a raccourci la balle.
(2) belle	(1) Le joueur a dévié la	Le joueur a dévié la belle.
	(2) Le prince a charmé la	Le prince a charmé la belle.
	(0) La salade a raccourci la	La salade a raccourci la belle.

Les phrases ont été constituées à l'aide de mesures de plongements de mots (Word2vec, Mikolov *et al.*, 2013) issues d'un modèle du corpus Wikipedia en français (Fauconnier, 2015). Ces mesures de plongements de mots (ang. *word embeddings*) fournissent une estimation de la similarité sémantique entre deux mots exprimée par un vecteur allant de 0 (aucune proximité) à 1 (proximité maximale). Sur la base d'essais progressifs, nous avons fixé des valeurs de similarité types comme seuils d'acceptabilité minimal ou maximal pour les mots composant les phrases de chaque contexte. Dans un premier temps, nous avons d'abord cherché à sélectionner les verbes afin d'obtenir une valeur de similarité supérieure à 0.250 avec la cible de leur contexte et inférieure à 0.150 avec la cible opposée

ou présentant un différentiel de similarité d'au-moins 0.2 entre les deux contextes. Par exemple, pour la paire « balle » / « belle » le contexte 1 correspond au verbe « dévier » dont la proximité avec « balle » est de 0.334 alors que sa similarité avec le mot « belle » n'est que de 0.096 (différence = 0.238). Pour le contexte 2 lié à « belle », le verbe « charmer » a une similarité de 0.343 avec la cible mais de seulement 0.044 avec le mot « balle » (différence = 0.299).

Les noms-sujet de chaque phrase ont ensuite été choisis de manière à respecter l'une des deux conditions suivantes :

— soit répondre aux mêmes seuils d'acceptabilité que pour le verbe (au-moins 0.250 pour le contexte relié et au-plus 0.150 pour le contexte non-relié),
— soit, si cette condition n'était pas possible, présenter une différence *positive* en faveur du contexte supposé relié.

Ainsi, pour la même paire « balle » / « belle », le contexte 1 a pour sujet « joueur » dont la proximité avec la cible « balle » est de 0.374 alors que sa similarité avec la cible « belle » est de 0.095. Le contexte 2 a pour sujet « prince » qui présente une valeur de similarité de 0.186 avec la cible, et de seulement 0.054 avec « balle ». On voit que le seuil de 0.250 entre le sujet et la cible n'est pas respecté. Par contre, la différence entre les deux contextes est bien positive.

Pour la construction des phrases correspondant au contexte 0 (aucun lien avec l'un des deux membres de la paire), nous avons sélectionné un sujet et un verbe dont les valeurs de similarité sont systématiquement inférieures à 0.100 avec les deux cibles. Toujours pour la même paire, le sujet « salade » correspond à une proximité de 0.051 avec « balle » et de 0.099 avec « belle » ; le verbe « raccourcir » présente des valeurs de similarité de 0.075 avec « balle » et de 0.052 avec « belle ».

Sur les 46 paires de mots initiales, 18 n'ont pas permis de trouver des combinaisons de 3 phrases porteuses respectant ces conditions. Au final, nous obtenons un ensemble de 28 triplets de phrases porteuses associés à 28 paires de mots. Le biais sémantique induit par la phrase porteuse peut être vu comme la combinaison des valeurs de proximité du sujet et du verbe avec le mot-cible.

Sur la base de ces éléments, les phrases sont composées de manière à ce que, pour une paire de mots, les 3 contextes soient exprimés avec un accord en genre et en nombre. Les verbes sont conjugués au même temps. L'uniformité des déterminants / genre / nombre / temps verbal au sein des trois contextes d'une même paire permet d'interchanger ces contextes devant les cibles pour opérer des manipulations du contexte sémantique. Les phrases sont relativement courtes, de 6 à 12 syllabes, ce qui permet une interprétation rapide de la structure tout en conservant une quantité suffisante de matériel acoustique et sémantique pour générer les effets de contexte prédits.

2.3 Enregistrement audio

L'enregistrement de l'intégralité des phrases a été effectué en chambre sourde par deux locuteurs natifs du français (français urbain de l'ouest de la France), une femme de 20 ans et un homme de 22 ans étudiants à l'Université de Nantes. Les phrases étaient présentées dans un ordre aléatoire et répétées trois fois en imposant un débit relativement soutenu à travers l'interface de présentation visuelle des phrases.

Pour chaque phrase enregistrée, nous avons ensuite déterminé deux informations temporelles : le moment de transition entre la phrase porteuse et la cible, ainsi que la position correspondant au milieu de la voyelle. Le point de transition a été déterminé en deux phases. Dans un premier temps, nous avons positionné ce point à partir de l'étude acoustique (spectrogramme et forme d'onde) des signaux.

Dans un second temps, un script permettait d'écouter les séquences en isolant la phrase porteuse du mot cible afin de déterminer dans quelle mesure ce point temporel permettait de séparer les deux parties de la phrase de manière satisfaisante. Cette position temporelle pouvait alors être modifiée de manière à améliorer le découpage. Un script Python a été conçu afin de séparer les phrases porteuses des cibles avant de les recombiner par *cross-splicing*, en suivant un *design* de carré-latin : une cible enregistrée à l'origine en contexte 0 est rattachée à la porteuse de contexte 1, une cible enregistrée en contexte 1 est rattachée à la porteuse de contexte 2 et une cible enregistrée en contexte 2 est rattachée à la porteuse de contexte 0. Cette méthode permet de neutraliser complètement les effets de coarticulation (courte et longue distance) entre la phrase porteuse et le mot-cible.

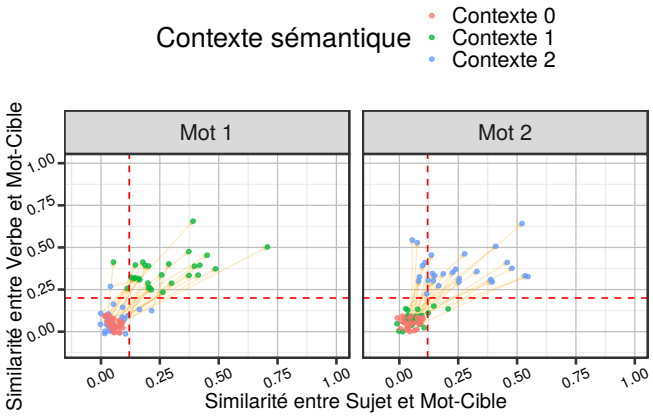
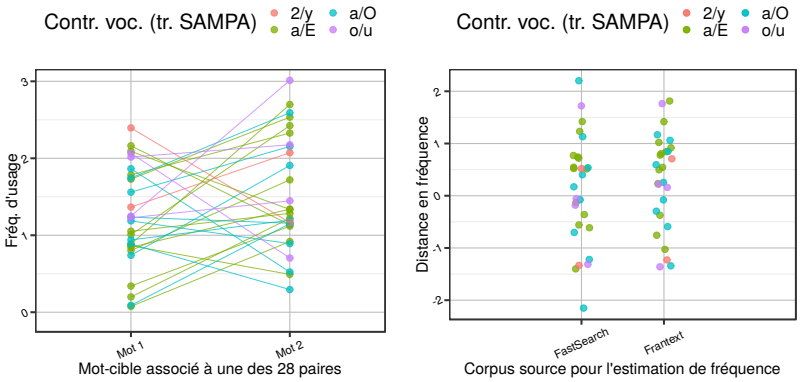


FIGURE 2 – Valeurs de similarité entre le mot-cible et respectivement le sujet (abcisse) / le verbe (ordonnée). Chaque point correspond à un mot d'une paire (Mot-cible 1 dans le graphique de gauche, mot-cible 2 dans le graphique de droite). Les points verts correspondent aux phrases porteuses associées au mot 1 (contexte 1), les points bleus correspondent aux phrases porteuses associées au mot 2 (contexte 2), les points rouges correspondent aux phrases porteuses de contexte 0. Les traits rouges en pointillés sont des repères visuels pour distinguer les nuages de points.

3 Résultats

Le corpus final est constitué de 28 paires de mots qui sont réparties de manière déséquilibrée entre les 4 contrastes vocaliques : 13 paires [a / ɛ], 9 paires [a / ɔ], 2 paires [y / ø], et 4 paires [o / u]. Les mesures de similarité caractérisant les relations entre mot-cible, verbe et nom-sujet en fonction du type de contexte sont présentées dans la figure 2. Ces données permettent de vérifier le caractère opérant des 3 catégories de contexte considérées : les valeurs de similarité du sujet et du verbe sont regroupées en position haute et / ou droite du graphique pour le contexte sémantiquement relié (contexte 1 / mot 1, contexte 2 / mot 2) alors que les valeurs sont regroupées en position basse et / ou gauche pour le contexte sémantiquement non-relié (contexte 1 / mot 2, contexte 2 / mot 1). Les valeurs associées au contexte 0 sont dans tous les cas localisées en bas / à gauche.

Pour chaque paire de mots, nous avons également recueilli les fréquences d'occurrence issues de la base de données Lexique (New *et al.*, 2001). Ces mesures sont représentées dans la figure 3. La différence de fréquence d'usage entre membres d'une paire est non-significative aussi bien sur le corpus *Frantext* (textes écrits, $\overline{x} = 0.307, sd = 0.898$; $t_{27} = 1.81, p = 0.082$), que sur le corpus *FastSearch* (pages web, $\overline{x} = 0.109, sd = 1.03, t_{27} = 0.559, p = 0.58$).



(a) Valeurs individuelles de fréq. (b) Distance en fréq.

FIGURE 3 – Fréquences d'occurrence ($\log_{10}$ sur 1 million) des mots-cible relevées dans la base de données Lexique (New *et al.*, 2001). (3a) valeurs individuelles comparées entre mots d'une même paire pour la fréquence d'occurrence sur des textes écrits (Frantext). (3b) distance de fréquence entre le mot 1 et le mot 2 issues de pages web (Fastsearch) et de textes écrits (Frantext).

4 Discussion

L'objectif de cette base de données est de mettre à disposition un matériel expérimental permettant de confronter les informations liées aux sources d'incertitude acoustique et sémantique dans les mécanismes de perception de la parole en fournissant des mesures objectives de relations sémantiques fondées sur des modèles de plongements de mots (Mikolov *et al.*, 2013). Ce travail fournit une liste de $3 \times 28 \times 2 = 168$ phrases distinctes correspondant à la combinaison de 28 paires de mots acoustiquement proches avec 3 contextes sémantiques distincts. Ces phrases ont été enregistrées par deux locuteurs francophones et manipulées afin de produire des croisements entre phrase porteuse et mot-cible qui permettent de supprimer la contribution des effets de coarticulation de la phrase porteuse vers le mot-cible.

Des mesures supplémentaires sont en cours sur un modèle alternatif du corpus Wikipedia en français (Gaudrain & Crouzet, 2019) afin de vérifier les valeurs issues du modèle de Fauconnier (2015). Nous avons également soumis l'ensemble des phrases possibles à un échantillon de locuteurs natifs du français et recueilli leurs jugements d'interprétabilité de 1 (non-interprétable) à 5 (totalement interprétable) afin de comparer nos mesures avec des estimateurs issus de réponses fournies par des locuteurs. Ces données sont en cours de récolte.

Cette base de données, grâce aux méta-informations fournies (valeurs de similarité, fréquences des mots, position temporelle du milieu de la voyelle du mot-cible…) et aux manipulations réalisées (*cross-splicing*), est multidisciplinaire et d'autres applications et expériences pouvant utiliser ce type de données peuvent être envisagées, notamment autour de questions d'acquisition et de *machine learning*.

L'intégralité des enregistrements (originaux, segmentés et réassemblés par *cross-splicing*) ainsi que la liste des phrases et les mesures réalisées sont mis à disposition sur un dépôt Zenodo (https://doi.org/10.5281/zenodo.3818582).

Remerciements

Ce travail a reçu le soutien de la « Mission pour les Initiatives Transverses et l'Interdisciplinarité » (MITI, CNRS, FR) et du programme Marie Skłodowska–Curie (PRESTIGE–2017–2–0044, UE).

Références

CALLIOPE (COLL.) (1989). *La Parole et son traitement automatique*. Paris : Masson.

CONNINE C. M., BLASKO D. G. & WANG J. (1994). Vertical similarity in spoken word recognition : Multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics*, **56**(6), 624–636. DOI : 10.3758/bf03208356.

CONNINE C. M. & CLIFTON C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology : Human Perception and Performance*, **13**(2), 291–299. DOI : 10.1037/0096-1523.13.2.291.

CONTENT A., MOUSTY P. & RADEAU M. (1990). Brulex. une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique*, **90**(4), 551–566. DOI : 10.3406/psy.1990.29428.

FAUCONNIER J.-P. (2015). French word embeddings. http://fauconnier.github.io.

GASKELL G. & MARSLEN-WILSON W. D. (2001). Lexical ambiguity resolution and spoken word recognition : Bridging the gap. *Journal of Memory and Language*, **44**(3), 325–349. DOI : 10.1006/jmla.2000.2741.

GAUDRAIN E. & CROUZET O. (2019). word2vec model trained on lemmatized French Wikipedia 2018. type : dataset, DOI : 10.5281/zenodo.3241447.

JOOS M. (1948). *Acoustic Phonetics*. Language monographs. Linguistic Society of America.

LADEFOGED P. & BROADBENT D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, **29**(1), 98–104. DOI : 10.1121/1.1908694.

MEUNIER C. (2005). Invariants et variabilité en phonétique. In N. NGUYEN, S. WAUQUIER-GRAVELINES & J. DURAND, Éds., *Phonologie et Phonétique : Forme et Substance*, chapitre 13, p. 349–374. Paris : Lavoisier.

MIKOLOV T., CHEN K., CORRADO G. S. & DEAN J. (2013). Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781v3.

NEAREY T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, **85**(5), 2088–2113. DOI : 10.1121/1.397861.

NEW B., PALLIER C., FERRAND L. & MATOS R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE. *L'année psychologique*, **101**(3), 447–462. DOI : 10.3406/psy.2001.1341.

PETERSON G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research*, **4**(1), 10–29. DOI : 10.1044/jshr.0401.10.

SJERPS M. J., MCQUEEN J. M. & MITTERER H. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, & Psychophysics*, **75**(3), 576–587. DOI : 10.3758/s13414-012-0408-7.

TABOSSI P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, **27**(3), 324–340. DOI : 10.1016/0749-596x(88)90058-7.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Introduction d'informations sémantiques dans un système de reconnaissance de la parole

Stéphane Level , Irina Illina, Dominique Fohr
Equipe MultiSpeech
Université de Lorraine, CNRS, Inria, F-54000 Nancy, France
{stephane.level,irina.illina,dominique.fohr}@loria.fr

Résumé

Malgré les avancés spectaculaires ces dernières années, les systèmes de Reconnaissance Automatique de Parole (RAP) commettent encore des erreurs, surtout dans des environnements bruités. Pour améliorer la RAP, nous proposons de se diriger vers une contextualisation d'un système RAP, car les informations sémantiques sont importantes pour la performance de la RAP. Les systèmes RAP actuels ne prennent en compte principalement que les informations lexicales et syntaxiques. Pour modéliser les informations sémantiques, nous proposons de détecter les mots de la phrase traitée qui pourraient avoir été mal reconnus et de proposer des mots correspondant mieux au contexte. Cette analyse sémantique permettra de réévaluer les *N* meilleures hypothèses de transcription (*N-best*). Nous utilisons les *embeddings* Word2Vec et BERT. Nous avons évalué notre méthodologie sur le corpus des conférences TED (TED-LIUM). Les résultats montrent une amélioration significative du taux d'erreur mots en utilisant la méthodologie proposée.

Abstract

Despite spectacular advances in recent years, the Automatic Speech Recognition (ASR) systems still make mistakes, especially in noisy environments. In order to reduce these errors, we suggest moving towards a contextualization of a ASR system, because semantic information is important for the performance of ASR. Current ASR systems mainly take into account only lexical and syntactic information. To model the semantic information, we propose to detect the words of the recognised sentence, which could have been badly recognized and to propose words corresponding better to the context. This semantic analysis will allow to re-evaluate the N-best hypotheses of recognition. We use Word2Vec embedding and Google's BERT model. We evaluated our methodology on the corpus of TED conferences (TED-LIUM). The results show a significant improvement of the word error rate using the proposed methodology.

Mots-clés : reconnaissance automatique de la parole, contexte sémantique ,*embeddings*, Word2Vec, BERT.

Keywords: automatic speech recognition, semantic context, embeddings, Word2Vec, BERT.

1 Introduction

Grace aux réseaux de neurones profonds, les systèmes de reconnaissance automatique de la parole commencent à être utilisables dans les conditions réelles de notre vie de tous les jours. D'ailleurs des nombreux industriels proposent déjà des systèmes vocaux pour nos maisons, nos voitures et nos smartphones.

Malgré des efforts constants et quelques avancées spectaculaires, la capacité d'une machine à reconnaître la parole est encore loin d'égaler celle de l'être humain. Les systèmes RAP actuels

Volume 1 : Journées d'Études sur la Parole, pages 362–369.
HAL : hal-02798559.

voient leurs performances diminuer de manière significative lorsque les conditions dans lesquelles ils ont été entraînés et celles dans lesquelles ils sont utilisés diffèrent. Les causes de variabilité existantes entre ces conditions peuvent être liées à l'environnement acoustique et/ou à l'acquisition du signal sonore. Le matériel de capture du son, le changement de microphone, l'environnement acoustique ajoutent au signal de la parole des composantes perturbatrices. L'approche classique comporte deux étapes : débruiter (rehausser) le signal puis le transmettre au système de RAP pour le décodage. Cependant, la performance d'un système de RAP sur un mot donné dépend toujours de la distorsion au moment précis où ce mot a été prononcé.

Pour résoudre ce problème, nous proposons de se diriger vers une **contextualisation** du système RAP. En effet, les informations lexicales, sémantiques et temporelles sont importantes pour qu'un système RAP soit performant. En revanche, les systèmes RAP actuels ne prennent en compte principalement que les informations lexicales et syntaxique (modèles de langage n-gramme locaux). Pour modéliser les informations sémantiques, plusieurs méthodes fondées sur des statistiques de cooccurrences, sur l'information mutuelle, sur un modèle vectoriel et sur les réseaux de neurones peuvent être utilisées (Sheihk, 2016).

Les **espaces sémantiques et thématiques** sont des espaces vectoriels utilisés pour la représentation numérique des mots, des phrases ou des documents textuels. Presque tous les modèles s'appuient sur l'hypothèse de la sémantique statistique qui stipule que: des schémas statistiques d'apparition des mots (contexte d'un mot) peuvent être utilisés pour décrire la sémantique sous-jacente (Turney & Pantel, 2010). La méthode la plus utilisée pour apprendre ces représentations est de prédire un mot en utilisant le contexte dans lequel ce mot apparaît : *embedding* (Mikolov *et al*., 2013 ; Pennington *et al*., 2014), et cela peut être réalisé avec des réseaux neuronaux. Ces représentations se sont avérées efficaces pour une série de tâches de traitement du langage naturel (Baroni *et al*., 2014). Elles sont devenues très populaires en raison de leur capacité à traiter de grandes quantités de données textuelles non structurées avec un faible coût de calcul. L'efficacité et les propriétés sémantiques de ces représentations nous motivent à explorer ces représentations sémantiques pour notre tâche de reconnaissance dans des conditions bruitées. Nous espérons que dans les parties très bruitées, le modèle de langage et le modèle sémantique peuvent permettre de lever les ambiguïtés acoustiques afin de trouver les mots prononcés par le locuteur.

Dans cet article nous proposons de compléter l'étape de RAP **par l'ajout d'informations sémantiques** afin de détecter les mots de la phrase traitée qui pourraient avoir été mal reconnus et de proposer des mots de prononciation similaire correspondant mieux au contexte. Cette analyse sémantique permet **de réévaluer** les *N* meilleures hypothèses de transcription (*N-best*) et peut être vue comme une forme d'adaptation dynamique dans le cadre spécifique des données bruitées. Les informations sémantiques sont introduites en utilisant des représentations prédictives à l'aide de vecteurs continus (*embeddings*). Toutes nos modélisations s'appuient sur les technologies performantes de DNN. Par rapport aux travaux précédents utilisant la réévaluation de la liste N-best (Song *et al.*, 2019 ; Shin *et al.*, 2019 ; Ogawa *et al.* 2018), nous nous appuyons uniquement sur des informations sémantiques. De plus, la spécificité de notre approche est l'utilisation de la partie contextuelle et des zones de possibilité de la liste des N-hypothèses: la partie contextuelle représente l'information sémantique du contexte thématique du document à reconnaître et la zone de possibilité correspond à la zone où nous voulons trouver les mots à corriger. Cela nous permet de donner moins d'importance aux mots de la zone de possibilité qui ne correspondent pas au contexte du document, et de donner un faible score sémantique à l'hypothèse correspondante.

2 Méthodologie proposée

Une façon efficace de prendre en compte les informations sémantiques est de réévaluer les meilleures hypothèses du système de reconnaissance. Le système de reconnaissance nous fournit

pour chaque mot de la phrase hypothèse un score acoustique $p_{ac}(w)$et un score linguistique $p_{lm}(w)$. La meilleure phrase est celle qui maximise la probabilité de la séquence de mots :

$$\widehat{W} = \underset{h_i \in H}{argmax} \prod_{w \in h_i} p_{ac}(w)^{\alpha} * p_{lm}(w)^{\beta} \quad (1)$$

$\widehat{W}$ est la phrase reconnue (le résultat final) ; H est l'ensemble des *N*-meilleures hypothèses de phrases ; h_i est la *i*-ème hypothèse de phrase ; *w* est un mot de l'hypothèse. α et β représentent les poids du modèle acoustique et du modèle de langage. Ils sont indispensables car les scores acoustiques et les scores linguistiques ne sont pas toujours normalisés (ce sont souvent des vraisemblances et non des probabilités). Ces poids sont ajustés sur un corpus de développement.

Nous souhaitons **ajouter de l'information sémantique** pour guider le processus de reconnaissance. L'approche la plus naturelle pour intégrer cette information consiste à modifier le calcul de la probabilité de la séquence de mots de la façon suivante :

$$\widehat{W} = \underset{h_i \in H}{argmax} \prod_{w \in h_i} p_{ac}(w)^{\alpha} * p_{lm}(w)^{\beta} * p_{sem}(w)^{\gamma} \quad (2)$$

Nous avons ajouté la probabilité sémantique de chaque mot : $p_{sem}(w)$. Pour avoir un bon équilibre entre les différents modèles, nous introduisons un troisième poids γ pour pondérer l'information sémantique. Il sera également ajusté sur un corpus de développement.

2.1 Définition de contexte et des zones des possibilités

Pour estimer cette probabilité sémantique, nous proposons d'introduire les notions de ***contexte et* des *zone des possibilités. Un contexte*** est constituée des mots qui sont communs à toutes les *N*-meilleures hypothèses générées par le SRAP. Ce contexte nous permet d'extraire des informations sémantiques sur le contexte thématique du document ou de la partie courante du document à reconnaître. Pour obtenir un contexte sémantique plus significatif, il peut être intéressant d'ajouter à ce contexte les mots des phrases précédemment reconnues. ***Une zone des possibilités*** est une zone située entre les parties contexte. C'est dans cette zone que nous souhaitons retrouver les mots pour corriger la phrase reconnue. À partir des *N*-meilleures hypothèses d'une phrase, nous allons extraire un contexte et une ou plusieurs zones des possibilités. Chaque zone peut être constituée de plusieurs mots. La Figure 1 illustre ces notions sur un exemple. Ici les *3*-meilleurs hypothèses sont les suivantes :

H1: le chat mange la souris grise
H2: le chat ange la souris grise
H3: le chat mange la sous rit grise

Dans ce cas le contexte est composée des mots *le, chat, la* et *grise*. Ce sont les mots qui sont communs aux trois hypothèses. Entre ces mots, nous définissons deux zones des possibilités : la première est constituée de deux possibilités *mange et ange*. La deuxième est aussi constituée de deux possibilités *souris* et *sous rit*. Nous supposons que les zones de possibilités correspondent aux zones où le système de reconnaissance hésite entre différentes solutions.

Pour obtenir le contexte, nous utilisons un algorithme de programmation dynamique qui va permettre d'apparier les hypothèses deux à deux afin de déterminer les mots communs à toutes les hypothèses.

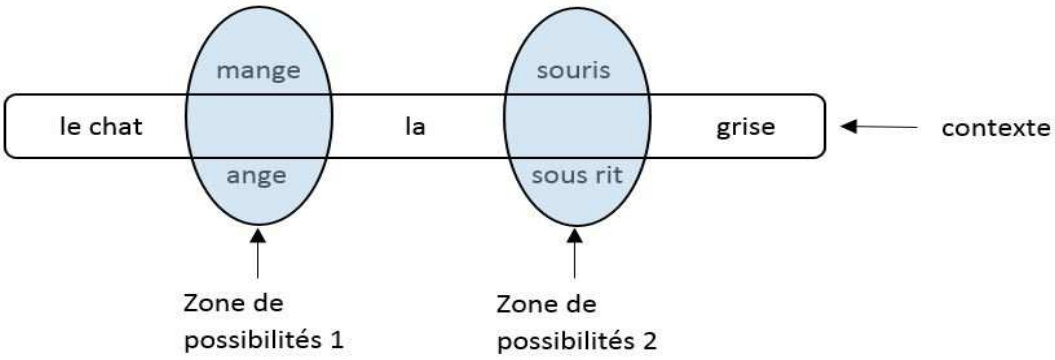


Figure 1. Illustration du contexte et des zones de possibilités pour un exemple.

2.2 Représentation sémantique du contexte et des zones des possibilités

Pour prendre en compte la sémantique du document à reconnaître, nous proposons de représenter chaque mot des N-meilleures hypothèses par un vecteur d'*embedding*. Dans notre approche, nous avons utilisé *word2vec (*Mikolov *et al*., 2013) et *BERT* (Devlin, 2018). ll est important de noter, que dans les représentations word2vec, l'*embedding* d'un mot est statique, c'est-à-dire un mot donné a un seul *embedding* quel que soit la phrase dans laquelle il apparaît. Dans le cas de représentations de mots à l'aide de BERT, l'*embedding* d'un mot dépend des autres mots de la phrase dans laquelle il apparaît et donc un mot peut avoir plusieurs *embeddings* en fonction du contexte de la phrase. Pour prendre en compte cette aptitude de BERT, lors de génération d'*embeddings* avec BERT pour le **contexte**, nous remplaçons chaque zone de possibilités par un masque spécial [*mask*] (comme dans le processus d'apprentisssage de BERT) et nous gardons inchangé le contexte. À partir de cette phrase avec les masques, BERT génère un *embedding* pour le contexte. Pour obtenir l'*embedding* pour une **zone de possibilités,** nous utilisons l'hypothèse de reconnaissance correspondante. Dans le cas où il y plusieurs zones de possibilités, toutes les zones de possibilités sont remplacées avec [*mask*] sauf la zone pour laquelle nous calculons l'*embedding*.

2.3 Calcul de la probabilité sémantique

A partir des représentations sémantiques du contexte et des zones de possibilités, nous pouvons calculer **une probabilité sémantique d'une hypothèse *h***. Cette probabilité sera utilisée dans la formule (2).

Pour prendre en compte la sémantique du document, nous représentons chaque mot des N-meilleures hypothèses par un vecteur d'*embedding,* comme décrit précédemment. Nous calculons un embedding moyen E_{cont} pour la partie contexte qui est égale à la moyenne des vecteurs d'*embedding* de tous les mots de la partie contexte. De la même manière, nous calculons un *embedding* moyen E_{pos} *(i, a)* pour la *i*-ème zone de possibilité de l'alternative a_h de l'hypothèse *h* comme la moyenne des vecteurs d'*embedding* de tous les mots dans cette alternative de la zone de possibilité. Une alternative correspond à un choix dans la zone de possibilité. Nous utilisons la similitude angulaire pour estimer un score sémantique entre chaque zone de possibilité et la partie contextuelle:

$$S_{sem}\left(E_{cont},, E_{pos}\left(i, a_h\right)\right) = 1 - \frac{\cos^{-1}\left(\cos\left(E_{cont}, E_{pos(i,\, a_h)}\right)\right)}{\pi} \qquad (3)$$

A partir des représentations sémantiques de la partie contexte et des zones de possibilité, nous calculons une probabilité sémantique $P_{sem}(h)$ d'une hypothèse h :

$$P_{sem}(h) = \prod_{i=1}^{N_p} S_{sem}\left(E_{cont}, E_{pos}(i, a_h)\right) \qquad (4)$$

où N_p est le nombre de zones de possibilité. Nous supposons que l'équation (2) peut être approximée comme suit:

$$\widehat{H} = \underset{h \in H}{argmax} \; P_{ac}(h)^{\alpha} P_{lm}(h)^{\beta} P_{sem}(h)^{\gamma} \qquad (5)$$

L'équation (5) est utilisée pour réévaluer la liste des N meilleures hypothèses. Pour chaque hypothèse, nous calculons le score sémantique et l'associons aux scores acoustiques et linguistiques selon (5). L'hypothèse qui obtient du meilleur score est considérée comme la phrase reconnue.

3 Expérimentations

3.1 Corpus utilisé

Nous avons utilisé le corpus TED-LIUM, la distribution standard (Hernandez *et al.,* 2018). Ce corpus contient les enregistrements des conférences TED. Le corpus est bien adapté à notre étude car chaque conférence est centrée sur un sujet particulier. L'ajout d'information sémantique avec un large contexte devrait permettre d'améliorer les performances de notre système de reconnaissance.

Le découpage du corpus TED en corpus d'apprentissage, développement et de test est proposé dans la distribution TED-LIUM et correspond à 452 heures d'apprentissage, 8 conférences pour le développement et 11 conférences pour le test. La Table 1 donne quelques statistiques sur le corpus de développement et de test car ce sont ces deux corpus qui nous intéressent pour l'introduction de l'information sémantique.

	Nbr. de documents audio	*Nbr. de phrases*	*Nbr. de mots*
Développement	8	500	17926
Test	11	1091	27021

Table 1 : Corpus de développement et de test du TED-LIUM.

3.2 Système de reconnaissance

Notre système de reconnaissance est fondé sur la boîte à outils de reconnaissance vocale *Kaldi* (Povey *et al*., 2011). Nous avons utilisé des modèles acoustiques triphones de type TDNN. L'apprentissage des modèles acoustiques TDNN a été réalisé en utilisant les 452 heures du corpus d'apprentissage de TED. Le lexique est composé de 150k mots et le modèle de langage contient 2 millions 4-grams, appris sur un corpus textuel de 250 millions de mots.

De façon classique, nous avons utilisé le corpus de développement pour choisir la meilleure configuration et ajuster les paramètres. Le corpus de test permet d'évaluer la méthode proposée avec les meilleurs paramètres obtenus sur le corpus de développement.

Pour se rapprocher des conditions réelles d'utilisation, nous avons décidé d'ajouter du bruit aux corpus de développement et de test. Nous avons ajouté un bruit additif de 10 dB et de 5dB (bruit de

F16 de la base NOISEX-92 (Varga, 1993)). La performance de notre système sur TED-LIUM sans ajout de bruit est autour de 8 % de taux d'erreur mots.

Les *embeddings word2vec* générés sont de taille 300 et modélisent 700 000 mots. Nous avons utilisé le modèle pré-entraîné *BERT Base* fourni par Google. Il est composé de 12 couches de *transformer,* chacun composés de 12 têtes d'attention. La taille de l'espace est 768. La métrique que nous avons utilisée est le taux d'erreur mots (WER).

4 Résultats expérimentaux

Le système de reconnaissance de base donne le taux d'erreur de 15.9 % WER pour le corpus de développement bruitée au Rapport Signal Bruit (RSB) de 10dB. Pour connaître le taux d'erreur minimal que nous pouvons obtenir en utilisant les *N* meilleures hypothèses, nous avons évalué le taux d'erreur *oracle* : 9.8 %. Ce taux est obtenu en sélectionnant l'hypothèse de la liste de *N*-meilleures hypothèses qui minimise le WER pour chaque phrase. Pour le corpus de développement bruité à 5dB nous avons obtenu un WER de 32,2 % et taux d'erreur mots *oracle* de 25,2 %.

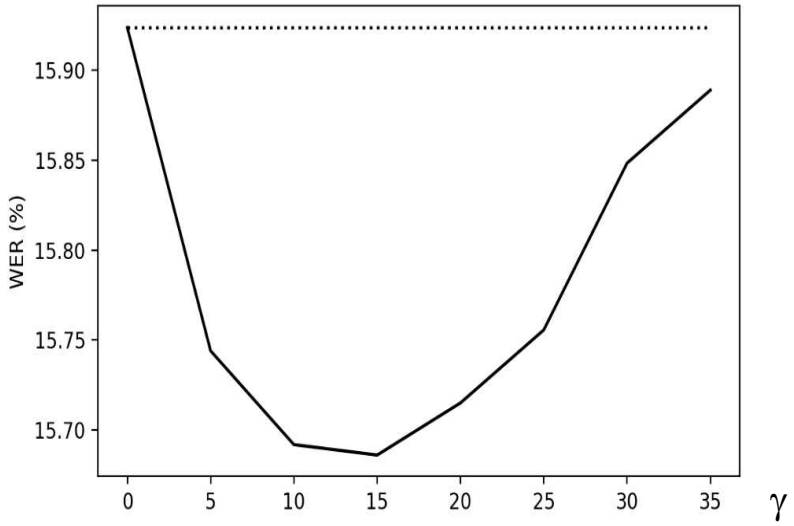


Figure 2. WER en fonction du coefficient sémantique γ. Méthode sémantique avec l'*embedding* Word2Vec. Corpus de développement TED-LIUM.

Nous avons évalué l'influence du paramètre γ (cf. formule 2) utilisé pour équilibrer les scores **acoustiques**, **langagiers** et **sémantiques**. La Figure 3 représente la courbe de ce paramètre pour l'*embedding* de Word2Vec en fonction de taux d'erreur obtenu. Nous observons que ce paramètre est important.

La Table 2 présente les résultats de reconnaissance sur le corpus TED-LIUM pour la partie développement et la partie test, ainsi que dans deux conditions de bruits : 10dB et 5dB. La première ligne de la table correspond au système sans le module sémantique, la dernière ligne à la performance maximale qu'on peut obtenir en recherchant dans N-meilleures phrase (oracle). Les lignes 2 et 3 correspondent aux approches proposées. Sur le corpus de test, nous obtenons une amélioration absolue de 0,4 % pour un RSB de 10dB (21,8 % versus 22,2 %) et de 0,8 % pour un RSB de 5dB (38,2 % versus 39 %) pour les parties test. Nous observons que les approches sémantiques proposées permettent de réduire le taux d'erreur mots. La meilleure performance est obtenue en utilisant l'*embedding* de Word2Vec. Dans les conditions plus bruitées (5dB) l'amélioration est un peu plus importante. Toutes les améliorations sont significatives par rapport au système de base.

Méthode	*RSB 10dB*		*RSB 5dB*	
	dev	*test*	*dev*	*test*
Système de base	15,9	22,2	32,2	39,0
Word2Vec *embedding*	15,7	**21,8**	**31,5**	**38,2**
BERT *embedding*	**15,5**	22,0	31,6	38,5
Oracle	9,8	14,0	25,2	29,8

Table 2 : Résultats de reconnaissance en terme de taux d'erreur de mots (WER %). Corpus de développement et de test TED-LIUM**,** RSB 10dB et 5 dB.

5 Conclusion et discussion

Dans cet article, nous voulions améliorer les performances d'un système de reconnaissance automatique de la parole en ajoutant des **informations sémantiques**. Nous proposons une méthodologie novatrice de la prise en compte de la sémantique à travers les représentations prédictives qui capturent les caractéristiques sémantiques des mots et de leur contexte. L'efficacité et les propriétés sémantiques de ces représentations récentes de type *embeddings* nous motivent à explorer ces représentations pour notre tâche de reconnaissance de la parole. Nous avons exploré les modèles **Word2Vec et BERT.** Les informations sémantiques sont prises en compte à travers le module de **réévaluation des N-meilleures hypothèses du système de reconnaissance.** Nous avons évalué notre méthodologie sur le corpus des conférences TED-LIUM. Les résultats montrent une amélioration significative du taux d'erreur mots en utilisant la méthodologie proposée.

Il existe de nombreuses extensions possibles de ce travail. Par exemple, il peut être possible d'améliorer les performances en explorant d'autres façons de calculer un *embedding* pour une zone. Il serait également intéressant d'étudier les différentes possibilités pour calculer le score d'une hypothèse.

Remerciements
Les auteurs remercient la DGA (Direction Générale de l'Armement), Thalès AVS et Dassault Aviation qui soutiennent le financement de cette étude et du programme scientifique «*Man-Machine Teaming*» dans lequel se déroule ce projet de recherche.

6 Références

BARONI M., DINU G., KRUSZEWSKI G. (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *In proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*
CORONA R., THOMASON J., MOONEY R.J. (2017). Improving Black-box Speech Recognition using Semantic Parsing. *In proceedings of the The 8th International Joint Conference on Natural Language Processing.*
DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Fernandez H., Nguyen H., Ghannay S., Tomashenko N., and Estève Y. (2018) TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *In proceedings of SPECOM.*
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. Dean, J. (2013) Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems 26, 3111–3119.*
Ogawa A., Delcroix M., Karita S., Nakatani t (2018) Rescoring N-best speech recognition list based on one-on-one hypothesis comparaison using encoder-classifier model. *In: Proceedings of the ICASSP.*
Pennington J., Socher R., Manning C.D. (2014) Glove: Global vectors for word representation. *In the Proceedings of the 2014 conference on empirical methods in natural langage.*
Povey D., Ghoshal A., Boulianne G.i, Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. (2011). The Kaldi Speech Recognition Toolkit. *In proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.*
Sheikh, I. (2016) Exploitation du contexte sémantique pour améliorer la reconnaissance des noms propres dans les documents audio diachroniques*, These de doctorat en Informatique, Université de Lorraine.*
Shin J., Lee Y., Jung K. (2019) Effective Sentence Scoring Method Using BERT for Speech Recognition. *In: Proceedings of Machine Learning Research*, pp.1081-1093.
Song Y., Jiangy D., Zhao X., Xuy Q., Wong R., Fany L., Yang Q. (2019) L2RS: a learning-to-rescore mechanism for automatic speech recognition. *arXiv:1910.11496v1.*
Turney P., Pantel P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. *In Journal of Artificial Intelligence Research,* 37, pp.141-188.
Varga A., Steeneken H. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, *Speech Communication,* Volume 12, Issue 3, pp. 247-251
Velikovich L., Williams I., Scheiner J., Aleksic P., Moreno P., Riley M. (2018) Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant, *In the Proceedings of Intersppech.*

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Production de la parole en réponse à de multiples perturbations du feedback auditif

Jinyu Li, Leonardo Lancia
Laboratoire de Phonétique et Phonologie, 19 rue des bernardins, Paris, France
jinyu.li@sorbonne-nouvelle.fr, leonardo.lancia@sorbonne-nouvelle.fr

RESUME

Des études antérieures ont montré que la production de la parole dépend des conditions du feedback auditif. Cette étude vise à investiguer les interactions entre les effets de trois facteurs différents sur la production de la parole : 1) le retard du feedback auditif (DAF), 2) le décalage de la f0 du feedback auditif et 3) la complexité des syllabes qui composent les énoncés. Nous avons manipulé le feedback auditif de 20 locutrices francophones pendant la répétition de trois phrases. Nous avons pu observer que plus de la moitié des participantes ont tendance à suivre la direction du décalage de la f0 du feedback auditif en recevant en continu cette perturbation. La position syllabique des voyelles est un facteur important affectant l'effet des perturbations du feedback auditif. Cependant les analyses décrites dans cette étude ne montrent pas un effet interactif du DAF et du décalage de la f0 sur la production de la parole.

ABSTRACT

Speech production in response to multiple perturbations of auditory feedback
Previous studies have shown that speech production depends on the conditions of auditory feedback. In the present study we investigate the interactions between the effects of three different factors: 1) the delay of the auditory feedback, 2) the f0 shift of auditory feedback, and 3) the complexity of syllables composing utterances. We analysed a corpus of French sentences, repeated several times by 20 French-speaking female speakers while their auditory feedback was manipulated. We could show that more than half of the participants tend to follow the direction of the f0 shift of auditory feedback. The vowels' syllabic position is an important factor affecting the effect of perturbations of auditory feedback. However, the analyses reported in this study do not support the idea that DAF and f0 shift have an interactive effect on speech production.

MOTS-CLÉS : multiples perturbations du feedback auditif, retard du feedback auditif, décalage de la f0 du signal du feedback auditif, agentivité, position syllabique

KEYWORDS: multiple perturbations of auditory feedback, delayed auditory feedback, pitch-shifted auditory feedback, agency, syllabic position

Volume 1 : Journées d'Études sur la Parole, pages 370–378.
HAL : hal-02798560.

1 État de l'art

Il est bien connu que les locuteurs réagissent rapidement au décalage de la fréquence fondamentale (f0) de leur feedback auditif. En général, ils produisent des réponses compensatoires. C'est-à-dire qu'ils changent leur f0 dans la direction opposée à celle du décalage, lors de la production des voyelles continues (ex : Burnett *et al*., 1998), des mots monosyllabiques ou multi-syllabiques (Natke & Kalveram, 2001 ; Donath *et al*., 2002), des phrases (Chen *et al*., 2007), et même en présence d'un retard du feedback auditif (DAF, ici pour *delayed auditory feedback*, Kalveram & Jäncke, 1989) ou d'une perturbation du volume du feedback auditif (Larson *et al*., 2000). De nombreux facteurs ont été signalés comme affectant la latence ou l'amplitude des réponses des locuteurs au décalage de la f0. Tout d'abord, la durée syllabique doit être suffisamment longue pour que les locuteurs puissent répondre au décalage (Natke & Kalveram, 2001). La vitesse de la réponse au décalage est proportionnelle à celle du déclenchement du décalage (Larson *et al*., 2000). De plus, les réponses à cette perturbation du feedback auditif dépendent du contrôle dynamique de la f0 par les locuteurs : dans les tâches qui nécessitent un contrôle précis et dynamique de la f0, l'amplitude des réponses est plus importante, la latence et le temps pour arriver au pic des réponses sont plus courts (Xu *et al*., 2004).

Il est également important de noter que certains locuteurs suivent parfois la direction du décalage de la f0. Cela se produit davantage lorsque le décalage de la f0 est prévisible (Behroozmand *et al*., 2012) ou lorsque la plage de décalage est importante (Burnett *et al*., 1998 ; Behroozmand *et al*., 2012). De plus, il a été montré que la direction de la réponse des locuteurs était dépendante de la tâche. Par exemple, il a été prouvé que les mécanismes contrôlant les réponses au décalage de la f0 du feedback auditif étaient sensibles à l'inflexion prévue de la f0 (Chen *et al*., 2007). La tendance à suivre ou contrer la perturbation de la f0 est également liée aux fluctuations du comportement du système de production (Franken *et al*., 2018) et à l'application de la perturbation à tous les essais de l'expérience ou seulement à certains essais (Franken *et al*., 2019).

Si les locutrices utilisent un modèle interne (Pickering & Garrod, 2013) pour contrôler leur production de la parole, lorsque les locuteurs suivent la direction de la perturbation de la f0, il semble que l'écart entre les valeurs prévues et perçues de la f0 augmente même. Pourquoi alors les locuteurs suivent-ils parfois la direction de la perturbation ? Ce comportement par ailleurs ressemble à celui observé quand les locuteurs produisent leurs énoncés de façon simultanée à ceux produits par une autre personne, car dans ce cas, ils tendent à imiter les caractéristiques de la voix de cette personne (Zheng *et al*., 2011). Le comportement du système sensorimoteur lors de la production de la parole pourrait donc dépendre à la fois des conditions du feedback sensoriel et de l'attribution de l'agentivité de la voix entendue. Il manque toutefois à la littérature une étude systématique de la manière dont les facteurs définissant les conditions du feedback sensoriel interagissent. Dans cette étude, nous étudions les interactions entre les effets du DAF, de la perturbation de la f0 du feedback auditif et de la complexité des syllabes composant les énoncés produits sur la production de la parole.

2 Expérimentation

Les données de cette étude ont été obtenues auprès de 20 locutrices francophones. Toutes les participantes étaient étudiantes universitaires ayant le français comme langue maternelle, et n'ayant pas de problèmes d'audition et de parole connus. Les participantes ont été divisées en deux groupes de dix selon la direction du décalage de la f0 dans leur feedback auditif (voir plus de détails ci-dessous). Trois phrases de cinq syllabes, se différenciant en complexité syllabique, ont été créées. Les trois phrases se composent respectivement et principalement de syllabes dont la structure est 1) CV : Vivien vit le vin ; 2) CVC : Jacqueline gère le jour ; 3) CCV : Bradley brise le bras. L'expérience s'est effectuée dans une chambre sourde. Les participantes, assises devant l'ordinateur donnant les consignes, portaient un micro et un casque. Les participantes percevaient leur voix, soit manipulée, soit normale, à travers le casque. L'expérience a commencé par une phase de familiarisation au DAF. Il a été demandé aux participantes de lire le court texte français « La bise et le soleil » avec un DAF de 120 ms. Pendant cette phase de familiarisation, le volume du feedback auditif dans le casque a été ajusté afin de minimiser la perception par les participantes de leur propre voix en dehors du casque (Burnett *et al*., 1998 ; Liu *et al*., 2012). Ensuite, dans la phase de test, nous avons demandé aux participantes de répéter les trois phrases en leur rythme confortable. Les essais expérimentaux (chacun consistant en une répétition des trois phrases dans un ordre aléatoire différent) ont été organisés en blocs de six. Au total, le test se compose de 16 blocs. Les essais composant le premier bloc ont été considérés comme des essais de contrôle, puisqu'il n'y avait aucune altération du feedback auditif. Pendant tous les essais de chaque bloc suivant, la f0 du signal du feedback auditif a été décalée d'une valeur constante tout le long du bloc (zéro, un ou deux demi-tons). Chaque degré de perturbation de la f0 a été appliqué à cinq blocs choisis au hasard. Pour les 10 participantes composant le groupe 1, le décalage de la f0 était toujours positif, tandis que pour les 10 participantes composantes le groupe 2, le décalage de la f0 était toujours négatif. Au cours de chaque essai, la valeur de DAF a été choisie au hasard parmi 0, 60 et 120 ms. Chaque valeur de DAF a donc été appliquée deux fois dans chaque bloc.

3 Analyse et résultats

Nous avons caractérisé les énoncés produits par nos participantes en analysant la durée des voyelles et la valeur médiane de la f0 de chaque voyelle. Nous avons d'abord étiqueté à la main une répétition de chaque phrase produite dans les essais de contrôle de chaque locutrice. Les frontières des phrases et des voyelles placées à la main ont été projetées sur les autres répétitions d'une même locutrice au moyen d'une procédure basée sur le *Dynamic Time Warping* (Sakoe & Chiba, 1978) s'appuyant sur une représentation cepstrale des énoncés (incluant les coefficients de 2 à 13 et leur deltas). En raison de la présence fréquente du dévoisement à la fin de chaque essai (comprenant une répétition des trois phrases), nous avons écarté les valeurs de f0 extraites des voyelles finales des essais.

3.1 Effet du DAF sur la durée des voyelles

La première analyse a été menée pour estimer les effets du DAF sur la durée des voyelles (voir la figure 1), afin de vérifier si les effets de cette perturbation étaient cohérents avec ceux observés dans la littérature. Cette analyse a été menée sur l'ensemble des participantes. Nous avons testé l'interaction entre l'effet du DAF, l'effet de la présence d'un accent et l'effet de la phrase par le biais d'une régression par modèles mixtes, séparément pour chaque groupe de locutrices. Aussi bien dans ce modèle, que dans les modèles décrits dans les sections suivantes, les interactions non significatives qui ne contribuaient pas de manière significative à l'adéquation du modèle aux données ont été exclues. La contribution d'une interaction à l'adéquation du modèle a été évaluée en comparant, au moyen d'un test de $\chi 2$, les résidus du modèle obtenus avec et sans l'interaction. Chaque modèle comprenait un intercept aléatoire par locutrice et une pente spécifique à la locutrice pour l'effet de chaque prédicteur. Dans tous les modèles qui suivent, l'effet du DAF a été codé avec un contraste par différences successives, tandis que l'effet des phrases a été codé avec un contraste par écart à la moyenne, ce qui a donné lieu à un intercept correspondant au comportement moyen observé sur l'ensemble des phrases.

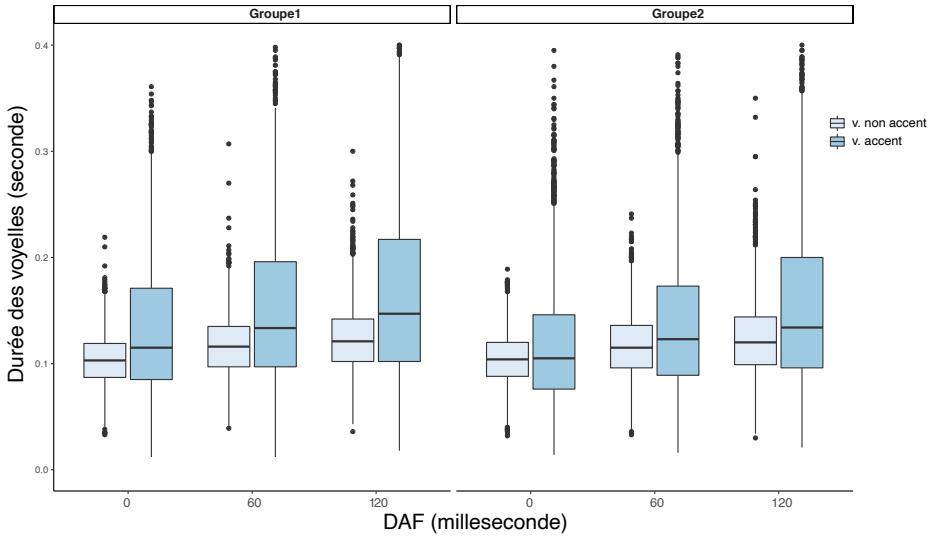


FIGURE 1 : Durée des voyelles accentuées et des voyelles non accentuées dans les conditions sans DAF (0 ms), avec le DAF de 60 ms ou de 120 ms du groupe 1 (qui est perturbé par le décalage positif de la f0 du feedback auditif) et groupe 2 (qui est perturbé par le décalage négatif de la f0 du feedback auditif)

Les voyelles accentuées sont généralement plus longues que les voyelles non accentuées pour le groupe 1 (estimée : 0,566, écart-type : 0,079, t-val. : 7,203, p-val. < 0,001) et le groupe 2 (estimée : 0,381, écart-type : 0,068, t-val. : 5.587, p-val. < 0,001). Les voyelles non accentuées sont allongées par le DAF de 60 ms par rapport à celles dans la condition sans DAF pour le groupe 1 (estimée : 0,202, écart-type : 0,046, t-val. : 4,444, p-val. < 0,001) et le groupe 2 (estimée : 0,239, écart-type : 0,063, t-val. : 3.818, p-val. : 0,001). Elles sont encore plus allongées par le DAF de 120 ms pour le groupe 1 (estimée : 0,113, écart-type : 0,046, t-val. : 2,468, p-val. : 0,019), mais non pas pour le groupe 2.

L'effet d'allongement est plus fort pour les voyelles accentuées avec le DAF de 60 ms à la fois pour le groupe 1 (estimée : 0,143, écart-type : 0,037, t-val. : 3.819, p-val. < 0,001) et pour le groupe 2 (estimée : 0,143, écart-type : 0,039, t-val. : 3.635, p-val. < 0,001). Ce même effet est aussi présent avec le DAF de 120 ms pour le groupe 1 (estimée : 0,159, écart-type : 0,038, t-val. : 4.218, p-val. < 0,001) et le groupe 2 (estimée : 0,191, écart-type : 0,040, t-val. : 4.802, p-val. < 0,001). Les voyelles dans la phrase qui est principalement composée des syllabes CV sont plus longues que celles dans les autres deux phrases, quelles qu'elles soient, non accentuées (estimée : 0,154, écart-type : 0,039, t-val. : 3.946, p-val. < 0,001) et cette différence augmente dans le cas des syllabes accentuées (estimée : 0,060, écart-type : 0,022, t-val. : 2.762, p-val. : 0,006) pour le groupe 1. L'effet de la structure syllabique est présent aussi dans les productions du groupe 2, mais n'augmente pas dans le cas des voyelles accentuées (estimée : 0,131, écart-type : 0,042, t-val. : 3.087, p-val. : 0,005). En revanche, les voyelles non accentuées dans la phrase qui est principalement composée des syllabes CCV sont plus courtes pour le groupe 1 (estimée : -0,247, écart-type : 0,039, t-val. : -6.337, p-val. < 0,001) et le groupe 2 (estimée : -0,261, écart-type : 0,042, t-val. : -6.160, p-val. < 0,001). Mais cet effet est réduit pour les voyelles accentuées aussi bien pour le groupe 1 (estimée : 0,263, écart-type : 0,022, t-val. : 12.165, p-val. < 0,001) que pour le groupe 2 (estimée : 0,293, écart-type : 0,023, t-val. : 12.857, p-val. < 0,001). En général, pour le groupe 1, la structure syllabique n'as pas d'impact sur l'effet du DAF et cela indépendamment de la présence de l'accent. Cependant pour le groupe 2, l'effet de DAF de 60 ms est moins important sur les voyelles non accentuées de la phrase qui est principalement composée des syllabes CCV (estimée : -0,060, écart-type : 0,027, t-val. : -2.190, p-val. < 0,029).

3.2 Effet du décalage de la f0 sur la f0 observée pour chaque locutrice

Afin d'évaluer les réponses des locutrices à la perturbation de la f0, nous avons effectué une régression linéaire pour chaque locutrice. Dans ces modèles, la variable dépendante était la valeur médiane de la f0 de chaque voyelle. Nous avons comparé les valeurs observées dans la condition de contrôle (enregistrées avant l'exposition aux perturbations de la f0) aux valeurs observées dans les blocs avec les perturbations d'un décalage positif ou négatif de la f0. Seules les voyelles produites au cours des essais sans DAF ont été prises en compte. En plus de l'effet du décalage de la f0, nous avons testé l'effet de l'accent et de la phrase, ainsi que les effets des interactions doubles et triples entre les prédicteurs. Dans le groupe 1, quatre locutrices montrent un effet significativement positif du décalage de la f0 quelle que soit la plage du décalage et à la fois dans les voyelles accentuées et non accentuées, dont une montre un effet plus fort dans les voyelles accentuées ; Une locutrice montre un effet significativement positif seulement lorsque la plage de la perturbation de la f0 était plus importante. Deux locutrices ont significativement modifié leur f0 en direction opposée de la perturbation. Dans le groupe 2, trois locutrices montrent un effet significativement négatif du décalage de la f0 quelle que soit la plage de décalage et à la fois dans les voyelles accentuées et non accentuées ; Une locutrice montre un effet significativement négatif seulement lorsque la plage de la perturbation était moins importante ; Une locutrice montre le même effet seulement lorsque la plage était plus importante ; Une locutrice montrent un effet significativement négatif quelle que soit la plage de décalage seulement pendant la production des voyelles accentuées. Pendant la production des voyelles non accentuées, cette locutrice montre ce même effet seulement lorsque la plage de la perturbation était moins importante. Une locutrice a significativement augmenté sa f0 en direction opposée de la

perturbation. En général, 55% des locutrices ont suivi la direction du décalage de la f0 au moins pendant la production des voyelles accentuée (voir dans la figure 2 la valeur médiane de la f0 observée dans les essais avec le décalage de la f0 mais sans DAF par rapport à celle observée dans les essais de contrôle de toutes les locutrices de tous les deux groupes). 30% des locutrices ne sont pas significativement affectées par le décalage de la f0. Seulement 15% des locutrices ont compensé le décalage de la f0.

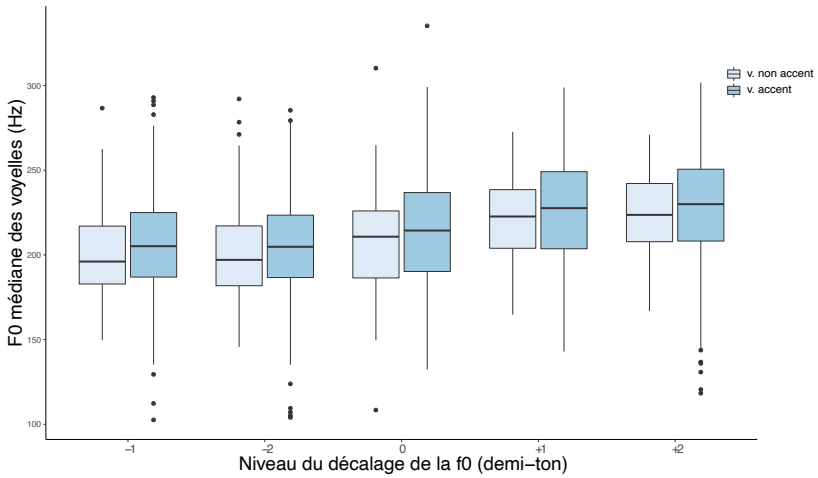


FIGURE 2 : La f0 médiane des voyelles accentuées et non accentuées des essais de contrôle (0) et des essais avec le décalage de la f0 négatif (-1, -2) ou avec le décalage de la f0 positif (+1, +2) du feedback auditif mais sans le retard du feedback auditif

3.3 Interactions entre les facteurs manipulés sur la durée des voyelles

Afin d'analyser les interactions potentielles entre les effets du décalage de la f0 et l'effet du DAF chez les locutrices qui ont suivi la perturbation de la f0, nous avons analysé comment la durée des voyelles était affectée par le décalage de la f0, la quantité du DAF, la phrase et par leurs interactions par le biais d'une régression par modèles mixtes, séparément pour chaque groupe de locutrices et pour les deux positions syllabiques (voyelles accentuées et non accentuées). Nous avons exclu les essais de contrôle. Chaque modèle (un modèle par groupe de locutrices et par position syllabique) comprenait un intercept aléatoire par locutrice et une pente aléatoire spécifique à la locutrice pour chaque prédicteur considéré. Les résultats des deux groupes sont assez homogènes. Les voyelles non accentuées de la phrase qui est principalement composée des syllabes CCV sont plus courtes à la fois pour le gourpe1 (estimée : -0,531, écart-type : 0,106, t-val. : -5.033, p-val. : 0,001) et pour le groupe 2 (estimée : -0,473, écart-type : 0,099, t-val. : -4.754, p-val. < 0,001). Les voyelles non accentuées sont allongées par le DAF de 60 ms par rapport à celles dans la condition sans DAF aussi bien pour le groupe 1 (estimée : 0,420, écart-type : 0,124, t-val. : 3.387, p-val. : 0,010) que pour le groupe 2 (estimée : 0,417, écart-type : 0,103, t-val. : 4.048, p-val. : 0,002). Cependant, pour tous les deux groupes, l'effet du DAF de 120 ms sur les voyelles non accentuées n'est pas significativement différent que celui de 60 ms. L'effet du DAF de 120 ms est réduit pour les voyelles non accentuées de la phrase CCV pour le groupe 1 (estimée : -0,132, écart-type : 0,054, t-val. : -2.454, p-val. : 0,014)

et aussi pour le groupe 2 (estimée : -0,139, écart-type : 0,049, t-val. : -2.832, p-val. : 0,005). Les voyelles accentuées sont allongées par le DAF de 60 ms par rapport à celles dans la condition sans DAF pour le groupe 1 (estimée : 0,287, écart-type : 0,073, t-val. : 3.949, p-val. : 0,004) et le groupe 2 (estimée : 0,353, écart-type : 0,079, t-val. : 4.486, p-val. < 0,001). Ces voyelles sont encore plus allongées par le DAF de 120 ms pour le groupe 1 (estimée : 0,231, écart-type : 0,073, t-val. : 3.171, p-val. : 0,013) et le groupe 2 (estimée : 0,239, écart-type : 0,079, t-val. : 3.023, p-val. : 0,006). Les voyelles accentuées de la phrase qui est principalement composée des syllabes CV sont plus longues que celles dans les autres deux types de phrase (estimée : 0,217, écart-type : 0,086, t-val. : 2.521, p-val. : 0,030). Pour les deux groupes, le décalage de la f0 et le DAF n'ont pas un effet interactif sur la durée des voyelles des trois types de phrase, indépendamment de la présence de l'accent.

4 Discussion

Contrairement à la plupart des études précédentes indiquant que les locuteurs produisaient surtout des commandes motrices compensatoires en réponse à des perturbations inattendues de la f0 du feedback auditif, dans cette étude, plus de la moitié des participantes ont tendance à suivre la direction du décalage de la f0, autrement dit, à adapter leurs commandes motrices au feedback auditif afin de stabiliser leur f0 lorsque le feedback auditif est perturbé de façon continue pendant la production de phrases. En plus, cette adaptation se produit rapidement. Plusieurs études montrent que dans la parole synchronisé (lorsque deux locuteurs produisent les mêmes énoncés de façon simultanée), un locuteur tend à imiter la voix qu'il entend. Cela se produit même lorsque la voix entendue masque la voix du locuteur (Zheng *et al*., 2011). De plus il a été démontré que dans les tâches de parole synchronisée, la suppression des activations du cortex auditif, qui est interprétée comme un index de la perception de notre propre voix, était absente (Jasmin *et al*., 2016). En combinant le retard du feedback auditif avec la perturbation de la f0, nous avons produits des conditions d'énonciation qui se rapprochent de la parole synchronisée (ou les locuteurs ne sont jamais parfaitement synchronisés). Il est donc raisonnable d'en déduire que nos participantes n'ont pas traité le signal acoustique perçu comme le produit de leur propre activité phonatoire, mais comme celle d'une autre locutrice. Il est donc possible que les locutrices aient intégré le signal entendu lors de la production des phrases dans leur modèle prédictif interne pour anticiper le déroulement au fil du temps du signal entendu (Pickering & Garrod, 2013). Notre recherche fournit donc un nouveau paradigme pour explorer le sens de l'agentivité en parole et les résultats obtenus nous montrent que l'attribution de l'agentivité dans la production de la parole est flexible et dépendante des conditions d'énonciation.

Conformément à ce qui a été observé dans les études précédentes (ex : Kalveram & Jäncke, 1989), en réponse au retard du feedback auditif, les locuteurs allongent leurs syllabes, surtout les syllabes accentuées. Cette réponse a pour but de réduire le décalage entre les conséquences attendues des commandes motrices et les caractéristiques du signal acoustique. L'étude de Kalveram et Jäncke (1989) a cependant analysé la production de logatomes isolés composés uniquement de syllabes CV. Par conséquent, la sensibilité accrue des syllabes accentuées au DAF, par rapport aux syllabes non accentuées, pourrait être due aussi bien au fait que les voyelles accentuées sont par nature plus longues que les non accentuées, qu'à leur position dans la structure métrique. En demandant à nos participants

de prononcer des phrases dans lesquelles la complexité syllabique a été systématiquement manipulée, nous avons pu montrer que la sensibilité accrue des voyelles accentuées au DAF est principalement due à leur position dans la structure métrique. Car autrement on aurait dû observer un effet de la structure syllabique, étant donné que les voyelles contenues dans les syllabes CV sont plus longues que celles contenues dans des syllabes dont la structure est plus complexe. De plus, en demandant aux locuteurs de prononcer des phrases contenant plusieurs syntagmes accentuels, nous avons montré que la réponse au DAF est modulée par la présence d'un accent au niveau du syntagme accentuel. Ces résultats ont des implications importantes sur le rôle de la structure métrique et du phrasé dans le contrôle moteur de la parole qui méritent des recherches supplémentaires. Pour les locutrices qui ont suivi la direction du décalage de la f0 du feedback auditif, le décalage de la f0 et le DAF n'ont pas d'effets interactifs sur la durée des voyelles des trois types de phrase. Cela pourrait signifier que le fait d'intégrer le feedback auditif dans leur système de prédiction pendant la production de la parole ne suffit pas à les aider à mieux résister au DAF. Cependant, nous observons une absence de l'effet de renforcement du DAF de 120 ms par rapport celui de 60 ms sur les voyelles non accentuées pour ces locutrices. Il sera donc intéressant de comparer les résultats de ces deux groupes de locutrices aux groupes de locutrices qui produisent ces mêmes phrases avec un seul type de perturbation du feedback auditif (soit le DAF, soit le décalage de la f0).

Remerciements

Cette étude a été financée par le projet « MoSpeeDi. Motor Speech Disorders : characterizing phonetic speech planning and motor speech programming/execution and their impairments », subside CRSII5_173711/1 Sinergia du Fond National Suisse de la Recherche Scientifique et par le programme "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL).

Références

BEHROOZMAND, R., KORZYUKOV, O., SATTLER, L., & LARSON, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control. *JASA*, 132(4), 2468–2477. DOI : doi.org/10.1121/1.4746984

BURNETT, T. A., FREEDLAND, M. B., LARSON, C. R., & HAIN, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *JASA*, 103(6), 3153–3161. DOI : doi.org/10.1121/1.423073

CHEN, S. H., LIU, H., XU, Y., & LARSON, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *JASA*, 121(2), 1157–1163. DOI : doi.org/10.1121/1.2404624

DONATH, T. M., NATKE, U., & KALVERAM, K. T. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *JASA*, 111(1), 357–366. DOI : doi.org/10.1121/1.1424870

FRANKEN, M. K., ACHESON, D. J., MCQUEEN, J. M., HAGOORT, P., & EISNER, F. (2018). Opposing and following responses in sensorimotor speech control : Why responses go both ways. *Psychonomic bulletin & review*, 25(4), 1458–1467. DOI : doi.org/10.3758/s13423-018-1494-x

FRANKEN, M. K., ACHESON, D. J., MCQUEEN, J. M., HAGOORT, P., & EISNER, F. (2019). Consistency influences altered auditory feedback processing. *Quarterly Journal of Experimental Psychology*, 72(10), 2371–2379. DOI : doi.org/10.1177/1747021819838939

HAIN, T. C., BURNETT, T. A., LARSON, C. R., & KIRAN, S. (2001). Effects of delayed auditory feedback (DAF) on the pitch-shift reflex. *JASA*, 109(5), 2146–2152. DOI : doi.org/10.1121/1.1366319

JASMIN, K. M., MCGETTIGAN, C., AGNEW, Z. K., LAVAN, N., JOSEPHS, O., CUMMINS, F., & SCOTT, S. K. (2016). Cohesion and joint speech : Right hemisphere contributions to synchronized vocal production. *Journal of Neuroscience*, 36(17), 4669–4680. DOI : doi.org/10.1523/JNEUROSCI.4075-15.2016

KALVERAM, K. T., & JÄNCKE, L. (1989). Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition. *Folia Phoniatrica*, 41(1), 30–42. DOI : doi.org/10.1159/000265930

LARSON, C. R., BURNETT, T. A., KIRAN, S., & HAIN, T. C. (2000). Effects of pitch-shift velocity on voice F0 responses. *JASA*, 107(1), 559–564. DOI : doi.org/10.1121/1.428323

LARSON, C. R., SUN, J., & HAIN, T. C. (2007). Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F0 and amplitude control. *JASA*, 121(5), 2862–2872. DOI : doi.org/10.1121/1.2715657

LIU, H., WANG, E. Q., METMAN, L. V., & LARSON, C. R. (2012). Vocal responses to perturbations in voice auditory feedback in individuals with Parkinson's disease. *PloS one*, 7(3). DOI : doi.org/10.1371/journal.pone.0033629

MAX, L., & MAFFETT, D. G. (2015). Feedback delays eliminate auditory-motor learning in speech production. *Neuroscience letters*, 591, 25–29. DOI : doi.org/10.1016/j.neulet.2015.02.012

NATKE, U., & KALVERAM, K. T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *Journal of Speech, Language, and Hearing Research*. DOI : doi.org/10.1044/1092-4388(2001/045)

PICKERING, M. J., & GARROD, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347. DOI : doi.org/10.1017/S0140525X12001495

SAKOE, H., & CHIBA, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1), 43–49. DOI : 10.1109/TASSP.1978.1163055

XU, Y., LARSON, C. R., BAUER, J. J., & HAIN, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *JASA*, 116(2), 1168–1178. DOI : doi.org/10.1121/1.1763952

ZHENG, Z. Z., MACDONALD, E. N., MUNHALL, K. G., & JOHNSRUDE, I. S. (2011). Perceiving a stranger's voice as being one's own : A 'rubber voice' illusion? *PloS one, 6*(4). DOI : doi.org/10.1371/journal.pone.0018655

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Prédiction continue de la satisfaction et de la frustration dans des conversations de centre d'appels

Manon Macary[1, 2] Marie Tahon[1] Yannick Estève[3] Anthony Rousseau[2]
(1) LIUM, Le Mans, France
(2) Allo-Media, Paris, France
(3) LIA, Avignon, France
m.macary@allo-media.fr, marie.tahon@univ-lemans.fr,
yannick.esteve@univ-avignon.fr, a.rousseau@allo-media.fr

RÉSUMÉ

Nous présentons un nouveau corpus, nommé AlloSat, composé de conversations en français extraites de centre d'appels, annotées de façon continue en frustration et satisfaction. Dans le contexte des centres d'appels, une conversation vise généralement à résoudre la demande de l'appelant. Ce corpus a été mis en place afin de développer de nouveaux systèmes capables de modéliser l'aspect continu de l'information sémantique et para-linguistique au niveau conversationnel. Nous nous concentrons sur le niveau para-linguistique, plus précisément sur l'expression des émotions. À notre connaissance, la plupart des corpus émotionnels contiennent des annotations en catégories discrètes ou dans des dimensions continues telles que l'activation ou la valence. Nous supposons que ces dimensions ne sont pas suffisamment liées à notre contexte. Pour résoudre ce problème, nous proposons un corpus permettant une connaissance en temps réel de l'axe frustration/satisfaction. AlloSat regroupe 303 conversations pour un total d'environ 37 heures d'audio, toutes enregistrées dans des environnements réels, collectées par Allo-Media (une société spécialisée dans l'analyse automatique d'appels). Les premières expériences de classification montrent que l'évolution de l'axe frustration/satisfaction peut être prédite automatiquement par conversation.

ABSTRACT

AlloSat : A New Call Center French Corpus for Affect Analysis

We present a new corpus, named AlloSat, composed of real-life call center conversations in French, continuously annotated in frustration and satisfaction. This corpus has been set up to develop new systems able to model the continuous aspect of semantic and paralinguistic information at the conversation level. The present work focuses on the paralinguistic level, more precisely on the expression of emotions. In the call center industry, the conversation usually aims at solving the caller's request. As far as we know, most emotional databases contain annotations in discrete categories or in dimensions such as activation or valence. We hypothesize that these dimensions are not task-related enough. To solve this issue, we propose a corpus enabling a real-time investigation of the axis frustration / satisfaction. AlloSat regroups 303 conversations with a total of approximately 37 hours of audio, all recorded in real-life environments collected by Allo-Media (an intelligent call tracking company). First classification experimentsshow that the evolution of frustration / satisfaction axis can be retrieved automatically at the conversation level.

MOTS-CLÉS : Corpus, Reconnaissance des Émotions, Centre d'appels, Satisfaction / Frustration.

KEYWORDS: Speech Corpus, Emotion Recognition, Call center, Frustration / Satisfaction.

Volume 1 : Journées d'Études sur la Parole, pages 379–387.
HAL : hal-02798561.

1 Introduction

Aujourd'hui, alors que nous sommes capables de stocker de plus en plus de données et notamment des données audio, leur valorisation est un sujet urgent. De nombreuses études se penchent donc sur l'extraction d'information, par exemple sémantique et para-linguistique. Dans les centres d'appels, des conseillers humains reçoivent des centaines d'appels par jour et tentent de répondre au mieux aux problématiques des appelants. Nous avons donc décidé de travailler sur la reconnaissance des émotions, appelé Speech Emotion Recognition (SER), car dans un contexte de centre d'appel, fournir à posteriori l'évolution de la satisfaction et de la frustration lors d'une conversation dans des gros volumes d'appels, peut intéresser l'entreprise afin d'améliorer sa qualité de service.

Si nous nous référons aux travaux menés en psychologie, il y a plusieurs modèles utilisés pour caractériser une émotion. Une approche consiste à décomposer les émotions en catégories discrètes telles que les " Big Six " (Ekman, 1999) ou les 32 émotions de la roue de Plutchik (Plutchik, 1980). Une autre approche décrit l'émotion comme un état continu, décrit par plusieurs dimensions notamment l'activation et la valence, mais aussi la dominance, l'intention ou l'axe conducteur/obstructif (Scherer, 2005) dont l'extrême positif est la satisfaction et l'extrême négatif est proche de la frustration.

Si on se replace dans notre contexte, l'objectif d'un appel est soit d'ouvrir un contrat, soit de résoudre des problèmes techniques ou financiers. Du coup la question de l'évolution de la frustration et de la satisfaction (appelé satisfaction par la suite) est cruciale. Nous avons donc cherché des corpus disponibles pour la reconnaissance d'émotion dans notre contexte. Les corpus existants sont souvent actés et ne sont généralement pas liés à des conversations de centre d'appels. Même si de nombreux efforts sont faits pour passer de corpus actés à des corpus de parole spontanée, il existe encore peu de corpus de parole spontanée et encore moins qui sont annotés en émotion continue. Dans les corpus SER, l'émotion est principalement représentée par des catégories discrètes, par exemple la colère, la joie dans des conversations de centre d'appels d'urgence (Devillers *et al.*, 2010) probablement parce que la collecte d'émotions discrètes est plus aisée que celle d'émotions continues (Campbell, 2008). On retrouve néanmoins des corpus comportant des annotations continues : l'activation et la valence pour le corpus SEMAINE (McKeown *et al.*, 2012) (composé de conversations simulées entre un utilisateur humain et une machine) et SEWA (Kossaifi *et al.*, 2019) (composé de conversations entre deux locuteurs à propos de publicités visualisées en amont).

Nous avons donc cherché à proposer un schéma d'annotation et de procéder à l'annotation d'un corpus existant en émotions continues. Cependant les corpus des centres d'appels sont en général très dépendants d'un domaine d'activité, par exemple on retrouve DECODA (Lailler *et al.*, 2016) (opérateur de transport parisien) qui est annoté avec des entités nommées, ou NATURAL (Morrison *et al.*, 2007) (compagnie d'électricité chinoise) qui est annoté avec deux classes : colère et neutre. À notre connaissance, aucun corpus de centre d'appels n'est composé de domaines hétéroclites.

N'ayant pas trouvé de corpus de centre d'appels suffisamment diversifié et annoté continûment en satisfaction de l'appelant, nous avons collecté des données provenant de différents domaines sur lequel nous avons mis en place notre propre schéma d'annotation en collaboration avec la société Allo-Media. Nous proposons donc un nouveau corpus, dédié à l'analyse de conversations en centres d'appels, annoté continûment en satisfaction. Les premières prédictions réalisées sur ce corpus nous montre des résultats encourageants qui seront développés au cours de cet article.

Le reste de cet article est organisé comme suit. La construction du corpus est introduite dans la section 2. La section 3 se concentre sur l'analyse de la cohérence des résultats tandis que la section 4 montre la mise en place des systèmes de prédiction et les résultats des premières expériences.

2 Construction du corpus

2.1 Contexte général

Le corpus est composé de conversations téléphoniques en français entre des interlocuteurs (les appelants) et des agents (les conseillers) où les locuteurs sont des adultes qui demandent des informations. Diverses informations sont demandées par les appelants : il peut s'agir de création de contrats, de demande d'informations globales sur l'entreprise, de plaintes, etc. Toutes les conversations ont été enregistrées entre juillet 2017 et novembre 2018 dans des centres d'appels situés dans des pays francophones. Les conseillers sont employés de diverses sociétés dans différents domaines. On retrouve notamment des entreprises du secteur de l'énergie, de la santé, du voyage, de la vente immobilière et de l'assurance.

2.2 Collecte de données

Comme nous avons récupéré un très grand nombre d'appels, nous avons dû décider quelles conversations devaient être annotées. En effet, nous ne pouvions pas annoter tous les appels reçus pendant la période de captation en raison du coût et du temps nécessaires pour traiter une telle quantité de données. De plus, nous savons que toutes les conversations ne sont pas porteuses d'émotions et encore moins de satisfaction, nous avons donc dû sélectionner des conversations. Nous avons donc mis en place trois critères pour sélectionner les conversations :

— **La durée** : nous avons décidé de ne prendre que des conversations de plus de 30 secondes contenant au moins trois tours de parole ;

— **Écart type (STD) de la fréquence fondamentale (F_0)** : extraite avec l'algorithme YAPPT (Zahorian & Hu, 2008) (adapté aux signaux téléphoniques), elle est un marqueur utilisé pour la détection des émotions. Cela nous a permis de conserver 500 conversations qui maximisaient l'écart-type F_0.

— **Score de valence** : calculé sur les transcriptions des conversations à l'aide du dictionnaire français FAN (Monnier & Syssau, 2014). Ce dictionnaire contient une valeur de polarité (entre 0 et 10) pour plus de 1000 mots français. Le score de valence est la valeur moyenne calculée pour chaque conversation.

Une vérification manuelle des conversations sélectionnées automatiquement a permis de sélectionner 253 enregistrements susceptibles de contenir des informations émotionnelles.

Comme la plupart des conversations en centre d'appels ne véhiculent pas d'émotions, nous avons ajouté 50 conversations neutres, sélectionnées au hasard afin que notre corpus reste cohérent avec la réalité des centres d'appel. Cette procédure aboutit à une base de données contenant 303 conversations.

2.3 Pré-traitement audio

Les deux canaux audio (interlocuteur et agent) ont été séparés, ce qui nous permet d'avoir des documents distincts pour l'appelant et l'agent. Pour des raisons éthiques et commerciales, le canal de l'agent a été supprimé. Par conséquent, le corpus contient uniquement la voix des appelants sans aucun chevauchement de signal des locuteurs. Puisque nous n'avons pas conservé la réponse de l'agent, il peut y avoir de longs moments de silence dans nos données. Afin de minimiser l'effort des annotateurs, nous avons décidé de remplacer ces silences par 2 secondes de bruit blanc, permettant

aux annotateurs d'identifier des silences potentiellement plus longs. Les conversations durent entre 32 secondes et 41 minutes, comme indiqué dans le tableau 1.

Il n'y a généralement qu'un seul locuteur par conversation. Au total, nous avons 308 locuteurs répartis en 191 femmes et 117 hommes. Les principales caractéristiques du corpus sont résumées dans le tableau 1.

Statistiques	**Value**
nombre de conversations	303
nombre de locuteurs	308
nombre de femmes	191
nombre d'hommes	117
durée totale	37h23m27s
durée min conversation	32s
durée max conversation	41m
durée moyenne conv.	7m24s
transcription automatique	303

TABLE 1 – Caractéristiques principales du corpus

Toutes nos conversations ont également des transcriptions automatiques grâce à un système basé sur Kaldi (Povey *et al.*, 2011) appartenant à Allo-Media.

2.4 Anonymisation

Afin de préserver l'anonymat des locuteurs, les données personnelles sont obfusquées , afin de respecter le règlement général sur la protection des données (RGPD). Nous avons également anonymisé tout ce qui peut identifier une entreprise, notamment les marques et les produits. Les informations personnelles sont supprimées et remplacées par des entités nommées dans les transcriptions, ce qui nous permet de savoir de quel type de données personnelles il s'agissait, et par un son jazzy dans l'audio.

2.5 Annotation de la satisfaction

Afin d'effectuer l'annotation continue, nous avons adapté CARMA (Girard, 2014), un toolkit dérivé de FeelTrace (Cowie *et al.*, 2000) nous permettant de faire une annotation sur un axe : de la frustration à la satisfaction, en utilisant les flèches d'un clavier. Nous avons personnalisé les paramètres afin de les faire correspondre à notre schéma d'annotation : une échelle commençant à 0 (extrêmement frustré) allant jusqu'à 10 (extrêmement satisfait). L'axe est initialisée à 5, censée correspondre à l'état neutre. Les émotions sont principalement détectables dans la seconde (Schuller & Devillers, 2010) contrairement aux mots qui sont généralement étudiés par des fenêtres de 30ms. Nous avons donc choisi de récupérer la position du curseur sous forme d'annotation toutes les 0,25 secondes. L'annotation a été faite par 3 annotateurs, 2 femmes et 1 homme. Ils ont reçus un guide d'annotation afin d'homogénéiser les annotations et de réduire l'aspect subjectif de celle-ci. Une annotation discrète a également été réalisée pour le début et la fin de la conversation, afin de vérifier la cohérence des annotations continues.

Deux exemples d'annotations de la satisfaction sont donnés sur la Figure 1 où les valeurs observées d'accord inter-annotateur sont respectivement de 0,851 et de 0,732. Dans la conversation A, nous

pouvons observer que l'appelant passe d'un état neutre (5) à frustré (presque 0) et reste relativement frustré (1-2) jusqu'à la fin de l'appel. La conversation B correspond à l'une des conversations neutres choisies au hasard.

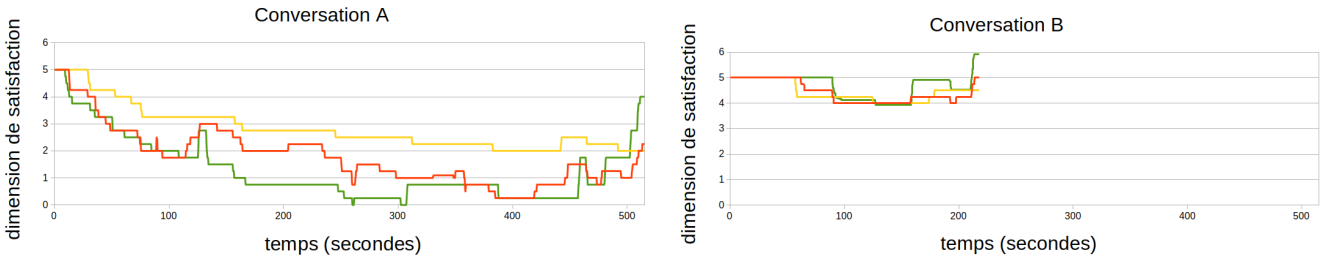


FIGURE 1 – Annotations continues de la satisfaction par les trois annotateurs pour deux conversations. Les étiquettes de fin discrètes sont "très frustrées" pour A et "neutres" pour B.

3 Analyse des données : cohérence des annotations

Afin d'évaluer l'accord inter-annotateur sur les annotations continues, nous avons utilisé le coefficient de corrélation. Ce coefficient est calculé au niveau de la conversation sur la satisfaction normalisée par rapport à l'ensemble des conversations entre les paires d'annotateurs. Les valeurs finales rapportées dans le tableau 2 montrent une bonne corrélation entre les annotateurs (en moyenne $R = 0,83$), ce qui signifie que les annotations continues sont cohérentes entre les annotateurs.

Paires d'annotateurs	Coefficient de corrélation R
a1-a2	0,82
a2-a3	0,87
a1-a3	0,80

TABLE 2 – Accords inter-annotateur par paire d'annotateur. a_i représente l'annotateur i.

L'une des raisons de ce fort accord est que le début de la conversation est presque toujours neutre. Cela peut s'expliquer de deux façons : d'abord, l'annotation continue est initialisée à 5, ce qui se traduit par un état neutre. Mais l'hypothèse principale est que l'interlocuteur est rarement frustré au début de l'appel : cette émotion est provoquée par les réponses de l'agent. Il en va de même pour la satisfaction. Comme nous le pensions, la plupart des conversations ont été perçues avec une frustration croissante, probablement parce que le conseiller n'est pas en mesure de donner une réponse suffisamment satisfaisante à l'interlocuteur.

En partant de ces résultats d'accord prometteurs, nous calculons une annotation de référence pour chaque conversation correspondant à la moyenne des trois annotations de la satisfaction et nous pouvons utiliser cette annotation de référence à des fins d'analyse et d'apprentissage. Cette annotation de référence est utilisée dans les expériences présentées par la suite.

4 Prédiction de la satisfaction

Comme nous l'avons dit dans l'introduction, notre objectif est de permettre de mieux comprendre l'état émotionnel des appelants dans un but d'analyse.Pour ce faire, il peut être utile d'avoir des indices sur la satisfaction de l'appelant tout au long de l'appel, et donc nous définissons une tâche de prédiction continue de cette dimension. Nous comparons deux modèles pour cette tâche. Le premier est le modèle de référence utilisé lors du challenge AVEC 2018 (Ringeval *et al.*, 2018). L'utilisation de ce modèle nous permettra de comparer les résultats obtenus sur AlloSat à ceux prédits sur SEWA (qui a été décrit en introduction), même si nous ne comparons pas exactement les mêmes dimensions, en effet AlloSat est annoté en satisfaction, alors que SEWA est annoté en valence. Le second est un modèle de réseau neuronal profond (DNN) avec des couches biLSTM (bidirectional Long Short Term Memory) déjà testé sur le corpus SEWA (Schmitt *et al.*, 2019). Ils utilisent tous deux des descripteurs audio comme entrée, que nous extrayons avec le framework OpenSMILE (Eyben *et al.*, 2010). Différents ensembles de descripteurs audio ont été testés afin de trouver celui qui était le plus adapté à notre corpus. Les modèles et les ensembles sont expliqués ci-dessous.

4.1 Descripteurs audio

Pour mieux comparer notre travail avec l'état de l'art dans le domaine du SER, nous avons décidé d'utiliser l'ensemble eGeMAPS (Eyben *et al.*, 2016). Cet ensemble a été conçu pour l'analyse automatique de la voix, en particulier l'analyse des émotions. Il contient 25 descripteurs de bas niveau (LLD) tels que le pitch, le jitter, les formants, etc. Une moyenne arithmétique et un écart-type (STD) sont calculés toutes les 0,1 secondes sur ces LLD. D'autres fonctions mathématiques calculées à partir de ces LLD sont également extraites pour un total de 88 descripteurs. Dans (Schmitt *et al.*, 2019) f_eGeMAPS a été défini avec 25 LLD et des fonctions mathématiques appliquées sur ces LLD (principalement moyenne et STD) extraits d'eGeMAPS totalisant 46 descripteurs. Un dernier descripteur, fonctionnant comme une détection de voix (voice activity detection i.e. vad), dénotant l'identité du locuteur (0 ou 1), est également incluse dans f_eGeMAPS.

Dans notre travail, les deux ensembles ont été extraits de nos données toutes les 0,25 seconde suivant le pas d'annotation d'AlloSat. Puisque nous ne gardons que le signal de l'appelant, nous modifions le vad pour indiquer si l'appelant parle (1) ou non (0). Le nombre de descripteurs des 4 ensembles est résumé dans le tableau 3.

4.2 Les architectures des DNN

4.2.1 Pré-traitement des inputs

Afin de s'aligner sur les architectures neuronales des articles de référence (Ringeval *et al.*, 2018; Kossaifi *et al.*, 2019), nous avons choisi d'utiliser une taille de séquence d'entrée fixe. Comme nous l'avons vu dans la section 3, les conversations ont des durées variant de 32 secondes à 41 minutes avec une moyenne (MOY) de 7m24s et un écart type (STD) de 4m58s. Habituellement, la taille d'entrée est fixée à $MOY + STD$ (ici 12m22s) pour couvrir plus de 95% du corpus. Les séquences longues sont alors coupées à la $MOY + STD$ tandis que les séquences courtes sont rallongées avec un padding. Afin de réduire l'effet du padding et la durée d'apprentissage, nous avons décidé de fixer la durée de la séquence d'entrée à 7 minutes. Nous avons appliqué un padding circulaire sur les courtes séquences.

Nous avons divisé notre corpus en trois sous-ensembles afin de respecter la répartition des conversations neutres : un apprentissage (201 conversations), un développement (42 conversations) et un test (60 conversations).

4.2.2 Les deux modèles neuronaux

Afin de pouvoir comparer nos résultats avec l'état de l'art, nous avons fait le choix de reproduire le système proposé dans le challenge AVEC 2018 (Ringeval *et al.*, 2018) sur la modalité "Cross-cultural Emotion". Le premier réseau neuronal est composé de 2 couches biLSTM de respectivement 64 et 32 unités. L'architecture bidirectionnelle est utilisée afin d'éviter les problèmes de délai d'annotation. En effet, il est possible que l'annotation présente des délais dans l'annotation, le temps que l'annotateur appuie sur les flèches du clavier ou qu'il décide s'il y a vraiment une variation à annoter.

Le second réseau est composé de 4 couches biLSTM comme décrit dans (Schmitt *et al.*, 2019). Les couches de biLSTM sont composées respectivement de 200, 64, 32, 32 unités.

Pour ces deux réseaux, la fonction d'activation utilisée est la fonction tangente hyperbolique. Un seul neurone de sortie est utilisé pour prédire une valeur toutes les 0,25 secondes.

4.3 Résultats des expériences

Les DNN sont implémentés avec le framework Keras [1] en utilisant Tensorflow [2]. L'apprentissage se fait par ensemble (batch) de 9 conversations en utilisant l'optimiseur ADAgrad. Le learning rate est initialisé à 0,001. Le nombre d'époques a d'abord été fixé à 500 avant d'être réduit à 200 puisque les réseaux ne s'amélioraient pas au delà d'environ 120 époques. Nous avons conservés les poids des réseaux donnant le meilleur score sur le développement afin de prédire les résultats sur le test. Le coefficient de corrélation de concordance (CCC) (Lin, 1989) a été utilisé comme fonction pour l'apprentissage du réseau et comme métrique d'évaluation pour déterminer le meilleur système. Ce score CCC varie de 0 (probabilité d'un tirage aléatoire) à 1 (corrélation parfaite).

Nous comparons deux réseaux appris sur deux axes émotionnels différents : la satisfaction avec AlloSat et la valence avec SEWA. Le tableau 3 donne un résumé des résultats obtenus avec les modèles et les sets de données étudiés.

		nb Descripteurs	**2 biLSTM**		**4 biLSTM**	
			dev	test	dev	test
SEWA	eGeMAPS	88	0.112*	-	-	-
(valence)	f_eGeMAPS	46	-	-	0,517*	0,410*
AlloSat	eGeMAPS	88	0,510	0,363	0,666	0,431
(satisfaction)	f_eGeMAPS	46	0,469	0,260	0,607	0,354
	eGeMAPS&vad	89	0,549	0,365	0,619	0,542
	f_eGeMAPS&vad	47	0,508	0,359	0,574	0,422

TABLE 3 – Résultats des expériences. *Ces résultats proviennent du challenge AVEC 2018 (Schmitt *et al.*, 2019)

Il semble que nous sommes en mesure de récupérer de bons scores de CCC pour notre corpus, comparables aux résultats de valence prédits sur le corpus SEWA. Le score CCC calculé sur l'ensemble

1. https ://keras.io
2. https ://www.tensorflow.org/

des données doit être pris avec précaution car, comme nous le montrons dans la Figure 2, le système est capable de faire de bonnes prédictions (conversation C) mais aussi de mauvaises prédictions (conversation D).

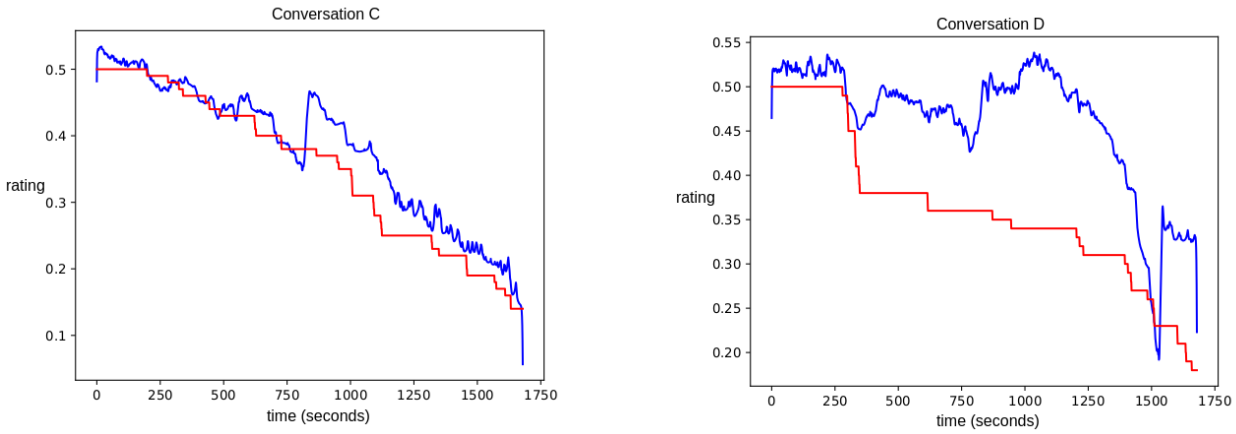


FIGURE 2 – Prédiction de la satisfaction sur des conversations issues du test. La référence est en rouge, la prédiction en bleu.

5 Conclusion

Dans cet article, nous présentons AlloSat, un nouveau corpus de conversations françaises en centre d'appels utilisable pour explorer la satisfaction (de la satisfaction à la frustration) dans de conversations téléphoniques réelles. Ce corpus contient 303 conversations pour un total de plus de 37 heures d'enregistrement et peut être obtenu en contactant les auteurs. L'objectif principal de ces travaux était de prédire la satisfaction tout au long d'une conversation. Cette prédiction ne pouvait être effectuée que si les annotations de ce nouveau corpus était cohérente, ce que nous avons vérifié. Les premières expériences montrent que les réseaux neuronaux biLSTM sont capables de prédire les valeurs de la satisfaction et donc de retracer cette dimension au cours d'un appel avec un score CCC correcte, comparable à celui calculé sur les prédictions de valence du corpus SEWA.

Par la suite, des investigations plus approfondies seront menées pour améliorer cette prédiction. Nous voulons notamment ajouter la modalité linguistique à nos descripteurs d'entrée. Nous prévoyons également d'aller plus loin dans nos expériences sur l'annotation continue et discrète en utilisant d'autres protocoles de classification (modèles, sets, niveaux de segmentation) tout en ajoutant des informations sémantiques supplémentaires.

Références

CAMPBELL N. (2008). *Expressive/Affective Speech Synthesis*, In *Springer Handbook of Speech Processing*, p. 505–518. Springer Berlin Heidelberg.

COWIE R., DOUGLAS-COWIE E., SAVVIDOU S., MCMAHON E. & AL. (2000). FEELTRACE : An instrument for recording perceived emotion in real time. In *ITRW on Speech and Emotion*, p. 19–24.

DEVILLERS L., VAUDABLE C. & CHASATGNOL C. (2010). Real-life emotion-related states detection in call centers : a cross-corpora study. In *Proc. of Interspeech*, p. 2350–2355.

EKMAN P. (1999). *Basic Emotions*, In *Handbook of Cognition and Emotion*, p. 301–320. Wiley, New-York.

EYBEN F., SCHERER K., SCHULLER B., SUNDBERG J. & AL. (2016). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, **7**(2), 190–202.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). OpenSMILE – the munich versatile and fast open-source audio feature extractor. In *Proc. of the ACM Multimedia 2010 International Conference*, p. 1459–1462.

GIRARD J. M. (2014). CARMA : Software for continuous affect rating and media annotation. *Journal of Open Research Software*, **2**(1), e5.

KOSSAIFI J., WALECKI R., PANAGAKIS Y., SHEN J., SCHMITT M. & AL. (2019). SEWA DB : A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence (Early Access)*.

LAILLER C., LANDEAU A., BÉCHET F., ESTÈVE Y. & AL. (2016). Enhancing the RATP-DECODA corpus with linguistic annotations for performing a large range of NLP tasks. In *Proc. of Language Resources and Evaluation Conference (LREC)*, p. 1047–1050.

LIN L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**(1), 255–268.

MCKEOWN G., VALSTAR M., COWIE R., PANTIC M. & AL. (2012). The SEMAINE Database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, **3**(1), 5–17.

MONNIER C. & SYSSAU A. (2014). Affective norms for French words (FAN). *Behavior research methods*, **46 4**, 1128–1137.

MORRISON D., WANG R. & DE SILVA L. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, **49**(2), 98–112.

PLUTCHIK R. (1980). Chapter 1 - a general psychoevolutionary theory of emotion. In *Theories of Emotion*, p. 3 – 33. Academic Press.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L. & AL. (2011). The kaldi speech recognition toolkit. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

RINGEVAL F., SCHULLER B., VALSTAR M., COWIE R. & AL. (2018). AVEC 2018 workshop and challenge : Bipolar disorder and cross-cultural affect recognition. In *Proc. of the 2018 on Audio/Visual Emotion Challenge and Workshop*, p. 3–13.

SCHERER K. R. (2005). What are emotions ? and how can they be measured ? *Social science information*, **44**(4), 695–729.

SCHMITT M., CUMMINS N. & SCHULLER B. W. (2019). Continuous emotion recognition in speech - do we need recurrence ? In *Proc. Interspeech 2019*, p. 2808–2812.

SCHULLER B. & DEVILLERS L. (2010). Incremental acoustic valence recognition : An inter-corpus perspective on features, matching, and performance in a gating paradigm. In *Proc. Interspeech 2010*, p. 801–804.

ZAHORIAN S. & HU H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, **123**, 4559–71.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Production de parole chez l'enfant porteur d'implant cochléaire : apport de la Langue française Parlée Complétée

Laura Machart[12], Anne Vilain[2], Hélène Lœvenbruck[1], Geneviève Meloni[123], Clarisse Puissant[2]
(1) Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France
(2) Univ. Grenoble Alpes, Grenoble INP, CNRS, GIPSA-lab, 38000 Grenoble, France
(3) Université de Montréal, Montréal, Canada
{laura.machart, anne.vilain, helene.loevenbruck, genevieve.meloni, clarisse.puissant}@univ-grenoble-alpes.fr

RESUME

La déficience auditive entraîne un retard sur le développement de la parole chez l'enfant sourd. La Langue française Parlée Complétée (LfPC), par le biais de 5 positions autour du visage et 8 configurations de la main, permet de rendre visibles tous les sons de la langue, sans confusion labiale. L'utilisation de ce système facilite la perception de parole et permet à l'enfant d'élaborer des représentations phonologiques stables. Cette étude s'intéresse à l'apport de la LfPC sur la production de parole chez l'enfant porteur d'implant cochléaire. A partir d'une tâche de dénomination d'images, nous observons que l'exposition à la LfPC (en perception) améliore significativement la production de parole chez l'enfant porteur d'implant cochléaire.

ABSTRACT

Speech production in children with cochlear implant(s): contribution of Cued French

Hearing impairment results in delayed speech development in deaf children. Cued French (LfPC) is a phonetic system, which can be used to support oral communication with deaf children: it supplements the auditory information with a manual cue. It has been shown to enhance speech perception and to help children build more stable phonological representations. In this study, we analyze the contribution of LfPC to the speech production abilities of children with cochlear implants. Using a picture naming task, we show that exposure to LfPC significantly improves speech production in children with cochlear implants.

MOTS-CLES : production de parole, déficience auditive, implant cochléaire, Langue française Parlée Complétée (LfPC)

KEYWORDS: speech production, deafness, cochlear implants, Cued Speech, Cued French

Volume 1 : Journées d'Études sur la Parole, pages 388–396.
HAL : hal-02798562.

1 Introduction

La déficience auditive est l'un des troubles sensoriels les plus fréquents et touche plus de 6,5% de la population mondiale. Les surdités congénitales représentent une naissance pour mille et dans 90% des cas, l'enfant sourd naît de deux parents normo-entendants (Berland, 2014). Chez l'enfant, la déficience auditive impacte et retarde le développement du langage : la perception de la parole étant lacunaire et déformée, l'enfant ne bénéficie pas d'un bain de langue suffisant. La perception, et donc la compréhension, sont altérées et les règles de construction de la langue ne peuvent pas être acquises de manière typique.

1.1 Perception de parole et déficience auditive

Dans un contexte de surdité profonde, la perception de parole se fait sans information auditive, c'est-à-dire uniquement à partir de l'information visuelle. La lecture labiale, qui repose sur le déchiffrage du mouvement des articulateurs de la parole, ne permet qu'une perception de parole limitée (Charlier & Leybaert, 2000) : hors contexte, seulement 10 à 30% d'un mot ou d'une phrase sont perçus (Bernstein et al, 2000). Les phonèmes peu visibles, les phénomènes de coarticulation ou encore les sosies labiaux (ex : /y/ et /u/) tout comme les variations bucco-faciales du locuteur ainsi que la qualité de l'image labiale sont autant de facteurs influençant la qualité de la perception. L'implant cochléaire, un des dispositifs de remédiation proposés en cas de surdité profonde, permet un accès de plus en plus précis aux sons de parole. Toutefois, l'information auditive transmise par l'implant cochléaire reste limitée, ce qui peut entraîner des troubles langagiers ultérieurs (Colin et al., 2017; Leybaert et al., 2010). Les recherches ont montré que les performances des enfants porteurs d'implant cochléaire restent en dessous de celles des enfants normo-entendants et que leurs représentations phonologiques demeurent peu détaillées (Nittrouer et al., 2018; Colin et al., 2017; Leybaert et al., 2010).

1.2 Perception de parole et Langue française Parlée Complétée

Afin de pallier le manque d'informations lié à la lecture labiale, certains font le choix de la Langue française Parlée Complétée (LfPC), équivalent francophone du *Cued Speech* (Cornett, 1967). Il s'agit d'un outil de réception du message oral qui aide la perception de la parole. L'association d'un geste manuel à la lecture labiale va rendre visibles tous les phonèmes de la langue. Le code LfPC se compose de 5 positions de la main autour du visage pour coder les sons vocaliques et de 8 configurations de la main pour coder les sons consonantiques du français (figure 1). Chaque position et chaque configuration de la main code plusieurs sons. La LfPC est donc indissociable de la lecture labiale : pour une même image labiale, on utilise différentes positions/configurations de la main et, à l'inverse, pour une même position/configuration de la main, l'image labiale est différente.

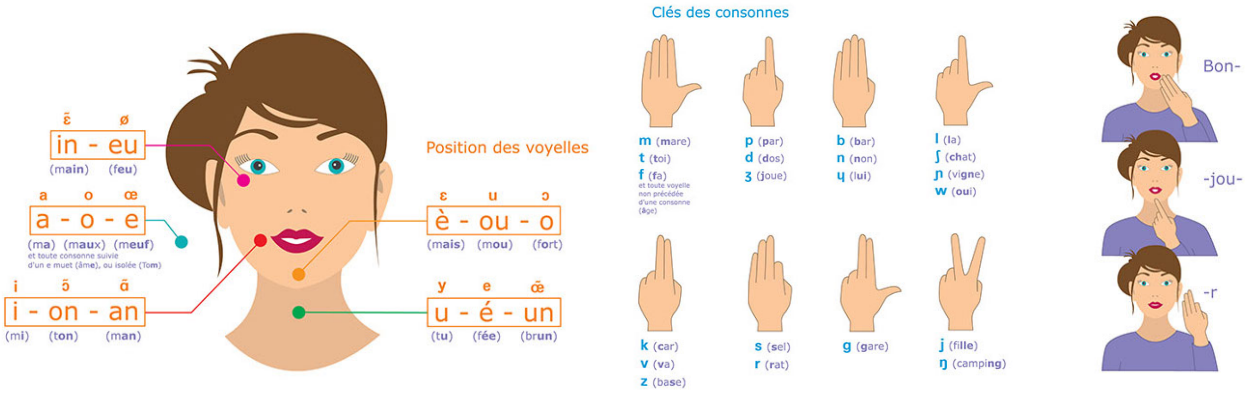


FIGURE 1: Positions et configurations de la main en LfPC (Source ALPC[1] https://alpc.asso.fr/les-cles-du-code-lpc/)

L'utilisation de la LfPC permet à la personne déficiente auditive de visualiser tous les sons du français sans ambiguïtés : la perception auditive est enrichie à l'aide de la perception visuelle. Il est à noter que dans la plupart des cas, ce sont les personnes entendantes (parents, enseignants, frères et sœurs) qui utilisent la LfPC pour permettre à l'enfant déficient auditif de percevoir le message oral. Les enfants eux-mêmes n'utilisent que très rarement le code en production.

Les recherches ont montré que la perception de syllabes, de mots et de phrases est améliorée avec l'utilisation du *Cued Speech*[2]. Plus l'exposition au code est prolongée plus les bénéfices sont observés (Bratakos et al., 1998; Périer et al., 1990). Par ailleurs, la précocité de l'exposition augmente encore davantage les capacités perceptives (Alegria et al., 1999). Enfin, la LfPC facilite l'émergence de représentations phonologiques stables (Charlier et Leybaert, 2000; Hage, et al., 1991) ce qui par conséquent améliore la perception auditive pure de la parole (Kos et al., 2008). En effet, chaque phonème du français étant systématiquement associé à une position ou configuration de la main, celle-ci fournit un indice complémentaire aux informations acoustiques pour distinguer et encoder les formes phonologiques. L'accroissement du lexique est favorisé (Hage, 1994) tout comme le développement de la morphosyntaxe grâce à la perception intégrale de la chaîne parlée (Hage & Leybaert, 2006).

1.3 Production de parole et Langue française Parlée Complétée

Très peu d'études se sont intéressées à l'apport de la LfPC à la production de parole. Or, comme mentionné ci-dessus, il a été mis en évidence que la LfPC facilite la mise en place des représentations phonologiques. De plus, Rvachew (1994) et Rvachew et al. (2004) ont mis en évidence que la production de parole est améliorée par l'entraînement perceptif chez les enfants présentant un retard

[1] Association pour la promotion de la Langue française Parlée Complétée

[2] Les études citées étant réalisées sur différentes langues, nous utilisons cette appellation générique

phonologique. Par conséquent, nous nous interrogeons ici sur le bénéfice éventuel apporté par la LfPC à la production de parole des enfants porteurs d'implant cochléaire (comme suggéré par les études présentées par Hage et Leybaert, 2006), et donc sur un possible transfert de ces représentations phonologiques vers la production de parole.

L'objectif de cette étude est d'analyser l'apport de la LfPC sur la production des phonèmes du français chez deux groupes d'enfants porteurs d'implant cochléaire : des enfants avec un faible niveau de décodage de la LfPC et des enfants avec un bon niveau de décodage de la LfPC. Pour ce faire, nous utilisons une tâche de dénomination d'images. Nous nous attendons à ce que l'exposition à la LfPC améliore la production de parole chez l'enfant porteur d'implant. Dans le cadre de cette expérience, il n'est pas demandé aux enfants d'utiliser la LfPC lors de la production, la LfPC n'est utilisée que pour la transmission des consignes.

2 Méthode

2.1 Tâche

Dans le cadre de cette étude, nous utilisons la tâche de dénomination d'images de la batterie EULALIES (Meloni et al., 2017). Cette batterie se compose de 5 tâches, chacune testant différents niveaux de traitement de la parole, en perception et en production. La tâche de dénomination d'images teste les représentations phonologiques en production : l'enfant voit des images sur l'écran d'ordinateur et doit produire spontanément le nom de chaque item. En cas de difficulté, l'expérimentateur propose soit un amorçage sémantique soit un amorçage phonologique. Lorsque l'enfant ne parvient pas à dénommer l'item présenté, et ce malgré les amorçages proposés, l'expérimentateur lui demande simplement de le répéter après lui. Cette tâche se compose de 68 mots[3] de 1 à 4 syllabes de différents degrés de complexité. Les items sélectionnés ont une haute fréquence d'occurrence dans la base de données Lexique.org (New et al., 2004) et sont accessibles même pour les plus petits (vêtements, animaux, objets du quotidien, aliments, moyens de transport, etc.). Ils incluent tous les phonèmes de la langue française, à différentes positions dans le mot.

2.2 Protocole

L'enfant est assis devant une table sur laquelle se trouvent un ordinateur et un microphone. Un micro-tête est installé afin d'enregistrer la voix de l'enfant. L'expérimentateur est installé à sa droite, face à l'écran d'ordinateur. La passation comporte des tests d'inclusion : une tâche d'empan visuel (PathSpan, Lefevre et al., 2010), une tâche d'empan de chiffres endroit (ODEDYS, Pouget, 2002) et le module morphosyntaxe en production du test ELO (Khomsi, 2008). Les 5 tâches de la batterie EULALIES, sont effectuées en débutant par la tâche de dénomination d'images. Pour les enfants normo-entendants, une audiométrie est réalisée pour éliminer un éventuel trouble de l'audition (perception à 20 dB sur les fréquences 250, 500, 1000, 2000, 4000 et 8000 Hz). Pour les enfants

[3] Un tiers des enfants a passé une première version de la tâche qui comportait 66 items

porteurs de déficience auditive, le niveau de décodage de la LfPC est évalué à partir du test TERMO (Busquet & Descourthieux, 2003). Deux niveaux sont déterminés : faible décodage (quelques syllabes sont décodées à vitesse lente) et bon décodage (décodage de mots isolés et de phrases simples à la vitesse de la parole). Un questionnaire de langage est rempli en amont par les parents afin de récolter des informations sur le développement langagier de chaque enfant (multilinguisme, âge du premier appareillage, catégorie socio-professionnelle des parents, etc.) et éliminer un possible trouble associé.

2.3 Matériel

L'enregistrement des données audio se fait à l'aide d'un enregistreur Zoom (H4n Pro). L'enfant porte un micro-tête SHURE (Beta 54R). L'audiométrie est réalisée à partir d'un audiomètre Electronica 9910. Les stimuli visuels sont présentés sur un ordinateur portable posé devant l'enfant.

2.4 Participants

Cette étude s'intéresse à 16 enfants avec surdité profonde âgés de 28 à 139 mois et porteurs d'implant cochléaire (groupe CI) (âge=98,44, écart-type=34,74). Ce groupe se compose de sept filles et neuf garçons. Treize enfants sur les 16 sont bi-implantés, les trois autres enfants portent une prothèse controlatérale. Tous les enfants CI sont en contact avec la LfPC. Un enfant du groupe CI sur les 16 est bilingue français/Langue des Signes française (LSF) (6,3%), quatre enfants bénéficient occasionnellement de la LSF en classe ou à la maison (25%) et le français signé[4] est utilisé ponctuellement pour quatre enfants (25%). Le groupe CI se divise en deux sous-groupes : huit enfants dont quatre filles avec un faible niveau de décodage de la LfPC (groupe LfPC-, âge=88,87, écart-type=40,74) et huit enfants dont trois filles avec un bon niveau de décodage de la LfPC (groupe LfPC+, âge=108, écart-type=26,78). Le groupe CI est mis en regard avec un groupe de 97 enfants avec audition typique (groupe NH) âgés de 35 à 135 mois et faisant partie de la large cohorte d'enfants typiques du projet EULALIES (Meloni et al. 2017) (âge=94,62, écart-type=22,33). Ce groupe est constitué de 58 filles et 39 garçons, et 40 enfants sont multilingues (41,23%). Les enfants du groupe NH n'ont aucune connaissance de la LfPC ni de la LSF. L'âge auditif est défini comme l'âge chronologique pour le groupe NH et comme le temps écoulé depuis le premier appareillage pour le groupe CI. La moyenne d'âge auditif du groupe LfPC- est 78,87 (écart-type=43,72) celle du groupe LfPC+ est 99 (écart-type=29,63). L'âge d'implantation moyen du groupe LfPC- est 21 (écart-type=8,02) et celui du groupe LfPC+ est 23,37 (écart-type=13,07). Les participants ont été recrutés dans les écoles de l'agglomération grenobloise et lors de stages d'été organisés par l'ALPC en 2018 et 2019, à destination des familles de la France entière. Cette étude a été approuvée par le Comité d'Ethique pour les Recherches Grenoble Alpes (CER Grenoble Alpes-Avis-2018-04-03-2-Amendement).

[4] Utilisation de signes de la LSF avec la syntaxe du français

2.5 Traitement des données

Tous les items produits à la tâche de dénomination sont transcrits et traités à partir du logiciel PHON (Hedlund & Rose, 2016). Une partie des données a été annotée en double aveugle avec un accord inter-juge de 82,4%. Une transcription multi-juges (5 transcripteurs) a ensuite permis de définir des consensus sur la transcription, les variations libres étant admises comme des productions correctes. Après alignement avec la cible, le nombre d'erreurs par mot est extrait avec PHON (somme des substitutions, élisions et épenthèses) puis la moyenne du nombre d'erreurs par sujet est calculée.

3 Résultats

La figure 2 représente le nombre d'erreurs moyen par mot en fonction du niveau de décodage de la LfPC et de l'âge chronologique pour les enfants des groupes CI et NH. Nous observons que les enfants CI avec un faible niveau de décodage de la LfPC semblent faire plus d'erreurs que les enfants CI avec un bon niveau de décodage de la LfPC, dont les résultats sont plus proches de ceux des enfants NH.

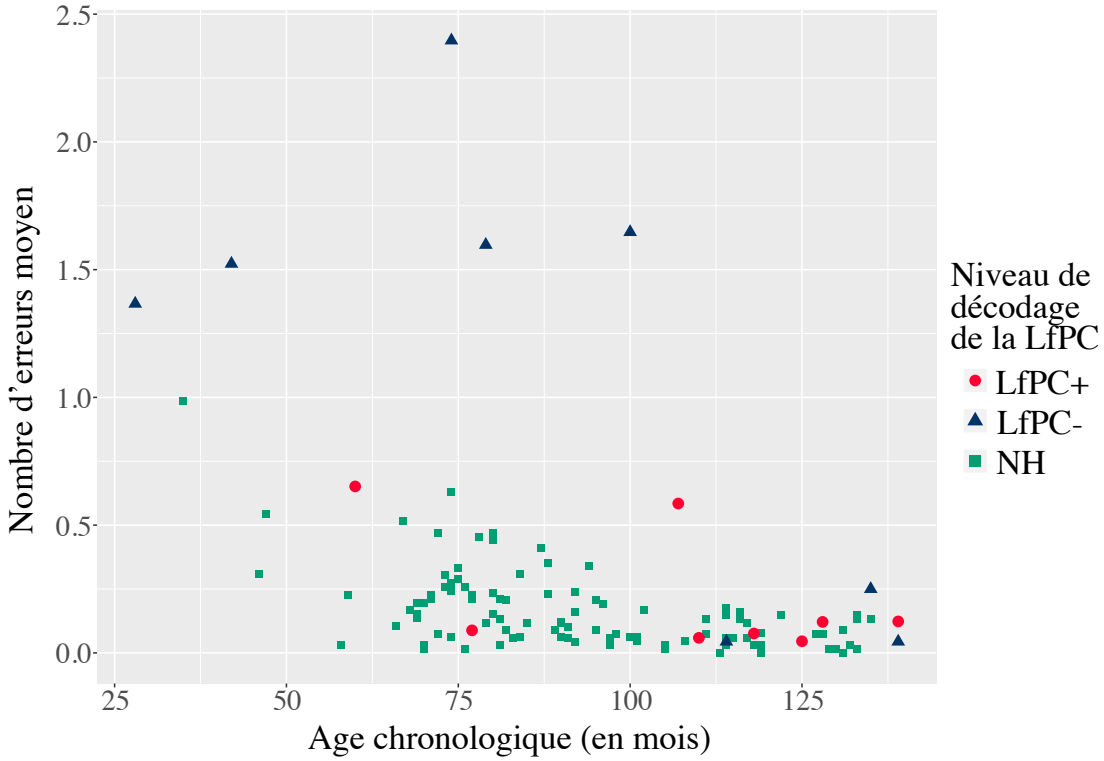


FIGURE 2: Nombre d'erreurs moyen pour les enfants CI et NH en fonction du niveau de décodage de la LfPC et de l'âge chronologique

Nous avons analysé l'effet du niveau de décodage de la LfPC sur les scores en production. Nos analyses statistiques révèlent que les tendances observées sur la figure 2 sont significatives. En effet, à partir d'un modèle linéaire incluant les facteurs principaux de groupe (LfPC-, LfPC+), d'âge

chronologique, d'âge d'implantation et de sexe ainsi que leurs interactions, une sélection par comparaison descendante de modèles a abouti au modèle incluant finalement le groupe (p=.029) et l'âge chronologique (p=.002). L'interaction entre le groupe et l'âge chronologique n'est pas significative.

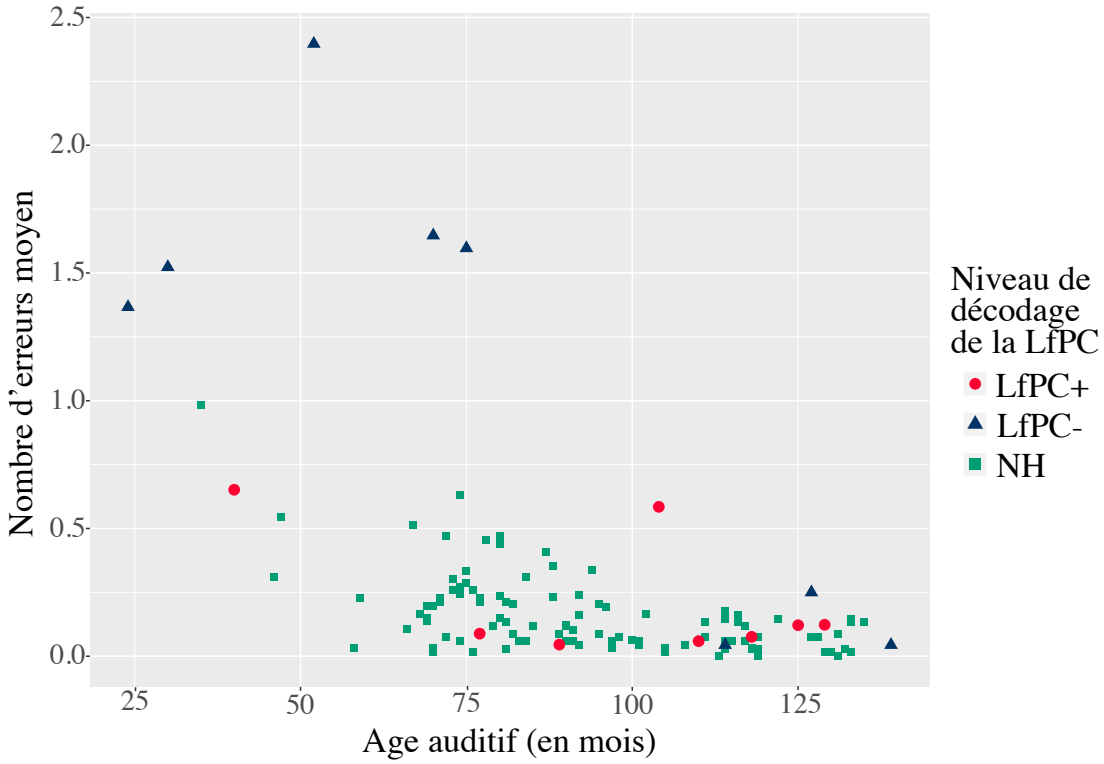


FIGURE 3: Nombre d'erreurs moyen pour les enfants CI et NH en fonction du niveau de décodage de la LfPC et de l'âge auditif

Nous avons réalisé les mêmes analyses en fonction de l'âge auditif. La figure 3 illustre le nombre d'erreurs moyen par mot en fonction du niveau de décodage de la LfPC et de l'âge auditif pour les enfants des groupes CI et NH. Notre sélection par comparaison descendante de modèles nous donne un modèle incluant le groupe (p=.019) et l'âge auditif (p<.001). L'interaction entre le groupe et l'âge auditif n'est pas significative.

Dans ces deux modèles, l'âge d'implantation n'a donc pas d'effet significatif. Les enfants implantés tôt n'ont pas de scores plus élevés que les enfants implantés tard. Ceci peut s'expliquer par la moyenne d'âge d'implantation plutôt basse (avant deux ans) dans notre groupe.

4 Discussion

Cette étude examine l'influence de la LfPC sur la production des phonèmes du français chez 16 enfants porteurs d'implant cochléaire. Le niveau de décodage de la LfPC a été évalué selon deux

niveaux : faible décodage (quelques syllabes sont décodées à vitesse lente) et bon décodage (décodage de mots isolés et de phrases simples à la vitesse de la parole).

Les résultats montrent qu'un bon niveau de décodage de la LfPC est associé à un nombre d'erreurs moyen en production plus faible. Ces résultats peuvent s'expliquer par la qualité des représentations phonologiques dont disposent les enfants profitant d'un bain de langue en LfPC. En effet, la LfPC permet l'émergence de représentations phonologiques stables et précises ce qui pourrait avoir un impact significatif sur la production de phonèmes en français. Par ailleurs, il est important ici de préciser que les enfants avec un bon niveau de décodage de la LfPC ont des scores similaires à ceux des enfants normo-entendants : la LfPC permettrait d'atteindre des productions acoustiques semblables à celles des enfants ne présentant aucun trouble de l'audition.

Cette étude exploratoire nous permet tout d'abord d'insister sur le fait que les enfants porteurs d'implant cochléaire ont des difficultés en production de parole, qui peuvent persister jusqu'à l'âge de huit ans, et qu'ils ont besoin d'une prise en charge adaptée. Elle met ensuite en évidence l'apport de la LfPC dans la production de parole chez l'enfant porteur d'implant cochléaire : une prise en charge s'appuyant sur la LfPC a un bénéfice significatif sur les compétences langagières, ce qui a nécessairement un effet positif sur le développement social et la progression scolaire.

Des travaux supplémentaires sont nécessaires pour compléter nos données et observer l'impact d'autres facteurs tels que le degré de surdité, le niveau de récupération auditive et les effets d'âge sur la production de phonèmes en français. Des analyses plus précises vont être menées pour observer l'influence du niveau de langage ainsi que l'effet de la longueur des mots sur les compétences en production chez l'enfant porteur d'implant cochléaire, bénéficiant ou non de la LfPC. D'autre part, les types de réponse produits (spontané, avec amorçage sémantique ou phonologique ou encore répétition) seront pris en compte. Nous envisageons aussi de quantifier plus précisément la fréquence d'exposition à la LfPC et à la LSF. D'autres analyses sont en cours pour mieux caractériser les types d'erreurs produits par ces enfants, ce qui permettra à terme de fournir des pistes d'interventions orthophoniques ciblées.

Références

ALEGRIA J., CHARLIER B. L. & MATTYS S. (1999). The role of lip-reading and Cued Speech in the processing of phonological information in French-educated deaf children. *European Journal of Cognitive Psychology,* 11, 451-472. DOI: 10.1080/095414499382255.

BERLAND A. (2014). *Le développement psychologique d'enfants sourds porteurs d'un implant cochléaire : étude longitudinale et transversale*. Thèse de doctorat, Université Toulouse 2 – Le Mirail, Toulouse.

BERNSTEIN L. E., TUCKER P. E. & DEMOREST M. E. (2000). Speech perception without hearing percept. *Psychophysiology,* 62, 233-252. DOI: 10.3758/BF03205546.

BRATAKOS M. S., DUCHNOWSKI P. & BRAIDA L. D. (1998). Toward the automatic generation of Cued Speech. *Cued Speech Journal,* 6, 1-37.

CHARLIER B. L. & LEYBAERT J. (2000). The rhyming skills of deaf children educated with phonetically augmented speechreading. *The Quarterly Journal of Experimental Psychology Section A,* 53, 349-375. DOI : 10.1080/713755898.
COLIN C. & RADEAU M. (2003). Les illusions McGurk dans la parole : 25 ans de recherches. *L'année psychologique,* 104, 497-542.
COLIN S., ECALLE J., TRUY E., LINA-GRANADE G. & MAGNAN A. (2017). Effect of age at cochlear implantation and at exposure to Cued Speech on literacy skills in deaf children. *Research in Developmental Disabilities,* 71, 61-69. DOI: 10.1016/j.ridd.2017.09.014.
CORNETT R. O. (1967). Cued Speech. *American Annals of the Deaf,* 112(1), 3-13.
HAGE C. (1994). *Développement de certains aspects de la morpho-syntaxe chez l'enfant à surdité profonde : rôle du Langage Parlé Complété*. Thèse de doctorat, Université Libre de Bruxelles, Belgique.
HAGE C., ALEGRIA J., PERIER O. & MARTIN D. S. (1991). Cued Speech and language acquisition: The case of a grammatical gender morpho-phonology. *Advances in Cognition, Education and Deafness,* 395-399.
HAGE C. & LEYBAERT, J. (2006). The effect of Cued Speech on the development of spoken language. In P. E. SPENCER & M. MARSCHARK, Eds., *Advances in the Spoken Language Development of Deaf and Hard-of-Hearing Children,* p. 193-211. Oxford University Press.
KOS, M.-I., DERIAZ, M., GUYOT, J.-P. & PELIZZONE, M. (2008). What can be expected from a late cochlear implantation? *Int. J. Pediatr. Otorhinolaryngol.* 73, 189–193. DOI: 10.1016/j.ijporl.2008.10.009
LEYBAERT J., COLIN C., HAGE C., & LASASSO C. J. (2010). Cued Speech for enhancing speech perception and first language development of children with cochlear implants. *Trends Amplification,* 14, 96-112. DOI: 10.1177/1084713810375567.
MELONI G., LŒVENBRUCK H., VILAIN A. & MACLEOD A. (2017). EULALIES, The France-Québec Speech Sound Disorders project. Poster présenté, IASCL, Lyon, France.
NITTROUER S., CALDWELL-TARR A., SANSOM E., TWERSKY J. & LOWENSTEIN J. H. (2014). Nonword repetition in children with cochlear implants: a potential clinical marker of poor language acquisition. *American Journal of Speech and Language Pathologies,* 23, 679. DOI: 10.1044/2014_AJSLP-14-0040.
PERIER O., CHARLIER B. L., HAGE C. & ALEGRIA J. (1990). Evaluation of the effects of prolonged Cued Speech practice upon the reception of spoken language. *Cued Speech Journal,* IV, 47-59.
RVACHEW, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech and Hearing Research*, *37*(2), 347–357. DOI: 10.1044/jshr.3702.347
RVACHEW, S., NOWAK, M., & CLOUTIER, G. (2004). Effect of phonemic perception training on the speech production and phonological awareness skills of children with expressive phonological delay. *American Journal of Speech-Language Pathology*, *13*(3), 250–263. DOI: 10.1044/1058-0360%282004/026%29

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Détection de la somnolence par estimation d'erreurs de lecture

Vincent P. Martin[1] Gabrielle Chapouthier[2] Mathilde Rieant[2] Jean-Luc Rouas[1]
Pierre Philip[3]
(1) LaBRI - Univ. Bordeaux - Bordeaux INP - CNRS - UMR5800 - F-33400 Talence, France
(2) CFUOB - Univ. Bordeaux Sengalen - F-33076 Bordeaux, France
(3) SANPSY - CNRS - USR 3413 - Univ. Bordeaux - CHU Pellegrin - F-33000 Bordeaux, France
{vincent.martin, rouas}@labri.fr, {gabrielle.chapouthier,
mathilde.rieant}@etu.u-bordeaux.fr, pierre.philip@u-bordeaux.fr

RÉSUMÉ

La détection automatique de la somnolence peut aider le suivi de patients souffrant de maladies neuro-psychiatriques chroniques. Des recherches précédentes ont déjà montré que cela est possible en utilisant des enregistrements vocaux. Dans cet article, nous proposons d'étudier les erreurs de lecture effectuées par des patients souffrant de Somnolence Diurne Excessive (SDE) sur le corpus TILE, enregistré à l'hôpital de Bordeaux. Avec des orthophonistes, nous avons défini et compté les erreurs de lecture des patients et les avons confrontées aux différentes mesures de somnolence du corpus. Nous montrons ici que relever ces erreurs peut être utile pour élaborer des marqueurs robustes de la somnolence objective mais aussi pour définir des critères d'exclusion des locuteurs n'ayant pas un niveau de lecture suffisant.

ABSTRACT

Sleepiness detection through reading errors estimation

Automatic detection of sleepiness can help to improve the follow-up of patients suffering from chronic diseases. Previous research on sleepiness detection has shown that this task is feasible using voice recordings. In this paper, we propose to study the reading errors made by patients suffering from Excessive Daytime Sleepiness (EDS) on the MSLT Corpus, collected at the Bordeaux hospital. With the help of speech therapists, we defined and counted reading errors and confront these numbers with sleepiness measurements. We show that evaluating these reading errors can be useful to elaborate robust markers of objective sleepiness but also to elaborate exclusion criteria of the speakers not having a sufficient reading level.

MOTS-CLÉS : Erreurs de lecture, Détection de la somnolence, Prosodie.

KEYWORDS: Reading Mistakes, Sleepiness detection, Prosody.

1 Introduction

L'un des défis majeurs actuels du diagnostic et du traitement des maladies chroniques en neuro-psychiatrie est la quantification des symptômes et le suivi des patients, afin d'adapter leur traitement et de détecter précocement les rechutes. Ce suivi est aujourd'hui possible grâce à des dispositifs médicaux connectés (mesurant par exemple le poids, la pression sanguine ou l'activité physique) mais des informations cruciales telles que le niveau de fatigue ou de somnolence sont difficiles à

Volume 1 : Journées d'Études sur la Parole, pages 397–405.
HAL : hal-02798563.

mesurer en milieu écologique. Par ailleurs, ces pathologies nécessitent des entretiens réguliers entre les médecins et les patients, mais ceux-ci ne permettent pas de mesurer les variations des symptômes en réponse au traitement lorsque les patients sont à leur domicile. De plus, le nombre grandissant de patients augmente le temps d'attente entre deux entretiens, ce qui implique des entretiens irréguliers avec les médecins et donc un suivi épisodique et incomplet des patients.

Les avancées récentes dans le domaine du traitement automatique de la voix rendent possible la détection des états du locuteur via l'analyse d'indices vocaux (Cummins *et al.*, 2018). Ces méthodes présentent de nombreux avantages : elles ne sont pas invasives, ne nécessitent ni capteurs spécifiques ni processus de calibration, et peuvent être mises en place dans des environnements variés. Elles permettent ainsi un suivi régulier et non restrictif des patients.

Notre objectif est d'estimer la somnolence objective de patients souffrant de Somnolence Diurne Excessive (SDE). Aucun corpus existant ne nous permettait d'élaborer un système répondant à cet objectif : la majorité des études portant sur la détection de la somnolence dans la voix sont basées soit sur le Sleepy Language Corpus (Martin *et al.*, 2019; Schuller *et al.*, 2011) soit plus récemment sur le SLEEP corpus (Schuller *et al.*, 2019). Or, ces corpus sont enregistrés sur des volontaires sains dont la seule mesure associée à la voix est la somnolence subjective. Afin d'avoir une annotation de la somnolence objective d'échantillons vocaux recueillis sur des patients souffrant de SDE, nous avons élaboré notre propre corpus, le corpus TILE (*MSLT corpus* en anglais) (Martin *et al.*, 2020), enregistré à la Clinique du Sommeil du Centre Hospitalier Universitaire de Bordeaux.

Par ailleurs, à notre connaissance, la majorité des travaux précédents ayant pour objectif de détecter la somnolence dans la voix sont basés sur des marqueurs vocaux concernant la qualité de la voix (énergie, fréquence, ...), généralement extraits avec la toolbox openSMILE (Eyben & Schuller, 2015). Cette étude tire bénéfice du fait que contrairement aux autres corpus, dans le corpus TILE tous les échantillons vocaux sont collectés lors d'une tâche de lecture, permettant d'annoter les erreurs de lecture faites par rapport au texte original. La nouveauté de cette approche réside dans le fait de se servir des erreurs de lecture comme nouveaux marqueurs de la somnolence. En effet, si les marqueurs de qualité vocale permettent d'étudier l'influence de la somnolence sur l'aspect neuro-musculaire de la voix (Krajewski *et al.*, 2009), nous pensons que les erreurs de lecture sont des marqueurs pertinents pour l'étude de l'influence de la somnolence sur les performances cognitives. Nous proposons ainsi une nouvelle méthode pour évaluer la somnolence d'un locuteur à partir de ses erreurs de lecture.

Cet article est organisé comme suit. Dans la Section 2, nous décrivons les cinq types d'erreurs de lecture que nous avons utilisées pour annoter notre corpus. Dans la Section 3, nous présentons brièvement notre corpus et les critères d'exclusion que nous avons mis au point. Les résultats sont présentés et discutés dans la Section 4. La Section 5 introduit une première tentative de classification. Enfin, la conclusion et nos futurs travaux sur le sujet sont présentés dans la Section 6.

2 Définition des erreurs de lecture

En travaillant avec des orthophonistes, nous avons sélectionné cinq types d'erreurs de lecture, qui sont à la fois suffisamment générales pour être représentées dans notre corpus et assez restrictives pour être spécifiques à certains comportements de lecture.

— Les achoppements (Ach) : "hésitation, coupure, dans le rythme de la parole" (Brin *et al.*, 2018).

Ces erreurs mesurent principalement la capacité d'*assemblage* du lecteur. L'*assemblage* est le fait de mettre bout à bout des syllabes pour former un mot : quand un lecteur commence à lire un mot, s'arrête, et reprend sa lecture, le processus d'assemblage a été interrompu, ce qui cause un achoppement. Nous avons choisi de ne pas prendre en compte les hésitations entre les mots (interruption dans le débit de parole), mais seulement les arrêts qui sont observés au milieu d'un mot ou les allongements artificiels de certaines voyelles, qui témoignent d'une hésitation. En effet, les différentes accents rencontrés lors des enregistrements induisent des hésitations plus ou moins appuyées entre les mots, ce qui n'est pas le cas des hésitations dans les mots. Les hésitations et reprises de bouts de phrases ou de phrases sont également prises en compte dans cette catégorie. Dans le cas d'une reprise, un seul achoppement est compté, quel que soit le nombre de mots repris.

— Les paralexies (Plx) : "erreur d'identification de mots écrits consistant à oraliser un mot écrit à la place d'un autre" (Brin *et al.*, 2018).

Contrairement aux achoppement, les paralexies reflètent les capacités d'*adressage* du lecteur. En effet, contrairement au processus d'*assemblage*, l'*adressage* est défini comme le fait de lire un mot dans sa globalité, sans le déchiffrer ou le découper en syllabes. Les paralexies sont des erreurs symptomatiques de ce type de lecture. Nous avons choisi de généraliser cette catégorie à toute prononciation d'un mot, existant ou non, qui est lu à la place du mot correct.

— Les oublis de mots (O) : cette erreur est observée lorsque le locuteur oublie de prononcer un mot et passe directement au suivant ou au début du mot suivant avant de se reprendre. Même s'il se corrige ensuite, l'oubli est compté.
— Les additions (Add) : cette erreur est observée lorsque le locuteur ajoute un mot qui n'est pas présent dans le texte. Même s'il se corrige ensuite, l'addition est prise en compte.
— Les inversions de mots (I) : cette erreur est observée lorsque le locuteur inverse plusieurs mots dans la phrase.

Si un locuteur se reprend après une paralexie, un oubli, une addition ou une inversion de mots, aucun achoppement n'est compté sauf s'il se trompe avec l'erreur correspondante lors de sa reprise.

3 Description de la base de données

3.1 Présentation du corpus

Le corpus utilisé dans cette étude est une version étendue du corpus TILE (Martin *et al.*, 2020). Il comprend les enregistrements de 115 patients enregistrés à la Clinique du Sommeil du Centre Hospitalier Universitaire de Bordeaux (France). Tous les patients ont émis des plaintes concernant des problèmes de sommeil et passent un Test Itératif de Latence d'Endormissement (Littner *et al.*, 2005) - TILE. Ce test consiste en 5 siestes espacées de 2h à partir de 9h du matin. Les patients sont largement phénotypés et leurs caractéristiques physiques ainsi que leurs résultats à des questionnaires subjectifs de dépression, de fatigue et de somnolence sur le long-terme sont collectés. Le principal avantage de ce corpus réside dans le fait qu'il associe à chaque échantillon audio deux mesures de somnolence : une mesure subjective (Karolinska Sleepiness Scale (Åkerstedt & Gillberg, 1990) - KSS) et une mesure objective (temps d'endormissement des patients à chaque sieste, appelé "valeur de TILE"). Cette double annotation est importante car si les sujets sains estiment correctement leur niveau de somnolence (Horne & Burley, 2010), ce n'est pas le cas des malades (Sangal, 1999). Les échantillons vocaux sont collectés durant la lecture d'un texte, qui est différent à chaque session mais le même pour tous les locuteurs à session constante. Afin d'éviter une trop grande valence émotionnelle et

pour avoir une grammaire et un vocabulaire simple, nous avons proposé aux sujets des extraits du Petit Prince d'un peu plus de 200 mots.

3.2 Sélection des sujets (critères d'exclusion)

Donnée	Femmes	Hommes	Total
Nombre de sujets	59	40	99
Nombre d'échantillons	295	200	495
Âge moyen (écart-type)	34,2 (11,6)	39,0 (17,1)	36,1 (14,3)
Niveau Social moyen (écart-type)	4,6 (2,4)	5,9 (2,6)	5,4 (2,6)
TILE moyenne (écart-type)	**11,8 (4,6)**	**10,3 (5,2)**	**11,2 (4,9)**
KSS moyen (écart-type)	4,2 (1,2)	4,6 (1,2)	4,4 (1,2)
Nombre de sujets somnolents - S	12	15	27
Nombre de sujets non somnolents - NS	47	25	72

TABLE 1: Statistiques concises du corpus une fois les patients exclus

En annotant la base de données avec les erreurs décrites précédemment, nous avons pu établir des critères d'exclusion pour cette étude. Exclure des patients de la base de données va certes en réduire la taille, mais cela va en contrepartie assurer que les marqueurs vocaux calculés et les erreurs de lecture mesurées sur les patients inclus seront principalement influencés par la somnolence, excluant les effets des pathologies et des troubles de la lecture sur les variables mesurées.

Tout d'abord, nous avons exclus trois patients présentant des troubles visuo-attentionnels ou d'alexie, séquelles répandues des AVC ou des Accidents Ischémiques Transitoires (AIT). Les patients exclus avaient effectivement de tels antécédents médicaux et avaient rythme de lecture très lent accompagné de très nombreuses erreurs. De même, trois patients produisaient de nombreuses erreurs de lecture suspectées d'être liées à un manque de contrôle musculaire lors de la production vocale, dûes à une maladie neuro-musculaire (comme par exemple une dysphonie, une myotonie, une chorée d'Huntington ou de l'épilepsie). Ces patients produisant un grand nombre d'erreurs ont tous les trois été diagnostiqué de maladies neuro-musculaires.

Trois patients ont été également été exclus après avoir omis ou répété une ligne, ou après avoir omis un trop grand nombre de mots, rendant la lecture incohérente. En effet, ces erreurs d'attention peuvent être dues à la somnolence, mais ce sont aussi des marqueurs des Troubles et Déficits de l'Attention avec ou sans Hyperactivité (TDAH). Différencier l'origine de ces erreurs étant très difficile, nous choisissons d'exclure ces patients, dont un était diagnostiqué TDAH.

Concernant la fluence verbale, un patient présentait des caractéristiques de bredouillement et a été exclus de notre corpus. Les pathologies concernant la fluence verbale sont importantes à prendre compte puisque leurs effets sont difficilement différentiables des effets de la somnolence. Finalement, nous avons également exclus quatre patients qui déchiffraient le texte lors de la lecture à voix haute malgré une précédente lecture dans leur tête quelques minutes avant, symptômes d'une éventuelle dyslexie ou d'un trouble associé.

Un patient supplémentaire soufrant de différentes maladies inflammatoires sérieuses (Maladie de Crohn, de Basedown et Spondylarthrite Ankylosante) et un autre souffrant lui de troubles anxieux sévères impactant sa lecture ont également été exclus.

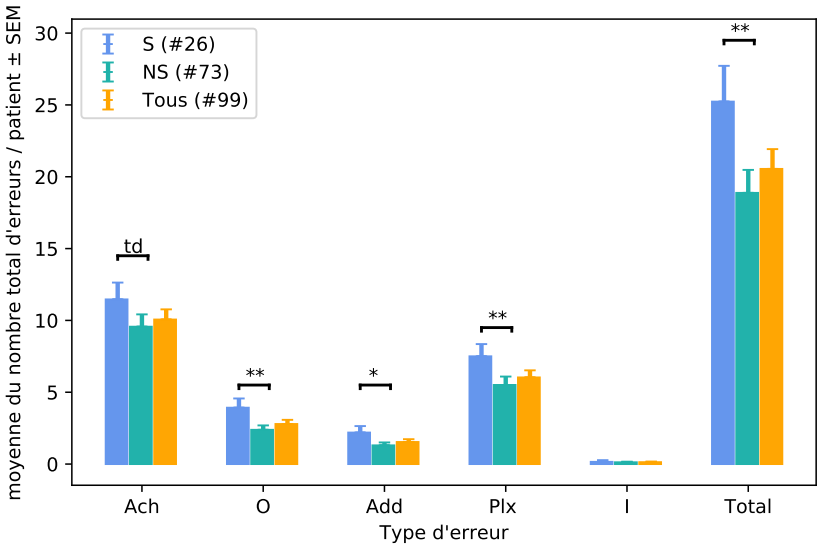


FIGURE 1: Distribution du nombre total d'erreurs par locuteur (moyenne ± SEM). Ach : Achoppements, O : Oublis, Add : Additions, Plx : Paralexies, I : Inversions de mots.
Tests de Mann-Whitney (td : $p < 6 \times 10^{-2}$, * : $p < 5 \times 10^{-2}$, ** : $p < 10^{-2}$).

Nous avons ainsi gardé un total de 99 locuteurs sur les 115 originaux. Des statistiques concises sur le corpus une fois ces patients exclus sont proposés dans le Tableau 1. Les patients sont séparés entre Somnolents (S) et Non-Somnolents (NS) grâce à la limite médicale de 8 minutes sur la moyenne des cinq valeurs de TILE qui est utilisée pour le diagnostic de la narcolepsie (Aldrich *et al.*, 1997).

4 Résultats

4.1 Caractéristiques des locuteurs

Afin de mesurer si les erreurs évaluées précédemment sont sensibles à la somnolence, nous avons tracé la distribution du nombre total de chaque type d'erreur par locuteur dans la Figure 1 (moyenne ± SEM - *Standard Error of the Mean*).

Excepté pour les Inversions de mots, les patients Somnolents produisent plus d'erreurs que leurs homologues Non-Somnolents, et ce quelle que soit la catégorie d'erreur (Test de Mann-Whitney. Ach : $U = 746,0\,; p = 5,3 \times 10^{-3}$. O : $U = 635,0\,; p = 5,7 \times 10^{-3}$. Add : $U = 702,5\,; p = 2,1 \times 10^{-2}$. Plx : $U = 616,5\,; p = 3,9 \times 10^{-3}$. I : $U = 921,5\,; p = 0,34$. Total : $U = 631,0\,; p = 5,8 \times 10^{-3}$). En raison du faible nombre d'Inversions de mots observés, ce type d'erreur n'est pas pris en compte dans la suite de l'étude.

Nous avons ensuite étudié si le nombre total d'erreurs dans chaque catégorie corrèle (ρ de Spearman) avec les différentes données médicales et sociales disponibles dans le corpus. Contrairement à nos intuitions, le niveau social (mesuré comme le nombre d'années d'étude après le Brevet des Collèges) ou encore l'âge ne corrèlent pas avec la production d'erreurs des 99 locuteurs du corpus. Il y a cependant une corrélation entre le nombre total d'additions par locuteur et la valeur moyenne de TILE ($\rho = 0,28\,; p = 5,3 \times 10^{-3}$) : plus les patients sont affectés par une maladie du sommeil

(somnolence objective plus élevée), plus ils font d'additions dans les textes. Ces erreurs corrèlent également avec le score du questionnaire Index de Sévérité de l'Insomnie (Morin *et al.*, 2011) ($\rho = 0,24\,; p = 1,6 \times 10^{-2}$) : plus les patients ont de troubles du sommeil et des insomnies, plus ils produisent d'additions.

Les paralexies corrèlent non seulement avec la somnolence objective, i.e. la valeur moyenne de TILE ($\rho = 0,25\,; p = 1,3 \times 10^{-2}$) mais aussi avec la somnolence subjective, i.e. la valeur moyenne des KSS ($\rho = 0,25\,; p = 1,1 \times 10^{-2}$). Le fait que les paralexies corrèlent positivement avec à la fois les mesures de somnolence objectives et subjectives indique que la production de ce type d'erreurs augmente avec la sévérité de leur somnolence objective mais aussi avec la perception que les locuteurs en ont. Cela a l'avantage de pouvoir détecter les deux types de somnolence grâce à ce type d'erreurs mais a le principal inconvénient de ne pas pouvoir différencier les deux types de somnolence : un patient produisant un grand nombre de paralexies pourra soit avoir une somnolence objective haute ou alors seulement le ressenti qu'elle est haute. En comparaison, le nombre d'additions semble être un biomarqueur du niveau de somnolence des patients plus précis puisqu'il ne corrèle qu'avec la somnolence objective.

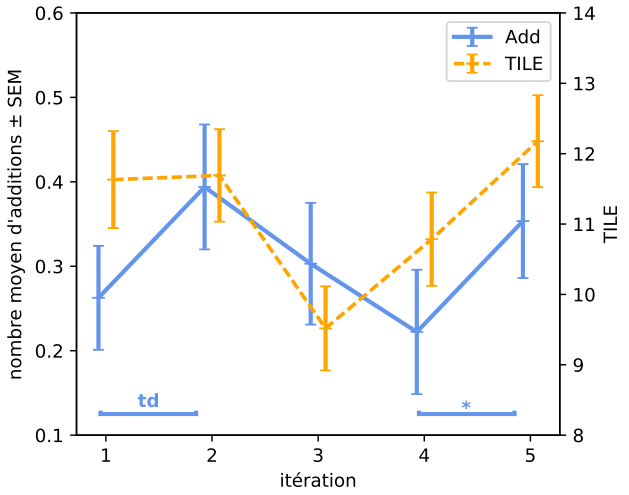
4.2 Étude des sources d'influences de production d'erreurs

Aucun des textes choisis n'a exactement la même taille, la même difficulté ou la même quantité de dialogues. Il est donc nécessaire d'étudier l'influence des textes sur cette production d'erreurs et de la séparer de l'influence de la somnolence. De plus, des variables dépendant du temps peuvent affecter à la fois le niveau de somnolence et la production. Parmi ces variables comptent le fait que les patients petit-déjeunent avant la première itération du test (à 9h du matin), qu'ils déjeunent peu avant la troisième itération (à 13h) ou qu'ils expriment généralement de la fatigue et un ennui vis à vis du test durant la dernière itération (17h). Dans la suite, "influence de l'itération" fera indépendamment référence à l'influence du texte ou des variables précédemment décrites, les deux effets n'étant pas séparables.

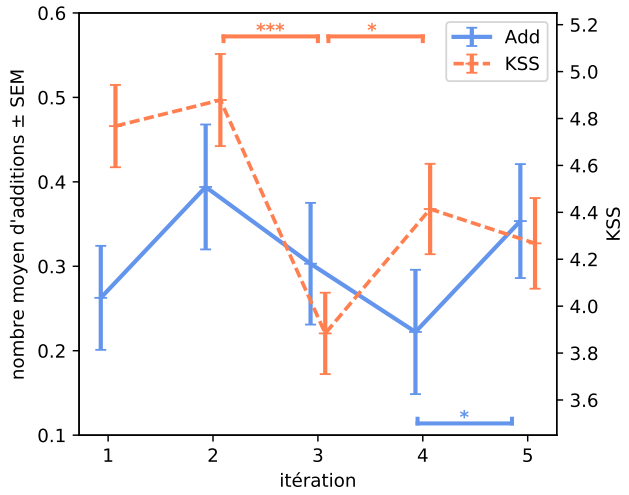
Afin de séparer la contribution de l'itération, de la valeur de TILE et du KSS sur les variations du nombre d'additions et de paralexies, nous avons appliqué une ANOVA multivariée à mesures répétées avec R (R Core Team, 2017).

En étudiant les variations des valeurs de TILE et du nombre d'additions chez les deux classes de patients (Figure 2a), la somnolence objective semble avoir une influence sur la production de ce type d'erreurs. Une ANOVA permet de mettre en évidence une influence significative du TILE sur les variations inter-sujets ($F = 6,0\,; p = 1,6 \times 10^{-2}$) et une influence presque significative du KSS sur les variations intra-sujets ($F = 3,4\,; p = 6,3 \times 10^{-2}$). Cela signifie que les différences observées entre les sujets indépendamment du temps sont principalement expliquées par leurs différences de TILE (ce qui confirme le lien entre TILE et additions) tandis que celles observées entre les sujets au cours du temps (influence conjointe de la session et du locuteur) sont principalement expliquées par les différences de variation de KSS au cours des itérations du test. La session n'a aucun effet significatif sur la production des additions. Nous émettons donc l'hypothèse que les variations du nombre d'additions sont principalement dues aux variations de somnolences objectives et subjectives, et qu'elles sont donc indépendantes du texte et des autres effets temporels.

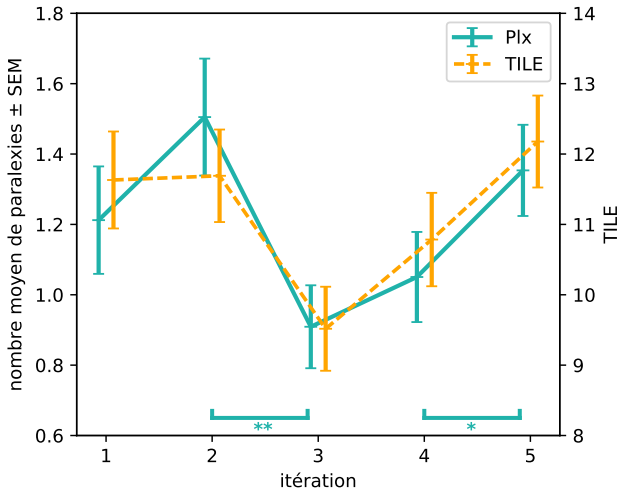
L'influence de la somnolence sur le nombre de paralexies est plus complexe à analyser. Malgré une forte corrélation entre l'évolution des valeurs de TILE et de KSS avec le nombre de paralexies



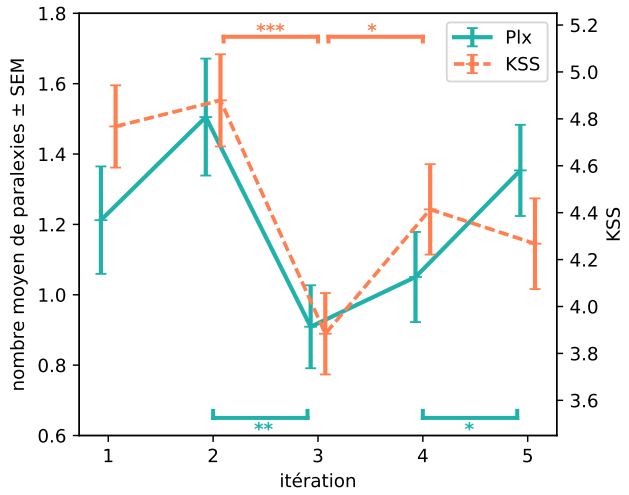
(a) Additions et TILE en fonction des itérations du test



(b) Additions et KSS en fonction des itérations du test



(c) Paralexies et TILE en fonction des itérations du test



(d) Paralexies et KSS en fonction des itérations du test

FIGURE 2: Additions (a,b) et paralexies (c,d) comparées au TILE et au KSS (moyenne±SEM). Tests de Mann-Whitney
(td : $p < 6 \times 10^{-2}$, * : $p < 5 \times 10^{-2}$,** : $p < 10^{-2}$, *** : $p < 10^{-3}$).

observées sur les Figures 2c et 2d, l'effet dominant lors de l'analyse par ANOVA est l'influence de l'itération ($F = 5,2\,; p = 4,2 \times 10^{-4}$). En effet, nous avons observé lors de l'annotation de la base de données que certains mots sont systématiquement mal prononcés, menant à une paralexie (par exemple le mot "méditatif" est très souvent prononcé "médiatif"). L'influence du texte induit ici une influence de l'itération sur la production des paralexies. Les corrélations entre le nombre total de paralexies et les deux mesures différentes de la somnolence restent malgré tout des indications optimistes concernant l'utilité des paralexies dans la détection de la somnolence.

Une ANOVA permet d'expliquer les variations intra-sujets des achoppements par le KSS ($F = 5,8\,;$ $p = 1,7 \times 10^{-2}$) et l'itération ($F = 5,9\,; p = 1,2 \times 10^{-4}$), ce qui exclut ce type d'erreur des biomarqueurs robustes pour estimer la valeur de TILE. Nous avons choisi d'ignorer les arrêts entre les mots, puisqu'ils sont difficilement différentiables des accents des locuteurs ou de respirations inhabituelles, ce qui pourrait être la cause de l'absence de corrélation entre le nombre total d'achoppements et la

somnolence objective des patients.

Concernant les oublis, nous avons remarqué que les mots oubliés sont souvent les mêmes. Ce sont des petits mots de liaison qui sont en général omis à l'oral (par exemple "Il me répéta alors" à la place de "Et il me répéta alors"). Une ANOVA confirme que les variations du nombre d'oublis dépendent fortement du texte (influence de l'itération : $F = 12,6$; $p = 1,2 \times 10^{-9}$), les empêchant d'être des biomarqueurs robustes de l'état de somnolence du locuteur. Une autre version de ces erreurs pourrait toutefois avoir du potentiel en tant que biomarqueur de la somnolence puisque les variations inter-locuteurs sont expliquées par la valeur de TILE ($F = 4,1$; $p = 0,05$). Ces observations soulèvent la nécessité d'une étude en profondeur du contenu des textes, pour éviter la reproduction de ces problèmes mais également pour s'assurer qu'ils sont visuellement équivalents (les dialogues présents dans notre corpus de textes semblent impliquer des erreurs visuo-attentionnelles) et que la difficulté des textes n'est pas la source des erreurs faites par les sujets.

5 Élaboration d'un classificateur

Une première approche pour élaborer un classificateur entre sujets Somnolents et Non-Somnolents consiste à concaténer les erreurs des cinq itérations du test et d'utiliser ce vecteur comme entrée d'une Machine à Vecteur Support - SVM (noyau linéaire, $C = 1 \times 10^{-2}$). Comme la taille de notre corpus est relativement faible (99 locuteurs), nous utilisons de la validation croisée *Leave One Speaker Out Cross Validation (LOSOCV)* : chaque locuteur est tour à tour isolé pour servir de test, tandis que les autres forment la base d'entraînement. Le résultat de la classe estimée pour le locuteur de test est ajouté dans une matrice de confusion globale, qui sert d'évaluation moyenne de l'ensemble des systèmes. Après avoir normalisé l'ensemble d'entraînement, entraînés les paramètres du SVM et évalué le système obtenu pour chaque itération de la LOSOCV, nous obtenons une performance (*Unweighted Accuracy Recall - UAR*) calculée sur la matrice de confusion globale de 61,0% (Sensibilité : 57,7%, Spécificité : 64,4%). Même si les performances obtenues par ce système sont en dessous de celles obtenues dans l'état de l'art, il sera intéressant d'étudier leur combinaison avec des systèmes classiques en utilisant différents ensembles de biomarqueurs.

6 Conclusions et Perspectives

En conclusion, nous avons proposé de nouveaux marqueurs pour la détection de la somnolence à partir des erreurs de lecture. La mesure de ces erreurs permet à la fois de proposer des critères d'exclusion pour notre corpus (basés sur leur niveau de lecture ou sur leurs pathologies) et de mesurer la somnolence des locuteurs grâce à leur voix. Par ailleurs, nous avons aussi montré que ces nouveaux marqueurs peuvent être utilisés dans des tâches de classification d'état de somnolence de locuteur. Cela pourrait aider à améliorer les systèmes classiques de détection de la somnolence dans la voix, actuellement basés sur des marqueurs acoustiques, en les utilisant avec différents types de marqueurs.

Nos futurs travaux comprennent l'élaboration d'un système de détection automatique des erreurs présentées dans cet article grâce à des systèmes de transcription automatique. D'autre part, nous projetons de mettre au point des marqueurs équivalents pour la parole spontanée et d'étudier les possibilités de fusion des marqueurs des systèmes classiques avec les nôtres.

Références

ALDRICH M. S., CHERVIN R. D. & MALOW B. A. (1997). Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy. *Sleep*, **20**(8), 620–629.

BRIN F., COURRIER C., LEDERLE E. & MASY V. (2018). *Dictionnaire d'orthophonie - 4ème édition*. Orthoedition édition.

CUMMINS N., BAIRD A. & SCHULLER B. (2018). Speech analysis for health : Current state-of-the-art and the increasing impact of deep learning. *Health Informatics and Translational Data Analytics*, **151**, 1–54.

EYBEN F. & SCHULLER B. (2015). Opensmile. *ACM SIGMultimedia Records*, **6**, 4–13.

HORNE J. & BURLEY C. (2010). We know when we are sleepy : Subjective versus objective measurements of moderate sleepiness in healthy adults. *Biological Psychology*, **83**(3), 266–268.

KRAJEWSKI J., BATLINER A. & GOLZ M. (2009). Acoustic sleepiness detection : Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods*, **41**(3), 795–804.

LITTNER M. R., KUSHIDA C., WISE M., DAVILA D. G., MORGENTHALER T., LEE-CHIONG T., HIRSHKOWITZ M., LOUBE D. L., BAILEY D., BERRY R. B., KAPEN S. & KRAMER M. (2005). Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test. *Sleep*, **28**(1), 113–121.

MARTIN V. P., ROUAS J.-L., MICOULAUD-FRANCHI J.-A. & PHILIP P. (2020). The Objective and Subjective Sleepiness Voice Corpora. In *12th Language Resources and Evaluation Conference*.

MARTIN V. P., ROUAS J.-L., THIVEL P. & KRAJEWSKI J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue*. DOI : 10.1109/SPED.2019.8906577.

MORIN C. M., BELLEVILLE G., BÉLANGER L. & IVERS H. (2011). The Insomnia Severity Index : Psychometric Indicators to Detect Insomnia Cases and Evaluate Treatment Response. *Sleep*, **34**(5), 601–608.

R CORE TEAM (2017). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing.

SANGAL R. (1999). Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy. *Clinical Neurophysiology*, **110**(12), 2131–2135.

SCHULLER B., BATLINER A., BERGLER C., POKORNY F. B., KRAJEWSKI J., CYCHOCZ M., VOLLMAN R., ROELEN S.-D., SCHNIEDER S., BERGELSON E., CRISTIA A., SEIDL A., WARLAUMONT A., YANKOWITZ L., NÖTH E., AMIRIPARIAN S., HANTKE S. & SCHMITT M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech 2019*.

SCHULLER B., STEIDL S., BATLINER A., SCHIEL F. & KRAJEWSKI J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *Interspeech 2011*, p. 3201–3204.

ÅKERSTEDT T. & GILLBERG M. (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci*, **52**, 29–37.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Détection de la somnolence objective dans la voix

Vincent P. Martin[1] Jean-Luc Rouas[1] Pierre Philip[2]
(1) LaBRI - Univ. Bordeaux - Bordeaux INP - CNRS - UMR5800 - F-33400 Talence, France
(2) SANPSY - CNRS - USR 3413 - Univ. Bordeaux - CHU Pellegrin, F-33000 Bordeaux, France
vincent.martin@labri.fr, rouas@labri.fr, pierre.philip@u-bordeaux.fr

RÉSUMÉ

Le suivi des patients souffrant de maladies neuro-psychiatriques chroniques peut être amélioré grâce à la détection de la somnolence dans la voix. Cet article s'inspire des systèmes état-de-l'art en détection de la somnolence dans la voix pour le cas particulier de patients atteints de Somnolence Diurne Excessive (SDE). Pour cela, nous basons notre étude sur un nouveau corpus, le corpus TILE. Il diffère des autres corpora existants par le fait que les sujets enregistrés sont des patients souffrant de SDE et que leur niveau de somnolence est mesuré de manière subjective mais aussi objective. Le système proposé permet détecter la somnolence objective grâce à des paramètres vocaux simples et explicables à des non spécialistes.

ABSTRACT

Objective sleepiness detection through voice

The following-up of patients suffering from chronical neuro-psychiatric diseases can be improved by sleepiness detection through voice. This article draws from state-of-the-art systems to estimate sleepiness level from voice, for the specific case of patients suffering from Excessive Daytime Sleepiness (EDS). To this end, we base our study on a new corpus, the MSLT corpus. It differs from other existing corpora by the fact that recorded subjects suffer from EDS and that their sleepiness level is measured by both subjective and objective means. The proposed system allows to detect objective sleepiness with simple vocal markers that are explainable to non-specialists.

MOTS-CLÉS : Détection de la somnolence, Paramètres vocaux, Prosodie, Lecture de textes.

KEYWORDS: Sleepiness detection, Vocal markers, Prosody, Read speech.

1 Introduction

L'un des défis majeurs actuels du diagnostic et du traitement des maladies neuro-psychiatriques chroniques est la quantification des symptômes et le suivi des patients afin d'adapter leur traitement et de détecter précocement les risques de rechute. Une telle surveillance est possible en milieu écologique grâce à des dispositifs médicaux connectés (mesurant par exemple le poids, la pression sanguine ou l'activité physique) mais des informations pourtant cruciales telles que la fatigue ou la somnolence sont difficiles à mesurer par ces dispositifs. Des entretiens fréquents entre les médecins et les patients sont nécessaires, mais la quantité grandissante de patients ne permet pas un suivi régulier et personnalisé. Par ailleurs, les entretiens ne permettent pas de mesurer les variations des symptômes en réponse au traitement lorsque les patients sont à domicile. Ainsi est née l'idée de proposer aux patients un suivi à domicile en utilisant un médecin virtuel. Des études précédentes ont

Volume 1 : Journées d'Études sur la Parole, pages 406–414.
HAL : hal-02798565.

montré que l'utilisation d'un tel médecin virtuel est bien accepté par les patients (Philip *et al.*, 2020, 2017). Nous désirons compléter l'analyse des réponses aux questions posées par le médecin virtuel en y ajoutant l'analyse de paramètres vocaux. En effet, il semble désormais possible de détecter des indices dans la voix permettant d'évaluer l'état des locuteurs pour des tâches de suivi ou de diagnostic médical (Cummins *et al.*, 2018). Cette méthode présente de nombreux avantages, puisqu'elle n'est pas invasive et ne nécessite ni de capteurs spécifiques ni de processus complexe de calibration. De plus, elle peut être mise en place dans des environnements variés et permet un suivi régulier et non restrictif des patients.

Si des études précédentes ont montré qu'il est possible d'estimer la somnolence subjective dans la voix (Martin *et al.*, 2019; Schuller *et al.*, 2011; Cummins *et al.*, 2018), la plupart étaient basées sur le *Sleepy Large Corpus* (Schuller *et al.*, 2011) qui ne contient que des enregistrements de sujets sains. Plus récemment, le corpus *SLEEP* (Schuller *et al.*, 2019), élaboré pour le challenge Interspeech 2019, ouvre la voie à l'utilisation de l'apprentissage profond pour la détection de la somnolence dans la voix grâce à sa taille importante (16462 échantillons). Cependant, il est lui aussi composé d'enregistrements de sujets sains dont la somnolence est évaluée de manière subjective.

Puisque notre objectif est d'estimer la somnolence objective de patients souffrant de Somnolence Diurne Excessive (SDE), les précédents corpus ne correspondent pas à nos besoins. En effet, la somnolence est uniquement mesurée par le questionnaire médical subjectif *Karolinska Sleepiness Scale* (Åkerstedt & Gillberg, 1990) - KSS) ; les enregistrements sont effectués sur des tâches extrêmement variées allant de la production de voyelles tenues à la lecture de textes divers, en Allemand et Anglais, rendant les échantillons difficilement comparables ; les sujets enregistrés sont des sujets sains, alors que nous souhaitons suivre à domicile des patients souffrant de SDE, pour lesquels les somnolences subjectives et objectives ne corrèlent pas (Sangal, 1999). Pour résoudre ce problème nous avons donc enregistré notre propre corpus, le corpus TILE (*MSLT corpus* en anglais) (Martin *et al.*, 2020b), enregistré à la Clinique du Sommeil du Centre Hospitalier Universitaire de Bordeaux.

Nous présentons ici un système de détection de la somnolence objective de patients souffrant de SDE, grâce à des traits spécifiques de leur voix. La collaboration avec des médecins exige que les paramètres vocaux extraits des échantillons audios soient interprétables et reliés à un phénomène physiologique. En conséquence, plutôt que de mettre au point un système complexe, nous cherchons à étudier si de bonnes performances de classification peuvent être obtenues avec des paramètres vocaux simples et facilement interprétables.

Cet article est organisé comme suit. La Section 2 présente brièvement le corpus utilisé. Nous présentons dans la Section 3 les paramètres vocaux élaborés pour cette étude et dans la Section 4 notre méthodologie de classification. La Section 5 présente et discute les résultats. Enfin, une conclusion et des perspectives sont proposées dans la Section 6.

2 Description du corpus

Le corpus utilisé dans cette étude est une version augmentée du corpus TILE (Martin *et al.*, 2020b). Enregistré à la Clinique du Sommeil au Centre Hospitalier Universitaire de Bordeaux, il comprend actuellement les enregistrements de 99 patients ayant des plaintes de Somnolence Diurne Excessive (SDE). Il est basé sur le Test Itératif de Latence d'Endormissement - TILE (Littner *et al.*, 2005), durant lequel les patients font cinq siestes espacées de deux heures à partir de neuf heures du matin.

Les patients sont largement phénotypés et différentes caractéristiques physiques sont collectées pour chaque patient. Ces données sont complétées par les réponses à des questionnaires médicaux subjectifs de dépression, de fatigue et de somnolence à long terme.

Le principal avantage de ce corpus réside dans le fait qu'il associe à chaque échantillon audio à la fois une valeur de somnolence subjective (Échelle de Somnolence de Karolinska (Åkerstedt & Gillberg, 1990) - KSS) et une valeur objective mesurée par EEG (temps d'endormissement à chaque sieste, appelée "valeur de TILE" dans la suite). Les patients sont assignés à la classe Somnolent (S) ou Non-Somnolent (NS) suivant si la valeur de la moyenne des TILE sur les 5 itérations est inférieure ou supérieure à 8 minutes, limite médicale pour le diagnostic de la narcolepsie (Aldrich *et al.*, 1997) (TILE moyen $\pm$ écart-type : SL : 4,8 min $\pm$ 2,0 min ; NSL : 13,6 min $\pm$ 3,3 min). La même limite de 8 minutes est utilisée pour étiqueter les échantillons, indépendamment du label du locuteur.

Les enregistrements vocaux sont collectés durant la lecture d'un texte, qui est différent à chaque itération du test mais qui est le même pour tous les patients à session constante. Afin d'éviter une trop grande valence émotionnelle et pour avoir une grammaire et un vocabulaire simple, nous avons proposé aux sujets cinq textes, extraits du Petit Prince, d'environ 230 mots (moyenne : 229 mots, écart-type : 16,4 mots). Des statistiques concises du corpus sont présentées dans le Tableau 1. Nous redirigeons le lecteur vers (Martin *et al.*, 2020b) pour plus d'informations sur ce corpus.

Donnée	**Femmes**	**Hommes**	**Total**
Nombre de sujets	59	40	99
Nombre d'échantillons	295	200	495
Âge moyen (écart-type)	34,2 (11,6)	39,0 (17,1)	36,1 (14,3)
Niveau Social moyen (écart-type)	4.6 (2,4)	5,9 (2,6)	5,4 (2,6)
TILE moyenne (écart-type)	**11,8 (4,6)**	**10,3 (5,2)**	**11,2 (4,9)**
KSS moyen (écart-type)	4,2 (1,2)	4,6 (1,2)	4,4 (1,2)
Nombre de sujets somnolents - S	12	15	27
Nombre de sujets non somnolents - NS	47	25	72

TABLE 1 – Statistiques concises du corpus TILE (Martin *et al.*, 2020b)

3 Description des paramètres vocaux

Notre étroite collaboration avec des médecins impose que les paramètres vocaux extraits des échantillons audio soient facilement interprétables et reliés à des phénomènes physiologiques. Dans ce but, nous extrayons des paramètres avec deux granularités temporelles. D'une part, des paramètres provenant des échantillons en entier, en utilisant soit la détection automatique de segments vocaliques (Pellegrino & Andre-Obrecht, 2000) ou la détection automatique de segments voisés grâce à l'extraction de fréquence fondamentale (Sjölander, 2004). D'autre part, nous calculons des paramètres sur chaque segment voisé pour caractériser la régularité de la production d'harmoniques. Ces paramètres sont ensuite moyennés sur chaque échantillon.

3.1 Paramètres calculés sur la totalité de l'échantillon

Les statistiques sur la durée et la proportions de segments voisés et des voyelles reflètent le comportement global de la voix du locuteur.

Les paramètres extraits en utilisant ce paradigme sont :
- *duvoiced* : la durée totale des parties voisées (en s.)
- *pervoiced* : le pourcentage en durée des parties voisées
- *durvowel* : la durée totale des segments vocaliques (en s.)
- *pervowel* : le pourcentage en durée des segments vocaliques

Nous obtenons ainsi 4 paramètres vocaux sur les statistiques des segments voisés et vocaliques calculés sur l'ensemble de l'échantillon.

3.2 Paramètres calculés sur les parties voisées

Les paramètres vocaux extraits sur les parties voisées comprennent des mesures de fréquence fondamentale et de courbes d'intensité :
- F0MEAN : la moyenne de la fréquence fondamentale sur les segments voisés
- F0VAR : la variance de la fréquence fondamentale sur les segments voisés
- F0SLOPE : le coefficient directeur de l'approximation linéaire de la fréquence fondamentale sur un segment voisé
- F0MAX : le maximum de la fréquence fondamentale sur un segment voisé
- F0MIN : le minimum de la fréquence fondamentale sur un segment voisé
- F0EXTEND : l'amplitude de la fréquence fondamentale sur un segment voisé

Les mêmes paramètres sont calculés sur les courbes d'intensité (NRJMEAN, NRJVAR, NRJMAX, NRJMIN, NRJEXTEND). Il en résulte 12 paramètres vocaux supplémentaires (6 sur F0, 6 sur l'intensité). Nous avons également calculé F0MEAN, F0VAR, NRJMEAN et NRJVAR sur les segments vocaliques, ajoutant ainsi 4 paramètres vocaux au groupe de paramètres.

Cet ensemble de paramètres est complété par des paramètres calculés avec le toolkit Matlab Covarep (Degottex *et al.*, 2014) que nous avons modifié pour les calculer seulement sur les segments voisés. Ces paramètres vocaux ont précédemment été utilisés pour caractériser les styles de chant (Rouas & Ioannidis, 2016) ou encore pour la classification d'attitudes sociales (Rouas *et al.*, 2019). Nous complétons ainsi notre ensemble de paramètres avec l'amplitude des harmoniques (H1,H2,H4), l'amplitude des formants (A1,A2,A3), leurs fréquences (F1,F2,F3,F4) et leurs bandes-passantes (B1,B2,B3,B4) ; la différence entre les amplitudes des harmoniques (H1-H2, H2-H4), la différence d'amplitude entre les harmoniques et les formants (H1-A1, H1-A2, H1-A3) ; la *Cepstral Peak Prominence* (CPP) ; les rapports harmoniques sur bruit dans différentes plages de fréquences (HNR05, HNR15, HNR25, HNR35). Tous ces paramètres sont moyennés sur chaque enregistrement, ce qui ajoute un total de 24 paramètres à notre ensemble de paramètres vocaux. Nous avons donc extrait un total de 44 paramètres.

4 Description de la méthodologie de classification

Contrairement aux précédents travaux sur la détection de la somnolence dans la voix (Martin *et al.*, 2019; Schuller *et al.*, 2019; Cummins *et al.*, 2018), le but de celui-ci n'est pas d'estimer la somnolence

instantanée du locuteur mais une somnolence à plus long terme.

Le TILE est un test reconnu médicalement pour le diagnostic de la narcolepsie (Littner *et al.*, 2005; Aldrich *et al.*, 1997). Lorsque la moyenne des valeurs de TILE des 5 siestes est inférieure à 8 minutes, le sujet est diagnostiqué comme narcoleptique. Même si la majorité de nos patients souffrent de maladies différentes de la narcolepsie (principalement d'hypersomnie idiopathique), nous choisissons de conserver la valeur limite de 8 minutes pour distinguer locuteurs Somnolents (S) et Non-Somnolents (NS), et ainsi obtenir une vérité terrain.

Pour estimer cette classe somnolence du locuteur, nous attribuons une classe (S ou NS) à chacun de ses cinq échantillons vocaux, indépendamment les uns des autres, avec la même limite de 8 minutes utilisée pour le label des locuteurs. Nous entraînons ensuite un classificateur à calculer les probabilités d'appartenance à la classe S et à la classe NS des échantillons (notés resp. p_i et $\overline{p_i}$ pour le $i$ème échantillon de chaque locuteur). En moyennant les probabilités des cinq échantillons d'un même locuteur, nous obtenons ainsi sa probabilité moyenne d'appartenir à chacune des classes p_{moy} (et $\overline{p_{moy}}$ pour les NS). En prenant ensuite le maximum entre p_{moy} et $\overline{p_{moy}}$, la classe du locuteur est déterminée. Cette procédure est résumée dans la Figure 1.

La méthodologie pour calculer les probabilités d'appartenance aux classes de somnolence à partir des paramètres vocaux est similaire à celle exposée dans (Martin *et al.*, 2019).

Afin d'avoir des résultats pertinents au regard de la faible taille du corpus, nous appliquons la méthode du *Leave One Speaker Out Cross Validation* : à chaque itération de la validation croisée, un locuteur est exclu pour servir de test. Le résultat de sa classification est additionné dans une matrice de confusion globale, sur laquelle est calculée le taux de précision (*Unweighted Accuracy Recall - UAR*). Le reste des locuteurs est divisé en deux sous-corpus d'entraînements (80%) et de développement (20%), équilibrés en valeurs de TILE moyenne, en sexe et en âge, afin de pouvoir déterminer les paramètres vocaux et les paramètres du classificateur les plus pertinents. Par ailleurs, le jeu de données étant déséquilibré, la base d'entraînement est augmentée grâce à la méthode *SMOTE* implémenté dans le module Python *SciKit Learn* (Pedregosa *et al.*, 2011).

Lors de chaque itération de la validation croisée, la même procédure est appliquée :

1. Calcul pour chaque marqueur vocal de la corrélation (ρ de Spearman) entre le marqueur vocal et la valeur de l'itération de TILE sur l'ensemble *entraînement* + *développement*. Cela permet d'ordonner les paramètres vocaux du plus corrélé au moins corrélé avec la somnolence objective.

2. Sélection du nombre de features. Pour cela, nous calculons les performances du système (en *entraînement* vs *développement*) pour les 1, 2, ..., 44 paramètres vocaux, et gardons le nombre de features et les paramètres de classificateur fournissant les meilleures performances. Le classificateur utilisé est une Machine à Vecteurs Supports - SVM, dont les paramètres sont le noyau (linéaire ou gaussien) et les paramètres C et γ. Durant cette phase, les performances sont mesurées grâce au score F1 (moyenne géométrique de la précision et du rappel).

3. Le système précédent est entraîné sur le sous-corpus *entraînement+développement* et nous obtenons les probabilités d'appartenance aux classes de somnolence des 5 siestes du locuteur de *test*.

4. Les probabilités sont moyennées et seuillées pour obtenir la classe d'appartenance du locuteur. La matrice de confusion globale moyenne du système est mise à jour et nous poursuivons la validation croisée.

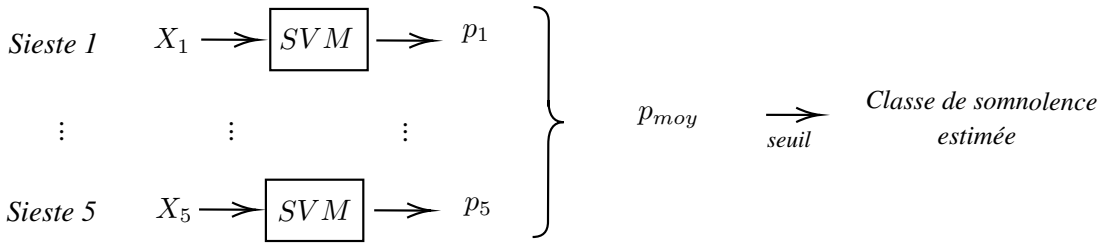


FIGURE 1 – Procédure pour l'estimation de la classe de somnolence. X_i : ensemble des paramètres vocaux triés pour la sieste i. p_i : probabilité que l'échantillon de la sieste i provienne d'un locuteur somnolent. Les classificateurs SVM représentés sont tous identiques et ont été entraînés sur tous les échantillons, toutes itérations confondues

5 Résultats et Discussions

5.1 Résultats du système

La précédente méthodologie conduit à la matrice de confusion présentée dans le Tableau 2a). Malgré les bonnes performance de cette méthode dans notre étude précédente sur la détection de la somnolence subjective (Martin *et al.*, 2019), le score UAR ainsi obtenu est seulement de 51,9%.

Pour améliorer ce système, nous décidons de modifier la procédure de sélection des paramètres vocaux : au lieu de les classer selon leur corrélation avec la valeur de TILE, nous choisissons de les classer selon leur pouvoir discriminant entre les deux classes S et NS, grâce à un test de Mann-Whitney. Plus le U de Mann-Whitney est faible, plus le marqueur vocal a des distributions différentes entre S et NS et donc un pouvoir discriminant élevé. En conservant à l'identique le reste du système, la matrice de confusion obtenue est représentée dans le Tableau 2b). Le score UAR correspondant est de 56,5%, ce qui représente une amélioration de 4,6% par rapport au système a) : cette stratégie de sélection de paramètres semble plus efficace que la précédente.

5.2 Fusion a posteriori des systèmes

Afin de tirer parti des deux différentes approches de hiérarchisation des paramètres (corrélation ou pouvoir discriminant), nous proposons une troisième stratégie. Pour chaque locuteur, nous déterminons sa probabilité d'être somnolent selon les deux systèmes, notés p^c_{moy} et p^m_{moy} (et leur équivalent pour la classe NS, $\overline{p^c_{moy}}$ et $\overline{p^m_{moy}}$). Nous effectuons la fusion des deux systèmes en prenant la classe, indépendamment du système, qui a la plus forte probabilité : $p_{moy} = max(p^c_{moy}, p^m_{moy})$ et $\overline{p_{moy}} = max(\overline{p^c_{moy}}, \overline{p^m_{moy}})$. Nous obtenons ainsi la matrice de confusion présentée dans le Tableau 2 c), conduisant à un score UAR de 60,0%, i.e. une amélioration de 3,5% par rapport au système b) et de 8,1% par rapport au système a).

5.3 Paramètres vocaux pertinents

Le meilleur système étant obtenu par la fusion de deux systèmes utilisant deux sélections de paramètres vocaux différents, il convient d'étudier quels sont les paramètres vocaux les plus pertinents pour chaque approche. Pour cela, nous séparons deux cas selon le paradigme de sélection de paramètres

a) *Corr*	S_{pred}	NS_{pred}
S_{th}	10	17
NS_{th}	24	48
UAR : 51,9%		

b) *Mann*	S_{pred}	NS_{pred}
S_{th}	11	16
NS_{th}	20	52
UAR : 56,5%		

c) *Fusion*	S_{pred}	NS_{pred}
S_{th}	11	16
NS_{th}	15	57
UAR : 60,0%		

TABLE 2 – Matrices de confusion et performances des trois systèmes étudiés : a) Tri des features par corrélation avec la valeur de somnolence (ρ de Spearman) ; b) Tri des features par leur pouvoir discriminant entre les deux classes (U de Mann-Whitney) ; c) Fusion des deux systèmes

vocaux choisi pour chaque locuteur lors de la fusion : corrélation de Spearman ou test de Mann-Whitney.

Lorsque la plus grande probabilité d'appartenance à une classe est observée pour un système de classification utilisant une corrélation de Spearman pour la sélection des features (ce qui est le cas pour 39 sujets), nous faisons la moyenne des corrélations des features avec les valeurs de TILE pour les 39 sujets concernés. Nous obtenons ainsi une corrélation moyenne des features avec les TILE pour les systèmes qui utilisent la corrélation de Spearman et qui sont efficaces dans la fusion. Les cinq paramètres les plus pertinents dans ce paradigme sont principalement des statistiques sur les parties voisées et vocaliques, ainsi que des paramètres concernant les harmoniques et les formants : CPP ($\rho = 6,7 \times 10^{-2}$), durvoiced ($\rho = 6,7 \times 10^{-2}$), durvowel ($\rho = 6,7 \times 10^{-2}$), B1 ($\rho = -6,4 \times 10^{-2}$), H1-H2 ($\rho = -6,1 \times 10^{-2}$).

Nous procédons de même avec les U de Mann-Whitney sur les 60 autres systèmes utilisant la sélection de features grâce à un test de Mann-Whitney. Les cinq paramètres vocaux les plus discriminants sont également des paramètres liés aux harmoniques et formants : H1-A1 ($U = 25981, p = 2,9 \times 10^{-2}$), H1-A2 ($U = 26186, p = 3,9 \times 10^{-2}$), CPP ($U = 26270, p = 4,3 \times 10^{-2}$), H1-H2 ($U = 26489, p = 5,9 \times 10^{-2}$), H1-A3 ($U = 26709, p = 6,1 \times 10^{-2}$).

Un point d'intérêt dans le fait de fusionner deux systèmes est l'origine des bonnes (ou mauvaises) performances de chacun des systèmes. Il est possible par exemple d'étudier l'efficacité de chaque type de système en comparant les distributions des valeurs de TILE moyenne pour les deux systèmes (les 39 sujets précédemment cités pour Spearman et 60 pour Mann-Whitney). Nous obtenons ainsi un TILE moyen pour les systèmes utilisant la discrimination par test de Mann-Whitney (moyenne : 10,3 minutes, écart-type : 5,25) significativement inférieur (test de Mann-Whitney : $U = 884,7, p = 2,1 \times 10^{-2}$) que pour les systèmes utilisant la corrélation de Spearman (moyenne : 12,4, écart-type : 4,0). Les systèmes utilisant la corrélation de Spearman sont donc plus performants sur la détection de la non-somnolence (6 Somnolents, 33 Non-Somnolents) tandis que ceux utilisant le test de Mann-Whitney comme processus de sélection de paramètres vocaux sont moins spécialisés (21 Somnolents, 39 Non-Somnolents). L'idée de faire la fusion de plusieurs systèmes est donc bénéfique dans le cadre de cette étude puisqu'elle permet de cumuler les avantages de chacun des classificateurs.

5.4 Discussion

Ce résultat est optimiste au regard de la faible taille de la base de données. Une étude plus fine de la matrice de confusion nous apporte des informations complémentaires. En effet, quel que soit le système, non seulement la majorité des sujets somnolents sont classés dans la mauvaise catégorie (60% dans le cas du meilleur système) mais la majorité des locuteurs classifiés comme somnolents

sont des patients qui ne le sont pas.

Il y a en effet un déséquilibre entre la classe NS et la classe S qui est très minoritaire. Malgré l'augmentation de données dans la classe minoritaire grâce au SMOTE, les améliorations successives du système proposé précédemment conduit à l'amélioration de la détection des patients non-somnolents plutôt que des sujets somnolents, comme en témoigne l'augmentation des performances sur cette classe là (67% pour le système (a), 72% pour le système (b) et 79,2% pour le système (c)). Un plus grand nombre de patients dans la classe S permettrait au système de mieux généraliser les caractéristiques de la voix somnolente et ainsi de surmonter ce problème.

6 Conclusion et perspectives

Pour conclure, nous avons proposé un système qui est prometteur dans la classification de la somnolence objective grâce à la voix chez des patients souffrant de Somnolence Diurne Excessive. Il s'appuie sur des paramètres vocaux qui sont explicables à des personnes qui ne sont pas spécialistes du traitement du signal vocal, et permet ainsi une collaboration étroite avec le monde médical.

Nos futurs travaux comprendront l'élaboration d'un meilleur classificateur basé sur d'autres techniques du traitement du signal, comme par exemple des techniques à base de Réseau de Neurones Récurrents pour prendre en compte les variations temporelles des paramètres vocaux. Par ailleurs, la fusion avec d'autres paradigmes de classification comme l'étude des erreurs de lecture (Martin *et al.*, 2020a) pourrait permettre de rendre le système plus robuste et d'en améliorer les performances. Enfin, la collecte de données sur des profils de patients permettant d'équilibrer les classes S et NS permettra d'obtenir des résultats plus signifiants.

Remerciements

Cette étude a été réalisée dans le cadre des projets IS-OSA, financé par la Région Nouvelle Aquitaine et SOMVOICE, financé par le Labex BRAIN (Université de Bordeaux).

Références

ALDRICH M. S., CHERVIN R. D. & MALOW B. A. (1997). Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy. *Sleep*, **20**(8), 620–629.

CUMMINS N., BAIRD A. & SCHULLER B. (2018). Speech analysis for health : Current state-of-the-art and the increasing impact of deep learning. *Health Informatics and Translational Data Analytics*, **151**, 1–54.

DEGOTTEX G., KANE J., DRUGMAN T., RAITIO T. & SCHERER S. (2014). COVAREP — A collaborative voice analysis repository for speech technologies. In *IEEE - ICASSP*, p. 960–964. DOI : 10.1109/ICASSP.2014.6853739.

LITTNER M. R., KUSHIDA C., WISE M., DAVILA D. G., MORGENTHALER T., LEE-CHIONG T., HIRSHKOWITZ M., LOUBE D. L., BAILEY D., BERRY R. B., KAPEN S. & KRAMER M. (2005).

Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test. *Sleep*, **28**(1), 113–121.

MARTIN V. P., CHAPOUTHIER G., RIEANT M., ROUAS J.-L. & PHILIP P. (2020a). Using reading mistakes as features for sleepiness detection in speech. In *The 10th International Conference on Speech Prosody*.

MARTIN V. P., ROUAS J.-L., MICOULAUD-FRANCHI J.-A. & PHILIP P. (2020b). The Objective and Subjective Sleepiness Voice Corpora. In *12th Language Resources and Evaluation Conference*.

MARTIN V. P., ROUAS J.-L., THIVEL P. & KRAJEWSKI J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue*. DOI : 10.1109/SPED.2019.8906577.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PELLEGRINO F. & ANDRE-OBRECHT R. (2000). Automatic language identification : an alternative approach to phonetic modelling. *Signal Processing*, **80**(7), 1231–1244.

PHILIP P., DUPUY L., AURIACOMBE M., SERRE F., DE SEVIN E., SAUTERAUD A. & MICOULAUD-FRANCHI J.-A. (2020). Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *npj Digital Medicine*, **3**(1), 2. DOI : 10.1038/s41746-019-0213-y.

PHILIP P., MICOULAUD-FRANCHI J.-A., SAGASPE P., DE SEVIN E., OLIVE J., BIOULAC S. & SAUTERAUD A. (2017). Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports*, **7**(1), 426–456. DOI : 10.1038/srep42656.

ROUAS J.-L. & IOANNIDIS L. (2016). Automatic Classification of Phonation Modes in Singing Voice : Towards Singing Style Characterisation and Application to Ethnomusicological Recordings. In *Interspeech 2016*, p. 150–154.

ROUAS J.-L., SHOCHI T., GUERRY M. & RILLIARD A. (2019). Categorisation of spoken social affects in Japanese : human vs. machine. In *ICPhS*.

SANGAL R. (1999). Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy. *Clinical Neurophysiology*, **110**(12), 2131–2135.

SCHULLER B., BATLINER A., BERGLER C., POKORNY F. B., KRAJEWSKI J., CYCHOCZ M., VOLLMAN R., ROELEN S.-D., SCHNIEDER S., BERGELSON E., CRISTIA A., SEIDL A., WARLAUMONT A., YANKOWITZ L., NÖTH E., AMIRIPARIAN S., HANTKE S. & SCHMITT M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech 2019*.

SCHULLER B., STEIDL S., BATLINER A., SCHIEL F. & KRAJEWSKI J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *Interspeech 2011*, p. 3201–3204.

SJÖLANDER K. (2004). The Snack Sound Toolkit.

ÅKERSTEDT T. & GILLBERG M. (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci*, **52**, 29–37.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Article retiré à la demande des auteurs

Volume 1 : Journées d'Études sur la Parole, pages 415–423.

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Article retiré à la demande des auteurs

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Représentation phonologique des signes à deux mains en LSF : faut-il reconsidérer l'orientation absolue dans les modèles phonologiques des langues des signes ?

Justine Mertz[1, 2]
(1) Laboratoire de Linguistique Formelle, UMR 7110, CNRS, Université de Paris, France
(2) Département d'études cognitives, ENS, EHESS, CNRS, PSL, UMR 8129, France
justine.mertz93@gmail.com

RÉSUMÉ

Cet article a pour objectif i) de mettre en évidence le besoin d'une orientation *absolue* en plus d'une orientation *relative* afin de décrire la phonologie segmentale des langues des signes, et ii) d'améliorer les modèles actuels et leur permettre de rendre compte de la phonologie de signes autrement problématiques. Dans cette étude sur la langue des signes française, nous nous concentrons sur une catégorie de signes en particulier : les signes à deux mains produits sur le corps du signeur (contact avec une partie du corps autre que la main non-dominante). Nous montrons que l'orientation *relative* ne permet pas une description adéquate de ces signes lorsque l'orientation des deux mains doit être spécifiée, puisqu'elle peut capturer soit l'orientation entre les deux mains, soit l'orientation vis à vis du corps, mais pas les deux. Afin de modéliser l'orientation de ces signes dans un cadre formel, nous proposons l'implémentation de plans secondaires. Tandis que cette implémentation requiert des ajustements minimes dans les modèles formels actuels, son impact quant à la théorie générale de la phonologie segmentale des signes est, elle, importante. Les plans secondaires imposent des restrictions géométriques et forcent l'orientation absolue ; le concept d'orientation comme simple classe phonémique relationnelle n'est donc plus suffisant (du moins pour ces signes).

ABSTRACT

Phonological representation of 2-handed signs in LSF : reconsidering absolute orientation in the phonological models of sign language ?

The goals of this paper are i) to provide evidence for the need of *absolute* orientation in addition to *relative* orientation in order to fully capture sign language segmental phonology, and ii) to enrich current models which are only based on *relative* orientation so that the phonology of otherwise problematic signs is also accounted for. We focus on one particular category of signs in French sign language : two-handed signs produced on the signer's body (contact with a body part other than the non-dominant hand). We show that *relative* orientation does not meet descriptive adequacy when the orientation between the two hands has to be specified, since it either captures the orientation between the hands or the orientation towards the body, but not both. We propose secondary planes as a formal step to model orientation for these signs. While the implementation of this solution requires minimal changes in current formal models, the impact on the whole theory of segmental phonology for sign is quite big. The core conceptualization of orientation as a purely relational phonemic class does not hold anymore (at least not for these signs), as secondary planes impose geometrical restrictions that force absolute orientation.

MOTS-CLÉS : phonologie des langues des signes, orientation absolue, signes à deux mains,

Volume 1 : Journées d'Études sur la Parole, pages 424–432.
HAL : hal-02798567.

géométrie des traits, Prosodic Model.

KEYWORDS: sign language phonology, absolute orientation, two-handed signs, feature geometry, Prosodic Model.

1 Introduction

À l'heure actuelle, la phonologie segmentale des langues des signes est façonnée par la géométrie des traits et les relations de dépendance (Clements, 1985; Anderson & Ewen, 1987). Depuis les premiers travaux de Stokoe (1961), les modèles phonologiques des langues des signes présupposent trois classes phonémiques comme primitives : la configuration manuelle (forme de la main dominante (H1), et celle de la main non-dominante (H2) dans les signes à deux mains), l'emplacement (lieu où le signe est produit), et le mouvement (taille et forme du mouvement des articulateurs). Une quatrième classe phonémique, l'orientation, a ensuite été intégrée aux modèles (Battison, 1978). On retrouve également l'orientation en relation avec le mouvement tel que dans les changements d'orientation (voir p.ex. Blondel & Miller, 2001; Brentari, 1998) ; bien que cela soit très intéressant, ce phénomène est traité dans la représentation phonologique du mouvement des signes et n'est donc pas pertinent pour cette étude puisque nous nous concentrons uniquement sur l'orientation comme classe phonémique.

D'un point de vue phonétique, l'orientation peut être définie à partir de termes de géométrie de l'espace : une fois qu'un objet, quel qu'il soit, est placé dans l'espace (physique), il est automatiquement positionné dans un plan de cet espace, et *orienté* par rapport à ce plan. Ainsi, une fois placées dans l'espace, les mains sont automatiquement orientées vis à vis d'un plan dans ce même espace. D'un point de vue phonologique, en linguistique des langues des signes, l'orientation est dérivée à partir de l'interaction entre la configuration manuelle et l'emplacement ; elle est définie par une relation binaire entre une partie de la main (déterminée par un trait de la configuration manuelle) et la partie de l'emplacement où le signe est produit (déterminée part un trait de l'emplacement). L'orientation est donc traitée comme une classe phonémique dérivée (Brentari, 1998, 2002; Crasborn & van der Kooij, 1997; Liddell & Johnson, 1989; Uyechi, 1994, 1995). De plus, on trouve des paires minimales comme illustré dans la Figure 1 avec les signes FAX et METRO en LSF. Dans le signe FAX, l'orientation est dérivée de la relation entre le trait de la configuration manuelle [paume de la main] ([*palm*]) de la main dominante et le trait de l'emplacement [paume de la main] de la main non-dominante. Pour le signe METRO, seul le trait de la configuration manuelle est différent puisque le [dos de la paume] ([*back of palm*]) de la main dominante fait face au même emplacement, tout le reste est identique. L'orientation est donc l'unique paramètre qui permet de différencier ces deux signes.

Parmi ces quatre classes phonémiques des langues des signes, la configuration manuelle est celle qui a suscité le plus d'intérêt auprès des linguistes et est donc plus finement décrite et repésentée dans la littérature (Sandler, 1986; van der Hulst, 1993; Uyechi, 1994, 1995). Cela peut être expliqué par le fait que les traits de la configuarion manuelle possèdent le plus de contraste phonémique dans la phonologie des langues signées (Brentari, 1998). L'intérêt des chercheurs pour la description des traits et le degré de contraste des autres classes phonémiques n'a pas été aussi important que pour la configuration manuelle. Par exemple, très peu de généralisations et de contraintes sur l'emplacement ou le mouvement ont été proposées dans la littérature, et le nombre de comparaisons interlinguistiques est d'autant plus faible. Prenons par exemple l'emplacement. Les signes peuvent être produits soit dans l'espace neutre, qui correspond à l'espace devant le signeur, soit sur le corps du signeur. L'espace neutre est divisé en trois plans principaux, le plan horizontal y, le plan vertical x et le plan sagittal

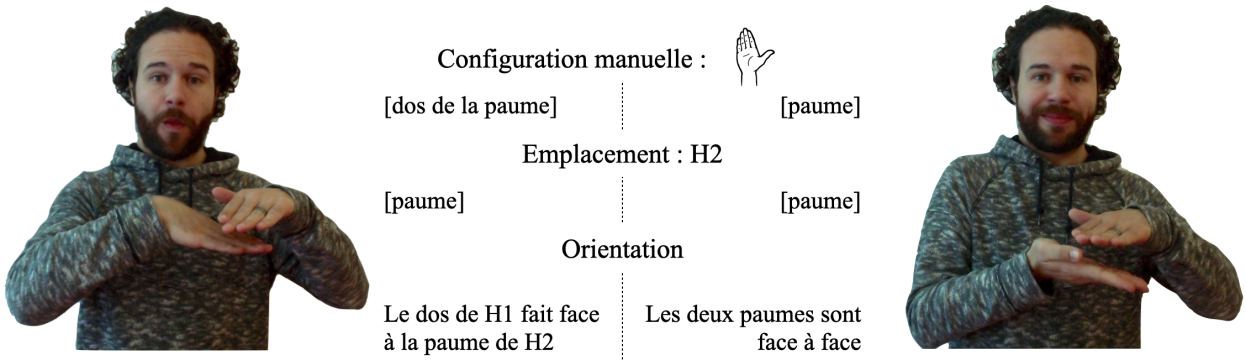


FIGURE 1 – Paire minimale MÉTRO (à gauche) ∼ FAX (à droite) en LSF. Les traits pertinents de la configuration manuelle et de l'emplacement sont indiqués pour chacun des signes, ainsi que l'orientation obtenue à partir de ces traits.

z ; s'il est produit dans l'espace neutre, le signe va être réalisé par rapport à l'un de ces plans. Le corps peut également être le plan lorsqu'il est nécessaire à la représentation phonologique du signe. Dans ce cas, le corps devient un plan, divisé en quatre grandes parties (la tête, le torse, le bras, et la main non-dominante), elles-même divisées en huit sous-parties, toutes justifiées phonologiquement (cf. Brentari, 1998). En ce qui concerne l'emplacement, seules deux contraintes générales ont été proposées (Brentari, 1998; Sandler & Lillo-Martin, 2006) :

1. Lorsqu'un signe est produit à un emplacement particulier sur le corps du signeur, le corps *doit être* l'emplacement ;
2. Pour les signes lexicaux, un seul plan peut être sélectionné.

Comme nous le verrons plus bas, ces contraintes génèrent un conflit systématique quant à l'analyse du contenu phonémique des signes, et ont des conséquences directes sur l'analyse d'une autre classe phonémique, l'orientation. Certains modèles définissent l'orientation en termes relatifs (Brentari, 1998; Crasborn & van der Kooij, 1997; Liddell & Johnson, 1989; Uyechi, 1994, 1995), c'est-à-dire à partir de traits des classes phonémiques primitives. Cette façon *relative* de dériver l'orientation remplace le concept d'orientation défini en des termes *absolus* présent dans d'autres modèles (Battison, 1978; van der Hulst, 1993; van der Hulst & Mills, 1996). Techniquement, l'orientation absolue était déterminée à partir du corps du signeur utilisé comme plan de référence, et la paume ou le poignet de la main dominante comme repère fixe de la main. Ainsi, en remplaçant l'orientation absolue par l'orientation relative, les modèles phonologiques des langues des signes sont devenus plus économes (moins de structure), et plus explicatifs (la nature d'une classe phonémique est saisie à partir de la définition des autres).

D'après Crasborn et van der Kooij (1997), l'orientation absolue reste nécessaire pour un nombre restreint de signes : les classificateurs et les "signes à deux mains asymétriques", soit les signes à deux mains avec des configurations manuelles différentes et dont la main non-dominante est statique (également connus sous le nom de "signes à deux mains de Type 3", Battison, 1978).

Dans cette étude, nous soulignons le caractère empirique et théorique trop stricte de cette conceptualisation de l'orientation et nous montrons que l'orientation absolue peut être étendue à d'autres catégories de signes. La suite de l'article est organisée de la façon suivante : dans la partie 2, nous utilisons des données de la Langue des Signes Française (LSF) pour montrer que i) l'implémentation de l'orientation absolue est nécessaire et n'est pas limitée aux deux catégories de signes évoquées précédemment (Battison, 1978), et que ii) l'orientation absolue n'est pas uniquement motivée par

l'iconicité dans les signes à deux mains mais également par des traits phonologiques. Dans la partie 3 nous proposons une analyse formelle capable de capturer l'orientation *absolue* sans apporter de changement radical à l'intuition initiale que l'orientation est principalement *relative* pour les signes. Afin de concrétiser cette analyse, nous l'implémentons dans le cadre du *Prosodic Model* (Brentari, 1998, 2002) bien qu'une stratégie similaire peut également être implémentée dans les autres modèles mentionnés précédemment. Enfin, dans la partie 4 nous présentons des extensions empiriques et théoriques possibles pour cet axe de recherche.

2 Le besoin d'orientation absolue

Le premier objectif de la modélisation du langage est de fournir les outils nécessaires à la description des propriétés interlinguistiques. En d'autres termes, si un modèle théorique particulier est construit à partir des propriétés et mécanismes d'une langue spécifique, il doit être assez robuste pour rendre compte de phénomènes identiques dans d'autres langues. La linguistique des langues signées, par opposition à celle des langues parlées, est une discipline relativement récente puisque les premiers travaux datent des années 1960 (Stokoe, 1960; Stokoe *et al.*, 1965). La construction de modèles théoriques représente donc un challenge pour la communauté scientifique. Généralement construits à partir des informations d'une seule langue signée (très souvent la langue des signes américaine), ils sont ensuite étendus aux autres langues des signes avec très peu d'adaptation. Cela concerne la plupart des modèles phonétiques et phonologiques, et touche toutes les classes phonémiques, dont l'orientation.

Comme mentionné dans la partie 1, certains modèles actuels définissent l'orientation en des termes relatifs (Brentari, 1998; Crasborn & van der Kooij, 1997; Liddell & Johnson, 1989; Uyechi, 1994, 1995) : elle est le résultat de l'interaction entre un trait de la configuration manuelle et un trait de l'emplacement. Un exemple est donné dans la Figure 2 avec le signe COLÈRE en LSF. L'orientation est définie en spécifiant le trait [bout des doigts sélectionnés] ([*tip of the selected fingers*] ou [*tip*]) orienté vers le trait [clavicule] ([*clavicle*]), un trait actif dès que le *corps* est sélectionné comme emplacement.

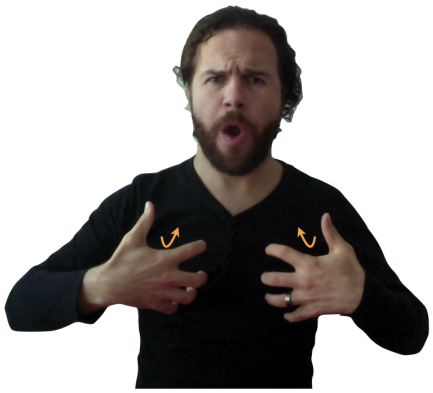


FIGURE 2 – Signe COLÈRE en LSF.

FIGURE 3 – Signe CEINTURE en LSF.

On remarque que le signe COLÈRE répond correctement aux deux contraintes de l'emplacement indiquées dans la partie 1 (Brentari, 1998; Sandler & Lillo-Martin, 2006). En effet, les mains sont en contact avec le corps, le *corps* est donc l'emplacement (cf. contrainte 1), et aucun autre plan n'est

sélectionné (cf. contrainte 2).

Cependant, ces deux contraintes restreignent fortement la forme des signes. Elles prédisent notamment que lorsque les mains des signes à deux mains sont articulées sur une région du corps (autre que la main non-dominante), celles-ci ne devraient jamais se toucher. Le cas contraire engendrerait un conflit dans la détermination et la spécification de l'emplacement du signe. Or, ce type de signes existe en LSF comme nous pouvons le voir avec le signe CEINTURE illustré dans la Figure 3.

D'une part, les deux mains touchent la taille et génèrent ainsi la sélection du corps comme plan pour établir l'emplacement. D'autre part, les deux mains se touchent et provoquent ainsi la production du signe sur le plan sagittal (qui est un plan de l'espace neutre). Un cas similaire impliquant le torse et le plan horizontal est représenté par le signe CODA en LSF (Figure 4). Ce phénomène est également observé dans d'autres langues signées comme nous pouvons le voir avec le signe CANDIDAT en Langue des Signes Italienne (LIS) présenté en Figure 5.

FIGURE 4 – Signe CODA en LSF.

FIGURE 5 – Signe CANDIDAT en LIS.

Pour rendre compte de ces signes, on pourrait stipuler que l'emplacement est non-spécifié (ou sous-spécifié) dans les formes sous-jacentes, et que la phonétique et la phonotactique complètent le matériel articulatoire (Crasborn & van der Kooij, 1997). Il est important de souligner que si la composante iconique du signe CEINTURE est présente, elle ne l'est pas pour les signes CODA (LSF) et CANDIDAT (LIS) [1].

Cette analyse ne permet cependant pas de résoudre le problème lié à l'orientation. En effet, les modèles actuels ne peuvent décrire que partiellement l'orientation de ces signes. Dans le cas de CEINTURE, si le trait [bout des doigts sélectionnés] (ici l'index et le pouce) et le trait [taille] ([*waist*]) sont spécifés, l'orientation par rapport au corps est bien définie, tandis que le contact entre les deux mains ne l'est pas. Dans ce cas, rien n'empêche les mains de se croiser puisque le plan sagittal n'est pas spécifié. Dans la situation où le trait [partie radiale des doigts sélectionnés] ([*radial*]) et le plan sagittal sont spécifiés, le contact entre les deux mains est représenté tandis que l'orientation par rapport au corps ne l'est pas. Le signe devrait donc obligatoirement être produit dans l'espace neutre puisque le corps n'est pas sélectionné (cf. contrainte 1), ce qui n'est pas le cas.

1. Le signe CEINTURE a été présenté à vingt entendants non-signeurs : aucun n'en a deviné le sens. Le lien entre signifiant et signifié est activé uniquement en connaissance du signifié, sa valeur iconique n'est donc pas maximale puisqu'il ne s'agit pas d'un signe transparent.

3 Analyse

Dans la partie 2 nous avons vu que la sous-spécification ne permet pas de résoudre le problème lié à l'identification de l'emplacement et de l'orientation dans certains signes à deux mains produits sur le corps. Dans cette partie, nous proposons une analyse alternative à partir de l'insertion de plans *secondaires* générée par une relation supplémentaire entre les deux mains. Dans le cas des signes à deux mains, ce dernier correspond à une orientation spécifique entre les deux mains pouvant se traduire par un trait [contact] supplémentaire.

La sélection secondaire dans les langues des signes a déjà été proposée dans la littérature pour résoudre des cas problématiques liés à la sélection des doigts (Eccarius, 2008). Ici nous proposons simplement de l'étendre au domaine de l'emplacement. Une fois que les plans secondaires sont introduits dans la représentation structurelle de ces signes, les traits de configuration manuelle peuvent y faire référence, résolvant ainsi à la fois les problèmes d'identification de l'emplacement approprié et de l'orientation. Nous illustrons cette solution avec le signe CEINTURE en LSF.

Supposons que le corps du signeur soit l'emplacement primaire d'après la contrainte générale 1 introduite dans la partie 1. Dans ce cas, le trait [bout des doigts sélectionnés] ([*tip*]) peut être spécifié en tant que partie de la main face à l'emplacement primaire. Le plan secondaire, ici le plan sagittal, empêche ensuite les deux mains de se croiser, et sert ainsi de point de référence pour l'"orientation secondaire", elle-même activée par un second trait de configuration manuelle. Dans ce cas, celui-ci correspond au trait [partie radiale des doigts sélectionnés] ([*radial*]) face au plan sagittal ; le point de contact spécifique entre les deux mains est ainsi garanti.

Une formalisation de ce développement dans le cadre du *Prosodic Model* est proposée dans la Figure 6. Le noeud de l'emplacement (POA pour *Place Of Articulation*) domine deux branches, une pour chaque plan : le plan du corps x avec le trait pertinent [taille] ([*waist*]), et le plan sagittal z. Dans la branche de configuration manuelle (A pour *Articulator*), le noeud H1 (main dominante) domine directement les traits pertinents nécessaires à l'orientation du signe, soit les traits [bout des doigts sélectionnés] et [partie radiale des doigts sélectionnés], chacun faisant référence à un plan spécifique (respectivement le plan du corps x et le plan sagittal z).

L'implémentation de plans secondaires pour éviter le croisement des mains dans les signes tels que CEINTURE crée une prédiction intéressante quant aux signes à deux mains dans lesquels les mains sont déjà croisées (c.-à-d. les signes où les mains sont dans la position contralatérale au début du signe). Si ces signes ne sélectionnent pas d'emplacement secondaire, les mains sont libres de "se décroiser" lors d'un mouvement de trajectoire. Ce phénomène est illustré dans la Figure 7 avec le signe CHÔMAGE en LSF, où le corps est le seul plan sélectionné.

Or, lorsqu'un plan secondaire est sélectionné, il est prédit que les mains restent du même côté contralatéral malgré un mouvement de trajectoire ; un exemple est donné avec le signe OS en LSF (Figure 8) où le corps est sélectionné pour représenter l'orientation entre [clavicule] et [avant des doigts] ([*finger front*]), et le plan sagittal est inséré comme plan secondaire pour représenter l'orientation de la [partie radiale] des deux mains, et représente une ligne imaginaire à ne pas croiser au niveau du contact entre les deux mains.

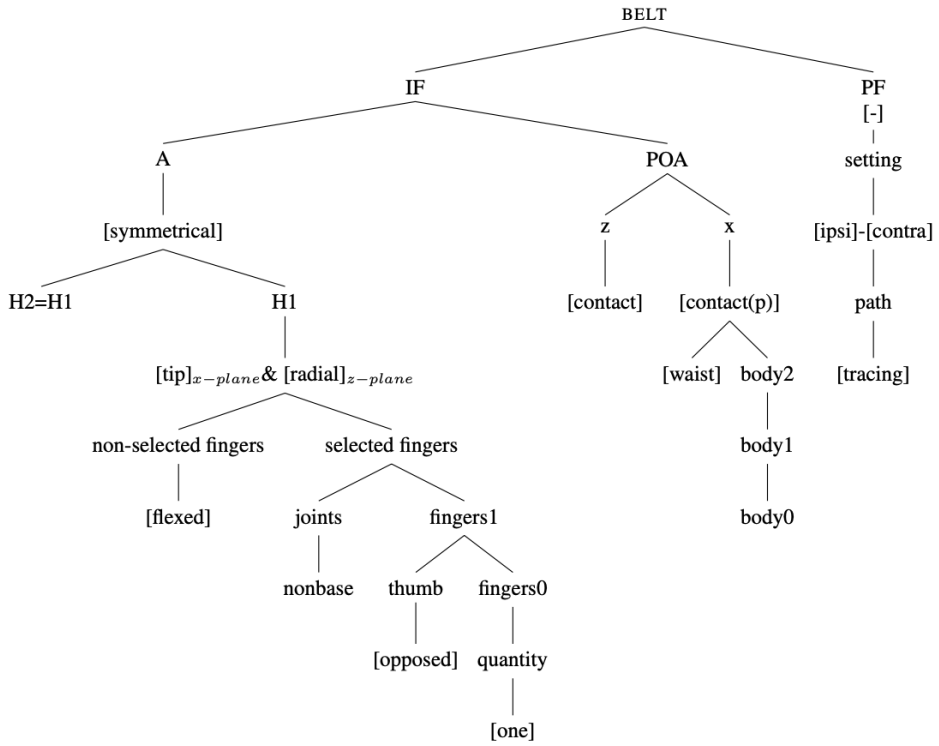


FIGURE 6 – Structure du signe CEINTURE dans le cadre du *Prosodic Model*.

4 Discussion

L'analyse présentée dans la partie précédente nous permet de représenter l'orientation des signes à deux mains grâce à l'introduction de plans secondaires dans la description formelle des signes. Notre implémentation de cette analyse dans le cadre du *Prosodic Model* (Brentari, 1998) n'est cependant pas sans conséquence. Ici, nous discutons deux point essentiels. Le premier concerne une surgénération tandis que le second traite du statut de l'orientation dans les signes qui nécessitent des plans secondaires.

Tout d'abord, l'analyse que nous offrons ne restreint pas l'utilisation de plans secondaires et prédit ainsi une surgénération importante, du moins pour les signes à deux mains dont ceux pour lesquels l'implémentation d'un plan secondaire n'est pas nécessaire. Limiter cette surgénération est possible en spécifiant le contexte dans lequel des plans secondaires sont introduits. Une solution pourrait être qu'un plan secondaire est généré dans tous les signes à deux mains puis supprimé dans les cas de redondance.

En ce qui concerne le statut de l'orientation, il est important de noter que l'impact de notre analyse est plus fort que ce que suggère la formalisation illustrée dans la Figure 6. En effet, tandis que l'implémentation de doigts sélectionnés secondaires (Eccarius, 2008) n'avait pas de conséquences majeures dans la conceptualisation de la phonologie segmentale des signes, l'extension d'une sélection secondaire à l'emplacement pourrait (ré)introduire un changement majeur et plus radical. Dans les anciens modèles de phonologie des langues des signes, l'orientation *absolue* était déterminée par le corps comme plan de référence et la paume comme point de référence (Sandler, 1986; Uyechi, 1994, 1995). Dans notre étude nous introduisons l'orientation *absolue* d'une façon plus indirecte puisque

FIGURE 7 – Signe CHÔMAGE en LSF.

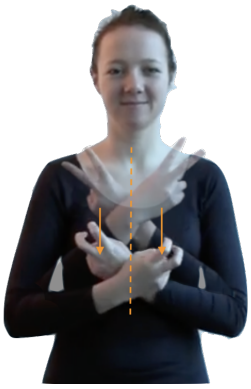


FIGURE 8 – Signe OS en LSF.

l'orientation dépend de deux plans (soit le corps et le plan sagittal dans l'exemple discuté dans cet article) ; notre implémentation permet donc de rendre compte de l'orientation des signes à partir d'une seconde coordonnée spatiale, c'est l'interaction de l'orientation établie entre les plans primaire et secondaire qui génère l'orientation *absolue*.

5 Conclusions

Les modèles théoriques sont extrêment importants pour définir et déterminer les généralisations linguistiques. Cependant, les étendre d'une langue à une autre sans adaptation peut mener à des inexactitudes empiriques. L'orientation *relative* étant, d'apparence, suffisante pour représenter la forme phonologique des signes, les chercheurs ont éliminé l'orientation *absolue* de la description des langues des signes. Les données de signes à deux mains en LSF et en LIS montrent cependant que l'orientation *relative* n'est pas suffisante, et que l'orientation *absolue* est également nécessaire à la description phonologique de certains signes.

D'un point de vue théorique plus général, en réintroduisant l'orientation *absolue*, notre solution représente une innovation majeure dans la description des signes. La modification et l'adaptation des modèles actuels reste cependant minime, puisque pour obtenir l'orientation absolue, l'unique condition est d'avoir un emplacement "secondaire". On notera que spécifier une simple paire de traits pour l'orientation absolue laisse une marge de manoeuvre importante pour tout ajustement phonétique, et garde ainsi la flexibilité nécessaire à l'étude des cas discutés par Battison (1978).

Il est également important de souligner que ces changements ne sont pas de simples ajustements phonétiques puisqu'ils sont motivés par la phonologie (p.ex. présence de paires minimales). L'iconicité n'est également pas une motivation suffisante puisque les exemples que nous avons vu en LSF avec le signe CODA ou encore en LIS avec le signe CANDIDAT (Figures 4 et 5) ne sont pas des cas isolés.

De nombreuses questions quant à la portée de ce besoin d'orientation absolue restent néanmoins en suspend. Une étude approfondie de la LSF sera menée afin d'identifier avec précision les catégories de signes affectés par ces modifications et ainsi déterminer les règles phonologiques sous-jacentes à l'orientation absolue. Nous nous pencherons également sur les études théoriques menées sur l'animation numérique des signes en LSF (Filhol, 2008). Cela nous permettra également d'en connaître les limites et ainsi éviter la surgénération actuellement présente dans notre analyse.

Références

ANDERSON J. M. A. & EWEN C. J. (1987). *Principles of dependency phonology*. Volume 47. Cambridge University Press.

BATTISON R. (1978). Lexical borrowing in American sign language.

BLONDEL M. & MILLER C. (2001). Movement and rhythm in nursery rhymes in LSF. *Sign Language Studies*, p. 24–61. Publisher : JSTOR.

BRENTARI D. (1998). *A Prosodic Model of Sign Language Phonology*. The MIT Press. DOI : 10.7551/mitpress/5644.001.0001.

BRENTARI D. (2002). Modality differences in sign language phonology and morphophonemics. In R. P. MEIER, K. CORMIER & D. QUINTO-POZOS, Éds., *Modality and structure in signed and spoken languages*, p. 35–64. Cambridge : Cambridge University Press. DOI : 10.1017/CBO9780511486777.003.

CLEMENTS G. N. (1985). The Geometry of Phonological Features. *Phonology Yearbook*, p. 225–252.

CRASBORN O. & VAN DER KOOIJ E. (1997). Relative orientation in sign language phonology. *Linguistics in the Netherlands*, **1997**, 37–48.

ECCARIUS P. N. (2008). *A constraint-based account of handshape contrast in sign languages*. Purdue University.

FILHOL M. (2008). *Modèle descriptif des signes pour un traitement automatique des langues des signes*. PhD Thesis.

LIDDELL S. K. & JOHNSON R. E. (1989). American sign language : The phonological base. *Sign language studies*, **64**(1), 195–277.

SANDLER W. (1986). The spreading hand autosegment of American Sign Language. *Sign Language Studies*, p. 1–28.

SANDLER W. (2006). From phonetics to discourse : the nondominant hand and the grammar of sign language. *Laboratory Phonology*, **8**.

SANDLER W. & LILLO-MARTIN D. (2006). *Sign language and linguistic universals*. Cambridge University Press.

STOKOE W., CASTERLINE D. & CRONEBERG C. (1965). *A dictionary of ASL on linguistic principles*. Gallaudet College Press Washington, DC.

STOKOE W. C. (1960). Sign Language Structure : An Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education*, **10**(1), 3–37. DOI : 10.1093/deafed/eni001.

UYECHI L. (1994). Local and global signing space in American Sign Language. In *Proceedings of NELS.*, volume 2, p. 589 : GLSA, UMass/Amherst.

UYECHI L. (1995). *The Geometry of Visual Phonology*. Thèse de doctorat, Stanford, California.

VAN DER HULST H. (1993). Units in the analysis of signs. *Phonology*, **10**(2), 209–241.

VAN DER HULST H. & MILLS A. (1996). Issues in sign linguistic : Phonetics, phonology and morpho-syntax. *Lingua*, **98**(1-3), 3–17.

VAN DER KOOIJ E. (2002). *Phonological categories in Sign Language of the Netherlands : The role of phonetic implementation and iconicity*. Netherlands Graduate School of Linguistics.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La mobilisation du tractus vocal est-elle variable selon les langues en parole spontanée ?

Christine Meunier[1], Morgane Peirolo[12], Brigitte Bigi[1]
(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(2) Aix Marseille Univ, Institut Convergence ILCB, Marseille, France
christine.meunier@univ-amu.fr, morgane.peirolo@gmail.com,
brigitte.bigi@univ-amu.fr

RÉSUMÉ

L'objectif de ce travail est de quantifier les positions articulatoires théoriques lors de la production de la parole spontanée dans trois langues. Chaque langue dispose d'un inventaire phonologique spécifique. Mais ces spécificités ne sont pas représentées telles quelles en parole spontanée dans laquelle les phonèmes n'ont pas tous la même fréquence d'apparition. Nous avons comparé trois langues (polonais, français et anglais américain) présentant des différences notables dans leur inventaire phonologique. Des positions articulatoires ont été calculées sur la base des fréquences des phonèmes dans chacune des trois langues dans des corpus de parole spontanée. Etonnamment, les résultats tendent à montrer que les positions articulatoires majoritaires sont très similaires dans les trois langues. Il semble ainsi que l'usage de la parole spontanée, et donc la distribution des phonèmes dans les langues, gomme les disparités des systèmes phonologiques pour tendre vers une mobilisation articulatoire commune. Des investigations plus approfondies devront vérifier cette observation.

ABSTRACT

Does vocal tract use depend on language characteristics in spontaneous speech?
The aim of this work is to quantify the theoretical articulatory positions during the production of spontaneous speech for three languages. Each language has a specific phonological inventory. However, these specificities are not represented as such in spontaneous speech in which phonemes do not have the same relative frequency. We compared three languages (Polish, French and American English) with notable differences in their phonological inventory. Articulatory positions were calculated according to phoneme frequencies in the three languages through spontaneous speech corpora. Surprisingly, the results tend to show that preferred articulatory positions are very similar in the three languages. Thus it seems that spontaneous speech production, and therefore phonemes distribution in languages, erases the disparities of phonological systems in order to provide similar articulation. Further investigation should verify this observation.

MOTS-CLÉS : Position articulatoire, langues, corpus, parole spontanée, fréquences phonétiques.
KEYWORDS : Articulatory position, languages, corpus, spontaneous speech, phoneme frequencies.

Volume 1 : Journées d'Études sur la Parole, pages 433–441.
HAL : hal-02798569.

1 Introduction

L'ensemble des humains partagent un appareil vocal tout à fait semblable leur permettant de produire le langage oral, mais tous ne parlent pas la même langue. Or, chaque langue dispose d'un inventaire phonologique spécifique puisant dans l'ensemble des sons possibles que l'appareil vocal peut produire. Des tendances universelles ont pu être observées permettant de mettre en évidence la façon dont les systèmes s'organisent et se différencient (Vallée et al., 1999, Ladefoged & Maddieson, 1996). Par exemple, les dialectes arabes sont caractérisés par des positions très arrières (consonnes uvulaires, pharyngales ou glottales) ; le suédois possède de nombreuses voyelles antérieures ; le français n'a pas de consonnes très postérieures (sauf /R/, uvulaire) et un tiers des voyelles sont antérieures tandis qu'un autre tiers est arrondi. Ces particularités permettent de caractériser le système phonologique dans ces langues. Toutefois, il reste à déterminer si l'inventaire phonologique d'une langue est effectivement représentatif de la mobilisation du tractus vocal lors de la production de cette langue dans un usage courant. L'inventaire phonologique d'une langue attribue le même poids à chaque phonème et donc à chacune de ses propriétés articulatoires. Mais les phonèmes sont diversement présents dans l'usage courant de la langue orale et, en conséquence les *préférences* articulatoires d'une langue vont dépendre de la fréquence d'utilisation des phonèmes dans cette langue. Par exemple, le geste d'arrondissement en français pourrait être minoré dans la mesure où les voyelles d'arrière (arrondie) sont peu fréquentes en parole spontanée (Meunier & Espesser, 2011). Par conséquent, dans quelle mesure la mobilisation du tractus vocal dans une langue est-elle pondérée par la fréquence relative des phonèmes dans l'usage courant de la langue?

L'objectif de ce travail est donc de **quantifier les positions articulatoires théoriques** lors de la production de la parole spontanée. Nous comparons les positions articulatoires relatives aux phonèmes dans trois langues : le polonais, le français et l'anglais américain. Pour représenter les positions articulatoires, nous nous sommes inspirés de la terminologie fournie par Articulatory Phonology (AP, Browman & Goldstein, 1992). AP propose une représentation des positions articulatoires au cours de la production de la parole. La terminologie utilisée convient bien à ce travail dans la mesure où il est question de décrire précisément des positions et mouvements articulatoires plutôt que de proposer des traits distinctifs. Il n'est pas besoin en effet, dans ce travail, de distinguer les phonèmes, mais plutôt d'inventorier leurs caractéristiques articulatoires, même si elles sont redondantes. Notons que ce travail ne fournit pas d'informations sur les articulations *réelles* dans la production de la parole et elle ne fournit pas non plus d'informations sur l'articulation *dynamique* des sons. En fait, il ne s'agit pas d'une étude articulatoire. A cette étape, nous souhaitons fournir une indication sur les tendances des positions *attendues* dans les trois langues lors de leur production courante. C'est la raison pour laquelle nos résultats sont basés sur l'annotation phonétique de trois corpus de parole spontanée permettant d'obtenir les fréquences relatives de chaque phonème.

2 Systèmes phonologiques, corpus et affectation des poids articulatoires

Plusieurs critères ont déterminé le choix de ces trois langues, soit le polonais (PO), le français (FR) et l'anglais américain (AE). En premier, nous étions contraints par l'existence de corpus annotés phonétiquement pour lesquels il était possible d'obtenir la fréquence de chaque phonème. D'autre part, nous avons essayé de choisir des corpus assez similaires pour ce qui est du type de parole produite : pour les trois corpus il s'agit de parole spontanée non familière dont les conditions

représentent une forme de parole guidée (conversation téléphonique pour AE et PO, Maptask pour PO et FR). Ensuite, nous avons privilégié ces trois langues car nous disposions de locuteurs natifs phonéticiens susceptibles de pouvoir nous éclairer sur les systèmes phonologiques de leur langue et sur les réalisations effectives. Enfin, ces trois langues présentent des systèmes phonologiques différents : peu de voyelles pour le polonais en comparaison du français et de l'anglais américain; un grand nombre de fricatives et d'affriquées en polonais ; un grand nombre de voyelles arrondies en français. En conséquence, des différences phonologiques pouvant entrainer une mobilisation variée du tractus vocal. Pour chaque langue nous décrivons les caractéristiques du système phonologique, puis le corpus utilisé et enfin quelques informations sur la distribution des phonèmes.

2.1 Polonais (PO)

Le polonais (PO) possède 5 voyelles (/a/ /ɛ/ /i/ /u/ /ɔ/) et s'apparente ainsi au système « le plus populaire » des langues du monde (Vallée et al., 1999) avec, en plus, une voyelle centrale fermée /ɨ/. Deux voyelles nasales viennent compléter cet inventaire. Le système consonantique est plus riche et plus complexe. Il dispose d'un grand nombre de fricatives et affriquées combinant de nombreux lieux d'articulation (labiales, alvéolaires, rétroflexes palatales et vélaires), modes d'articulation (fricatives et affriquées) et modes de phonation permettant de produire 15 consonnes différentes[1]. Avec 6 plosives, 3 nasales et 4 approximantes, la totalité des consonnes représente 31 segments, soit 39 phonèmes au total avec un fort déséquilibre en faveur des consonnes. Notons toutefois que l'inventaire des consonnes (notamment des affriquées) pourrait être surestimé comparé à d'autres sources[2].

Le corpus polonais utilisé ici est *Paralingua* (Klessa et al., 2013). Il s'agit un recueil d'enregistrements téléphoniques de dialogues dans une tâche de type *Map Task*. Des paires de locuteurs ont été enregistrés dans des séances de 30mn. Le corpus a été transcrit et aligné en utilisant le code SAMPA[3]. Notre étude porte sur des séquences de 2 à 4mn pour 20 dialogues, soit 145mn d'enregistrement. Nous avons retenu 39 phonèmes-type pour une production de 38.236 segments phonétiques réalisés dans cet extrait. Il s'agit, nous le verrons, d'un corpus de taille très réduite comparé à ceux utilisés pour le FR et l'AE. Notre informatrice polonaise nous a indiqué que certaines consonnes affriquées apparaissent très rarement dans le lexique. Toutefois les phonèmes du corpus sont tout à fait comparables à ceux décrits dans l'inventaire.

2.2 Français (FR)

Le français (FR) possède 11 voyelles orales dont la majorité (7) est arrondie. Trois voyelles nasales dont une arrondie viennent compléter l'inventaire (Fougeron & Smith, 1999). La particularité du français est de posséder des voyelles antérieures arrondies (/y/ par exemple) ce qui est rare dans des langues du monde. Par ailleurs, les voyelles antérieures sont majoritaires (9 sur 14). Le français compte également 20 consonnes dont 6 plosives et 6 fricatives avec alternance de voisement. Les lieux d'articulation sont répartis entre les lèvres et le voile du palais, /R/ étant la consonne la plus

[1] https://en.wikipedia.org/wiki/Polish_phonology
[2] http://www.phonetics.ucla.edu/appendix/languages/polish/polish.html
[3] https://www.phon.ucl.ac.uk/home/sampa/polish.htm

postérieure. Trois glissantes sont également présentes aux lieux d'articulation correspondant aux 3 voyelles fermées. Le français compte donc en tout 34 phonèmes.

Le corpus français utilisé ici est le *Aix Map Task corpus* (Gorisch et al., 2014). Le protocole suivi et celui établi à Edinburgh (Anderson et al., 1992) dont l'objectif principal et de faire interagir des participants dans une tâche communicative ayant pour but la description d'un itinéraire via une carte. Quatre paires de locuteurs (2 hommes et 6 femmes) ont réalisé cette tâche pour un total de 2h18mn d'enregistrement, soit 26.706 mots. Le corpus a été transcrit et aligné. L'annotation phonétique est en SAMPA[4]. Pour notre étude nous avons utilisé 32 phonèmes-type et 96.090 segments phonétiques produits. La liste des phonèmes du corpus est très similaire à celle de l'inventaire phonologique à l'exception des voyelles /ɔ/ et /o/qui sont regroupées dans le corpus. De même la nasale palatale /ɲ/ n'est pas représentée.

2.3 Anglais Américain (AE)

L'anglais-américain (AE) est constitué d'un système vocalique riche et complexe. Outre les caractéristiques classiques des voyelles, on note la distinction entre *tense* et *lax* ainsi que de nombreuses diphtongues. La répartition entre voyelles antérieures et postérieures est équilibrée et seules les voyelles postérieures *tense* sont arrondies. Ladefoged (1999) distingue 11 monophtongues et 3 diphtongues. Le système de consonnes comporte 9 fricatives (avec alternance de voisement) et 2 affriquées. Six plosives et 3 nasales viennent compléter l'inventaire. On compte également 4 approximantes (/l/ /r/) dont 2 glissantes (/j/ /w/). L'anglais-américain compte donc 38 phonèmes.

Le corpus anglo-américain utilisé ici est le *Buckeye Corpus of conversational speech* (Pitt et al., 2005) contenant 40 locuteurs enregistrés lors de conversations téléphoniques. Le corpus, contenant environ 300.000 mots, a été transcrit et aligné dans un codage assez spécifique décrit dans le manuel du corpus[5]. La correspondance et l'identification des voyelles n'a pas été simple mais nous avons bénéficié de l'aide de locuteurs natifs. Parmi l'ensemble des symboles décrits, nous avons retenus 39 phonèmes-types représentants 1.536.801 segments phonétiques produits dans le corpus. Là encore, on est assez proche de l'inventaire dressé plus haut. La fréquence des phonèmes présents dans ce corpus est donnée par Yang (2012).

Les limites de la comparaison des productions des trois corpus concernent évidemment la taille très hétérogène des corpus avec un nombre de phonèmes nettement plus réduit pour le polonais. Toutefois, nous avons pu observer que les distributions des phonèmes et donc leur fréquence respective dans un type de corpus sont assez stables et apparaissent très rapidement avec peu de données et peu de locuteurs. Cela est probablement dû au fait que la fréquence des phonèmes est dépendante de la fréquence des mots dont la distribution est très contrainte par le discours.

Dans la mesure où la fréquence des phonèmes pourrait être déterminante dans la représentation des positions articulatoires, nous présentons ici les phonèmes les plus fréquents (50% des occurrences) dans chacune des trois langues (table 1). Notons que les phonèmes fréquents communs aux trois langues sont la plosive alvéolaire /t/ et la voyelle moyenne antérieure /e/ ou /ɛ/.

[4] https://www.phon.ucl.ac.uk/home/sampa/french.htm
[5] https://buckeyecorpus.osu.edu/BuckeyeCorpusmanual.pdf (page 22)

PO	a (11%)	**e** (13%)	o (9%)	**t** (6%)	j (5%)	n (4%)	ɪ (4%)		
FR	a (11%)	R (8%)	**t** (7%)	d (6%)	l (5%)	**ɛ** (5%)	s (5%)	**e** (5%)	
AE	ʌ (8%)	ɪ (8%)	n (6%)	s (5%)	**ɛ** (5%)	**t** (4%)	i (4%)	r (4%)	k (4%)

TABLE 1 : Phonèmes les plus fréquents dans les trois langues totalisant 50% des occurrences de phonèmes dans les corpus respectifs

On notera également que la répartition des phonèmes dans 50% des occurrences est différentes dans les trois langues : peu de phonèmes (surtout des voyelles) pour PO ; un nombre plus important pour FR et AE avec une proportion de voyelles et consonnes plus équilibré.

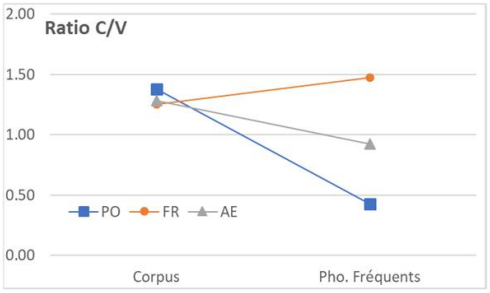


FIGURE 1 : Ratio Consonnes/Voyelles dans totalité de chacun des corpus et au sein des phonèmes les plus fréquents (50% des occurrences, voir table 1).

Ainsi, les trois langues ont un ratio C/V très similaire lorsque l'on considère la totalité des occurrences des corpus. Toutefois, si l'on ne prend en compte que les phonèmes les plus fréquents, les trois langues montrent un ratio C/V très différent (figure 1). Pour PO, 3 voyelles représentent 38% des occurrences du corpus, alors que AE et surtout FR montrent une proportion de consonnes plus importante dans les phonèmes fréquents.

2.4 Attribution des positions articulatoires

Notre objectif est de rendre compte des positions articulatoires les plus fréquentes dans chacune des langues et dans la production courante de la parole. Pour cela, il n'est pas nécessaire que ces critères articulatoires soient distinctifs, mais plutôt qu'ils expriment, le mieux possible, l'état *théorique* des articulateurs au cours de sa production. Nous précisons « théorique » car il ne s'agit pas de la réalisation effective des phonèmes, mais plutôt d'une réalisation attendue. Browman & Goldstein (1992) proposent d'exprimer la représentation phonologique des sons de la parole par un ensemble de *variables* et de *dimensions* du tractus vocal (*Tracts Variables & Dimensions*). Les dimensions précisent l'état de chaque variable. Parmi les différentes propositions des auteurs, nous avons sélectionné quatre variables (table 2) et nous avons effectué un regroupement (les variables *Tongue Tip* et *Tongue Body* ont été regroupées en *Tongue*).

Tract Variables	***Dimensions***
GLO - Glottis	closed, wide
VEL - Velum	closed, wide
LP - Lip Protrusion	protruded, not protruded
TCL - Tongue Constriction Location	labial, dental, alveolar, postalveolar, palatal, velar, uvular

TABLE 2 : représentation des *Tract Variables* et *Dimensions* selon AP et utilisées dans cette étude.

Pour les variables GLO, VEL et LP, l'attribution des dimensions est binaire et il n'y a pas d'ambiguïté concernant cette attribution. Pour TCL, chaque consonne se voit attribuer le lieu d'articulation qui la caractérise. L'opération est globalement assez simple excepté pour les consonnes rétroflexes de PO et AE qui ont été codées en *postalveolar* (comme il est souvent indiqué dans les inventaires phonologiques). En revanche, les voyelles sont généralement dites antérieures, centrales ou postérieures mais ne sont pas caractérisées par un lieu articulatoire précis. Il a été décidé que les voyelles antérieures étaient codées *palatal* et celles d'arrière *velar*, ce qui correspond à leur position effective. Notre méthode consiste donc à assigner à chaque phonème l'ensemble des dimensions qui le caractérise. Ensuite chaque dimension se voit attribuée une valeur correspondant à la fréquence de chaque phonème dans le corpus (table 3). En additionnant l'ensemble de ces valeurs dans chaque colonne nous obtenons la fréquence d'utilisation de chaque dimension. Cette opération est répétée dans chaque langue et pour chacune des variables.

TCL	alveolar	postalveolar	palatal	velar
e			10.75	
s	2.49			
k				3.51
dz'		0.65		
n	4.08			
…				

TABLE 3 : Exemple (pour PO) de l'attribution des fréquences des phonèmes (en %) pour la variable Tongue Constriction Location (TCL). Dans le corpus PO, la fréquence de la consonne /s/ est 2.49%. Cette valeur est reportée dans la colonne *alveolar*, caractéristique de cette consonne.

3 Résultats

3.1 Glottis, Velum et Lip Rounding

Pour ces trois variables l'attribution des dimensions est binaire. GLO et VEL sont en position *closed* ou *wide*. LP est *protruded* ou *not protruded*. Pour GLO et VEL on note une impressionnante similarité pour les trois langues : une forte préférence pour la position *closed* autour de 78% en moyenne pour GLO et 89% pour VEL (figure 2). Il n'y a quasiment pas de différence entre les trois langues. Notons que la dimension *wide* pour VEL représente 15% de l'inventaire de FR, alors que les occurrences *wide* dans le corpus sont à 10.7%. Pour LP, les tendances sont également très similaires pour les trois langues avec la dimension *not protruded* très majoritaire (87% en moyenne). On note une proportion légèrement moins importante de *not protruded* pour FR (81%), probablement due au nombre important de voyelles arrondies dans le système vocalique français. Toutefois, 28% des phonèmes FR sont arrondis et cette proportion se trouve minorée dans le corpus, ce qui suppose une fréquence moins importante des voyelles arrondies en FR.

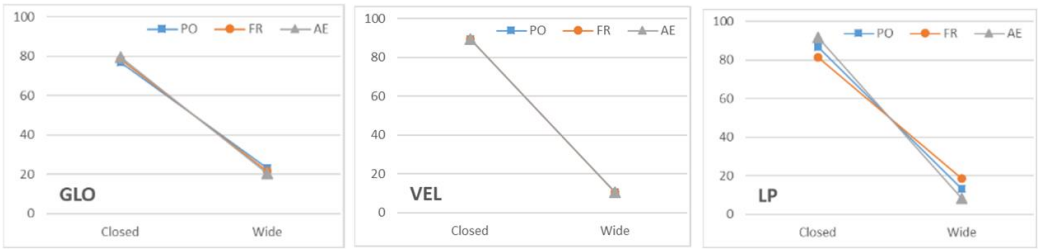


FIGURE 2 : Fréquence d'apparition des dimensions pour les trois variables GLO, VEL et LP

3.2 Tongue Constriction Location

La position de la langue lors de la constriction présente également de nombreuses similitudes dans les trois langues (figure 3). On observe une forte prépondérance des dimensions *alveolar*, *palatal* et *velar* (respectivement en moyenne 25%, 32% et 22%), la dimension *postalveolar* étant nettement moins utilisée (5%). La dimension *dental* n'est présente que pour AE et représente une faible proportion (3.5%). De même, la dimension *uvular* n'est présente qu'en français via le seul phonème /R/ qui, à lui seul, totalise 7.6% des occurrences. Ainsi, au cours de la production de parole spontanée, on observe que dans ces trois langues, la position de la langue est majoritairement positionnée dans trois régions : palais dur, alvéoles puis palais mou.

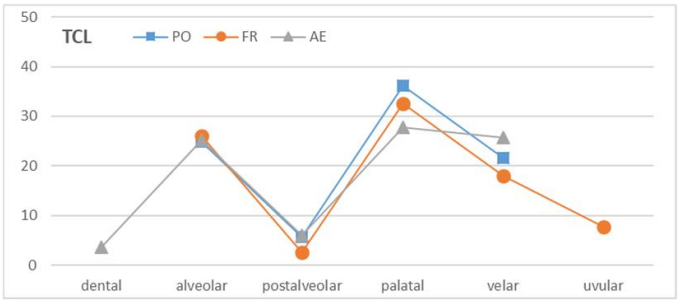


FIGURE 3 : Fréquence d'apparition des dimensions pour la variable TCL

Voyelles et consonnes sont présentées conjointement ici (figure 3) de façon à obtenir la configuration du tractus pendant la parole spontanée. Les voyelles n'étant présentes que sur les dimensions *palatal* et *velar*, il est clair que ces positions sont très sollicitées. Nous avons donc distingué voyelles et consonnes de façon à mieux cerner les parts respectives de ces deux types de phonèmes dans la configuration du tractus. La dimension *alveolar* est nettement plus sollicitée dans la production des consonnes pour les trois langues, ceci au détriment des autres dimensions (figure 4).

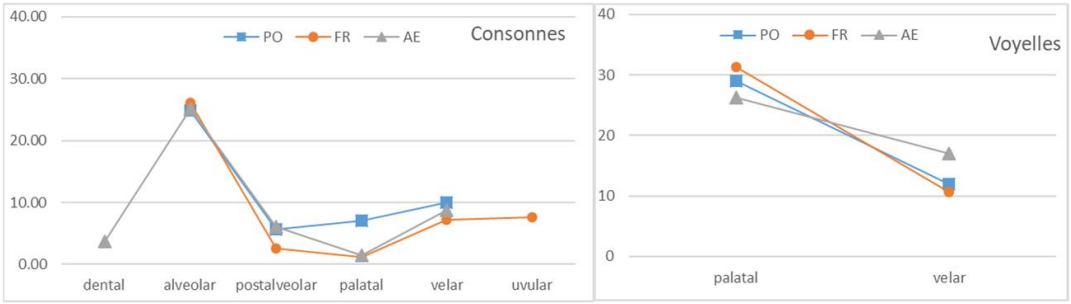


FIGURE 4 : Fréquence d'apparition des dimensions pour la variable TCL. Pour les consonnes (à gauche) et pour les voyelles (à droite)

Certes la dimension *alveolar* est également importante dans les inventaires phonologiques des trois langues, mais de façon variable : 26% pour PO, 19% pour FR et 15% pour AE avec, donc, une forte proportion pour PO. Mais si l'on regarde les phonèmes les plus fréquents (Table 1) dans chacune des trois langues, on trouve 1 alvéolaire en PO (/t/=6%), 4 alvéolaires en FR (/t/ /d/ /l/ /s/=23%) et 3 en AE (/n/ /s/ /t/=16%). La forte proportion des alvéolaires dans l'inventaire phonologique de PO est donc contrebalancée par sa faible représentation dans les phonèmes fréquents. On observe donc que les disparités observées dans les inventaires phonologiques semblent être *gommés* par l'usage des phonèmes dans la production courante. La dimension *alveolar* revêt probablement un caractère

spécifique car il s'agit du lieu d'articulation le plus fréquemment présent dans les inventaires phonologiques des langues du monde (position alvéodentale=15.3%, Vallée et al. 1999).

La dimension *palatal* est majoritaire dans la production des voyelles pour les trois langues (figure 4, droite). S'il est vrai que les voyelles antérieures sont majoritaires dans l'inventaire phonologique FR, ça n'est pas le cas pour PO ni pour AE. La préférence pour la dimension *palatal* dans ces deux langues est donc due à la fréquence des voyelles antérieures dans les corpus (table 1). Là encore on observe que les différences des inventaires phonologiques semblent être compensées par l'usage des sons de chaque langue tendant vers des positions articulatoires communes.

4 Discussion

Dans ce travail nous avons cherché à rendre compte de la configuration *théorique* du tractus vocal au cours de la production de parole spontanée et dans trois langues différentes. Les trois langues utilisées ici PO, FR et AE présentent des spécificités dans leur inventaire phonologique : un système vocalique riche et complexe en AE et réduit pour PO ; des voyelles antérieures plus nombreuses pour FR, idem pour l'arrondissement et la nasalité; un système consonantique riche en fricatives et affriquées pour AE et surtout PO. Malgré ces spécificités, on remarque une étonnante similitude concernant la fréquence des positions articulatoires pour les trois langues. Au cours de la production de parole spontanée, nous avons pu observer qu'en moyenne 78% des phonèmes sont produits avec la glotte fermée, 89% avec le voile du palais relevé et 13% avec un arrondissement des lèvres. Ces fréquences moyennes sont extrêmement proches pour chacune des langues. De même, la position de la langue lors de la constriction est majoritairement positionnée dans trois régions : palais dur, alvéoles puis palais mou. Là encore la fréquence des positions articulatoires est très proche dans les trois langues avec une prédominance de l'utilisation de la position alvéolaire dans la production des consonnes et de la position palatale pour les voyelles.

On retrouve donc une très forte similarité de la mobilisation du tractus vocal dans les trois langues et ce, malgré des différences notables dans la configuration des trois inventaires phonologiques. Ces tous premiers résultats tendent à montrer que, quelques soient les disparités et spécificités des systèmes phonologiques, la structure lexicale et l'usage du lexique dans le discours spontané tend à gommer ces différences, suggérant des contraintes biomécaniques et/ou phonologiques favorisant des configurations préférentielles du tractus vocal.

Nous devons toutefois indiquer que ce travail préliminaire présente plusieurs limites : 1/ nous avons utilisé des corpus de tailles très hétérogènes (notamment le corpus PO est probablement trop restreint pour représenter fidèlement les distributions de phonèmes en parole spontanée) ; 2/ le lien entre les inventaires phonologiques des trois langues et l'annotation phonétique des corpus a parfois été très difficile à établir et il faudrait probablement utiliser des corpus annotés de façon plus homogène ; 3/ il serait également nécessaire d'obtenir une description précise des réalisations articulatoires des langues de façon à attribuer correctement les dimensions ; 4/ il sera tout à fait nécessaire de poursuivre ce travail avec un nombre plus important de langues présentant des systèmes phonologiques plus diversifiés. Enfin, dans une étape ultérieure, nous envisageons de mettre en correspondance, au travers des corpus, ces positions attendues avec les réalisations effectives des segments phonétiques.

Références

ANDERSON, A., BADER, M., BARD, E., BOYLE., E., DOHERTY, G., GARROD, S., ISARD, S., KOWTKO, J., MCALLISTER, J., MILLER, J., SOTILLO, C., THOMPSON, H. S., AND WEINERT, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34:351–366. DOI: 10.1177/002383099103400404

BROWMAN, C. P., & GOLDSTEIN, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49 (3-4)*, 155–180.

FOUGERON, C., SMITH, C.L. (1999). French, on *The Handbook of the International Phonetic Association*, Cambridge University Press, 78-81.

GORISCH, J., ASTÉSANO, C., BARD, E. G., BIGI, B., & PRÉVOT, L. (2014). Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. *LREC Proceedings*, 5p. HAL: hal-02158880.

KLESSA, K., WAGNER, A., OLEŚKOWICZ-POPIEL, M., & KARPIŃSKI, M. (2013). Paralingua – A New Speech Corpus for the Studies of Paralinguistic Features. *Procedia - Social and Behavioral Sciences*, *95*, 48–58. DOI: 10.1016/j.sbspro.2013.10.621.

LADEFOGED, P. (1999). American English, on *The Handbook of the International Phonetic Association*, Cambridge University Press, 41-44.

LADEFOGED, P. & MADDIESON, I. (1996). *The Sounds of the World's Languages*, Wiley, 450 p.

MEUNIER, C., & ESPESSER, R. (2011). *Is Phoneme Inventory a Good Predictor for Vocal Tract Use in Casual Speech?* 1370–1373. HAL: hal-01514696

PITT, M. A., JOHNSON, K., HUME, E., KIESLING, S., & RAYMOND, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. DOI: 10.1016/j.specom.2004.09.001

VALLÉE, N., BOË, L.-J., & STEFANUTO, M. (1999). Typologies phonologiques et tendances universelles. Approche substantialiste. *Linx. Revue des linguistes de l'université Paris X Nanterre*, *11*, 31–54. DOI: 10.4000/linx.863.

YANG, B. (2012). Reduction and Frequency Analyses of Vowels and Consonants in the Buckeye Speech Corpus, *Phonetics and Speech Sciences* (말소리와 음성과학), Volume 4, Issue 3, p.75-83.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Interaction entre durée et position dans la perception des fricatives voisées chuchotées[1]

Yohann Meynadier, Noël Nguyen, Sophie Dufour
Aix Marseille Univ, CNRS, LPL, UMR7309, Aix-en-Provence, France
Aix Marseille Univ, ILCB, Marseille, France
{yohann.meynadier ; noel.nguyen-trong ; sophie.dufour}@univ-amu.fr

Résumé

Cette étude s'intéresse à la reconnaissance du trait de voisement en parole chuchotée. Nos travaux antérieurs (Dufour & Meynadier 2019) montrent une reconnaissance plus tardive du trait [+voisé] reposant sur un traitement probablement pré-lexical d'informations acoustiques, autres que la vibration laryngée, extraites du signal chuchoté et utilisées dans l'accès lexical. Via une tâche d'identification en perception catégorielle, cette étude révèle que la durée conditionne la perception du voisement des fricatives chuchotées : plus /ʃ/ est long, plus il est perçu [-voisé] ; plus /ʒ/ est bref, plus il est identifié [+voisé]. Cet effet de durée est modulé par le trait sous-jacent de voisement et la position dans le (non)mot. La fricative [+voisé] en position finale montre une frontière perceptive particulièrement décalée vers des durées beaucoup plus longues que les autres.

Abstract

Interplay between duration and word position in voicing perception of whispered fricatives
This study focuses on the recognition of the voicing-feature in whispered speech. Our previous works (Dufour & Meynadier 2019) show a later recognition of the feature [+voiced] probably based on a pre-lexical processing of acoustic information, other than the laryngeal vibration, extracted from the whispered signal and used in lexical access. Via an identification task in categorical perception, this study reveals that duration influences the perception of the voicing of whispered fricatives: the longer /ʃ/ is, the more it is perceived [-voiced]; the shorter /ʒ/ is, the more it is identified [+voiced]. This duration effect is modulated by the underlying voicing-feature and the position within the (non-sense) word. The fricative [+voiced] in word-final position shows a perceptual boundary particularly shifted towards durations much longer than the others.

Mots-clés : voisement, parole chuchotée, perception, fricative, français
Keywords: voicing, whispered speech, perception, fricative, French

1 Introduction

Le chuchotement crée une opacité du trait de voisement entre la forme phonétique et la représentation phonologique des mots en français. Cette opacité provient du fait qu'en parole chuchotée le trait de voisement ne peut plus être rattaché à la vibration des cordes vocales à un

[1] Une version résumée de ce travail a été acceptée à la 17è International Conference on Laboratory Phonology, 6-8 juillet 2020, Vancouver, Canada (LabPhon17).

Volume 1 : Journées d'Études sur la Parole, pages 442–450.
HAL : hal-02798571.

niveau phonétique, mais doit être associé à d'autres propriétés acoustiques secondaires pour être correctement reconnu.

Dans le versant de la perception, les travaux sur l'anglais rapportent que les obstruantes voisées chuchotées sont reconnues au-delà du seuil du hasard (Dannenbring 1980, Tartter 1989, Munro 1990, Gilichinskaya & Strange 2011). Pour le français, Dufour & Meynadier (2019) ont précisé le décours temporel de la reconnaissance des mots chuchotés à obstruantes voisées via un paradigme d'amorçage audiovisuel de décision lexicale. Les résultats montrent que la reconstruction du trait [+voisé] des obstruantes chuchotées prend plus de temps que la reconnaissance du trait [-voisé]. Sans être totalement homophones aux obstruantes sourdes, les obstruantes voisées chuchotées restent temporairement ambiguës. L'ambiguïté n'est levée que si un temps supplémentaire de traitement de 50 ms est accordé au sujet. L'hypothèse est que dans ce bref laps de temps, l'auditeur extrait du signal de parole chuchotée les informations acoustiques nécessaires à la reconnaissance de l'obstruante voisée.

Dans le versant de la production, des travaux récents sur le français montrent que les consonnes voisées chuchotées, bien que produites sans vibration des cordes vocales, conservent certaines traces phonétiques de leur identité sous-jacente [±voisé] (pour une revue Meynadier & Dufour 2016, 2018). Notamment, les durées acoustiques des consonnes et des voyelles préconsonantiques paraissent être les corrélats du voisement les plus robustes en parole chuchotée (Vercherand 2010, Meynadier & Gaydina 2013, Meynadier & Dufour 2018). Par ailleurs, en parole modale la manipulation de la durée des obstruantes influence l'identification de leur trait de voisement (Denes 1955). Reste qu'aucune étude, à notre connaissance, ne s'est intéressée à ce point en parole chuchotée.

Afin de clarifier l'opacité du contraste de voisement en parole chuchotée, trois hypothèses sont testées ici. Premièrement, sur la base des travaux précités, nous testons ici la possibilité que dans le bref laps de temps supplémentaire dans la reconnaissance des obstruantes voisées chuchotées, mis en évidence par Dufour & Meynadier (2019), le sujet est capable d'extraire de la parole les informations acoustiques relatives à la durée des consonnes afin d'identifier perceptivement leur trait de voisement. Nous nous attendons donc à ce que les résultats de perception soient en accord avec ceux de production rapportés par la littérature, à savoir que plus la consonne est longue, plus elle est associée au trait [-voisé], et inversement, plus elle est courte, plus elle est identifiée comme [+voisé]. Deuxièmement, nous cherchons à établir à quel niveau de traitement l'information de durée segmentale est recrutée pour recouvrer le voisement sous-jacent de la consonne. Suite à Dufour & Meynadier (2019), nous supposons qu'un traitement pré-lexical des indices acoustiques est impliqué dans la reconnaissance du 'voisement chuchoté'. Nous nous attendons donc à ce que l'effet de durée de la consonne soit similaire pour des mots et pour des non-mots. Enfin, nous testons ici également si un effet de position de la consonne dans le (non)mot influence l'identification de son voisement sous-jacent. Sur ce point, nous pensons que l'allongement spécifique des consonnes finales de (non)mots isolés (cf. Table 1), du fait de l'allongement final prosodique, pourrait décaler la frontière catégorielle d'identification du trait de voisement.

Pour répondre à ces questions, la durée des fricatives /ʃ/ et /ʒ/ en position initiale et finale de mots et de non-mots a été manipulée dans une tâche de perception catégorielle du trait [±voisé] de la fricative.

2 Protocole expérimental

Un test d'identification portant sur les fricatives /ʃ/ versus /ʒ/ a été mené pour évaluer si la durée de la consonne permet de faire basculer la perception catégorielle du voisement de la consonne chuchotée.

2.1 Matériel linguistique

Les fricatives ont été choisies pour représenter le contraste de voisement afin de simplifier la manipulation de la durée acoustique de la consonne. En effet, les plosives, présentant deux phases articulatoires distinctes, occlusion et explosion, obligeraient à un contrôle de leur durée plus complexe. Par ailleurs, les modifications de durée de l'occlusion silencieuse des plosives chuchotées ne sont pas perceptibles en initiale absolue de mot. Seul le contraste /ʃ/-/ʒ/ est ici utilisé.

/ʃ/-/ʒ/ ont été produites par un locuteur français dans des séquences CVC formant soit une paire minimale, en position initiale de mot : « *ch*ar » vs « *j*arre » et en position finale de mot : « ca*ch*e » vs « cage », soit une paire de non-mots se distinguant seulement par le voisement de la fricative, en initiale : « *ch*eur » vs « *j*eur » et en finale « queu*ch*e » vs « queuge ». Ces quatre paires ont été enregistrées en chambre sourde à l'aide d'un microphone Senheiser HD415 (échantillonnage : 32 kHz ; 32 bits). Une distance de 30 cm a été maintenue constante entre le micro et la bouche du locuteur qui a produit naturellement les mots en voix chuchotée non projetée ou forcée (comme chuchotés à l'oreille d'un interlocuteur à distance proximale). L'absence de vibrations glottiques du signal chuchoté, ainsi que la différence de durée attendue (pour des références en voix chuchotée et modale, Dufour & Meynadier 2019) entre les consonnes cibles et entre les voyelles préconsonantiques selon le voisement de la fricative (Table 1) ont été vérifiées via Praat (Boersman & Weenink 2017).

	C	V	C		C	V	C
*ch*ar	138	-	-	*ch*eur	154	-	-
*j*arre	106	-	-	*j*eur	111	-	-
c*ach*e	-	185	253	qu*euch*e	-	199	243
c*ag*e	-	240	161	qu*eug*e	-	230	156

TABLE 1 : Durée (ms) des fricatives cibles et des voyelles préconsonantiques des mots et non-mots utilisés pour la génération des stimuli de l'expérience

À l'aide du Vocal Toolkit (Corretge 2019) de Praat, la durée acoustique de la fricative a été modifiée par resynthèse selon un continuum de 11 pas de 20 ms allant de 50 à 250 ms. L'intensité globale (RMS) de chaque fricative cible a été neutralisée à la valeur moyenne calculée entre celle de la [+voisé] et celle de la [-voisé] de la paire de (non)mots concernée. Les portions restantes _CV (_ar, _eur) et CV_ (ca_, queu_) ont également été neutralisées en durée, en intensité et en fréquence. Les portions de chaque paire [±voisé] de (non)mots ont dans un premier temps été resynthétisées avec une durée neutre correspondant à la durée moyenne entre les deux membres de la paire. Un morphing acoustique (spectral) à mi-chemin entre ces deux mêmes exemplaires a été appliqué (via la fonction *Mix…* du Vocal Toolkit de Praat). Enfin, l'intensité a été homogénéisée de la même manière que pour les fricatives cibles. Les 4 paires x 2 (non)mots x 11 pas de durée, soit 88 stimuli, ont été contrôlés afin que seule la différence de durée de la fricative varie. À noter que le profil spectral (autre que l'intensité globale) de la fricative cible n'a pas été neutralisé par morphing, et que donc il est susceptible de différences selon le voisement sous-jacent de la consonne. Néanmoins, à notre connaissance, aucune différence spectrale systématique due au voisement de la consonne n'a encore été mis en évidence dans la littérature (Meynadier & Dufour 2018).

Enfin, les stimuli ont été reconstruits par concaténation de la fricative cible et de la portion restante manipulées du (non)mot. Afin de lisser la transition entre ces deux parties du stimulus, un amortissement initial et/ou final (fonction *Fade…* du Vocal Toolkit de Praat) sur 30 ms a été appliqué sur le signal. Ces deux parties ont alors été concaténées avec un chevauchement de même durée.

2.2 Expériences

Ainsi, 8 continua de durée consonantique, soit 2 (mots + non-mots) en paire de voisement x 2 positions x 11 pas de durée de la fricative, ont été soumis à une tâche d'identification dans un test de perception catégorielle. 32 sujets (21 F, 11 H) ont passé l'expérience sur les mots, 32 autres (26 F, 6 H) sur les non-mots. Chaque sujet a répondu sur les paires à fricative en position initiale (*ch*ar vs *j*arre ou *ch*eur vs *j*eur) et en position finale (ca*ch*e vs cage ou queu*ch*e vs queuge) dans deux sessions séparées par une pause ; l'ordre de ces deux sessions a été balancé entre les sujets. À l'intérieur de chaque session, les stimuli à fricative [+voisé] et [-voisé] ont été aléatoirement ordonnés différemment pour chaque sujet. Les tests ont été implémentés sur ordinateur via le logiciel Perceval (André et al. 2003).

Les 64 sujets de l'expérience ont été rémunérés 10 € pour un temps de passation d'environ 30 min. Aucun n'a rapporté de troubles de l'audition, de la parole ou du langage, ni de troubles attentionnels. Ils ont passé le test individuellement dans un box expérimental calme. Les stimuli ont été entendus au moyen d'un casque audio circum-aural Superflux HD681B et les réponses saisies sur un boitier à 2 boutons. La position des (non)mots à fricative [+voisé] et [-voisé] était indiquée orthographiquement sur le boitier et demeuraient constante lors de la session. Elle était inversée entre les deux sessions du test et balancée entre les sujets de l'expérience. Le sujet était face à un écran sur lequel il a lu les consignes et où apparaissait en son milieu un point de fixation durant les 750 ms précédant l'écoute du stimulus, afin de maintenir son attention. Un silence de 1,5 sec. séparait chaque stimulus. Le niveau d'écoute a été maintenu à un volume fixe, confortable et représentatif de la parole chuchotée durant toute l'expérience et entre les sujets. La tâche du sujet consistait à identifier le (non)mot entendu en appuyant sur le bouton correspondant du boitier-réponse immédiatement après la fin acoustique du stimulus, sans pression ou limite de temps. Chaque stimulus a été présenté 10 fois dans la même session. Chaque sujet a donc pris sa décision de voisement sur 220 stimuli pour chacune des deux sessions du test.

3 Résultats

Nous rapportons ici les résultats statistiques obtenus sur le pourcentage moyen de réponses correctes aux tests. Une réponse est dite 'correcte' quand l'auditeur a choisi le (non)mot comportant la fricative dont le trait sous-jacent de voisement correspond bien à celui de la fricative du stimulus entendu. Dans un premier temps, sont présentés les statistiques et les résultats relatifs à l'effet du continuum de durée de la fricative sur le pourcentage d'identification de son voisement. Ensuite, les analyses et les résultats statistiques concernant les différences en durée de frontière catégorielle d'identification du voisement sont présentés afin de rendre compte des interactions de la durée de la fricative avec son voisement sous-jacent, sa position dans le (non)mot et la lexicalité du stimulus.

3.1 Effet de la durée

Les réponses des participants (1 = réponse correcte, 0 = réponse incorrecte) ont été analysées grâce à un modèle de régression *logit* à effets mixtes (Baayen 2008 ; Baayen et al., 2008). Le modèle incluait la durée des consonnes (11 niveaux), le voisement ([+voisé], [-voisé]), la position (initiale, finale) et la lexicalité (mot, non-mot), ainsi que leurs interactions, comme effets fixes. Le modèle incluait également l'intercept et la pente aléatoires par participants pour l'effet de la durée des consonnes, du voisement et de la position (Barr et al., 2013). L'ensemble de données étant trop volumineux pour les paramètres par défaut de *glmer.nb*, l'option *nAGQ=0* a été implémentée afin de permettre au modèle de converger (Bates et al., 2016). Le modèle a été appliqué sur 28159 données (1 réponse n'a pas été enregistrée par le logiciel).

De façon intéressante le modèle a révélé une interaction significative entre la durée des fricatives et leur voisement [χ^2= 5341,61 ; p < .0001]. Plus la fricative /ʃ/ est longue, plus elle est perçue [-voisé], et inversement plus la fricative /ʒ/ est longue, moins elle est perçue [+voisé] (autrement dit, plus /ʒ/ est brève, mieux elle est identifiée comme [+voisé]). La Figure 1 montre que cet effet est vrai que la fricative soit produite en position initiale ou finale de mot ou de non-mot.

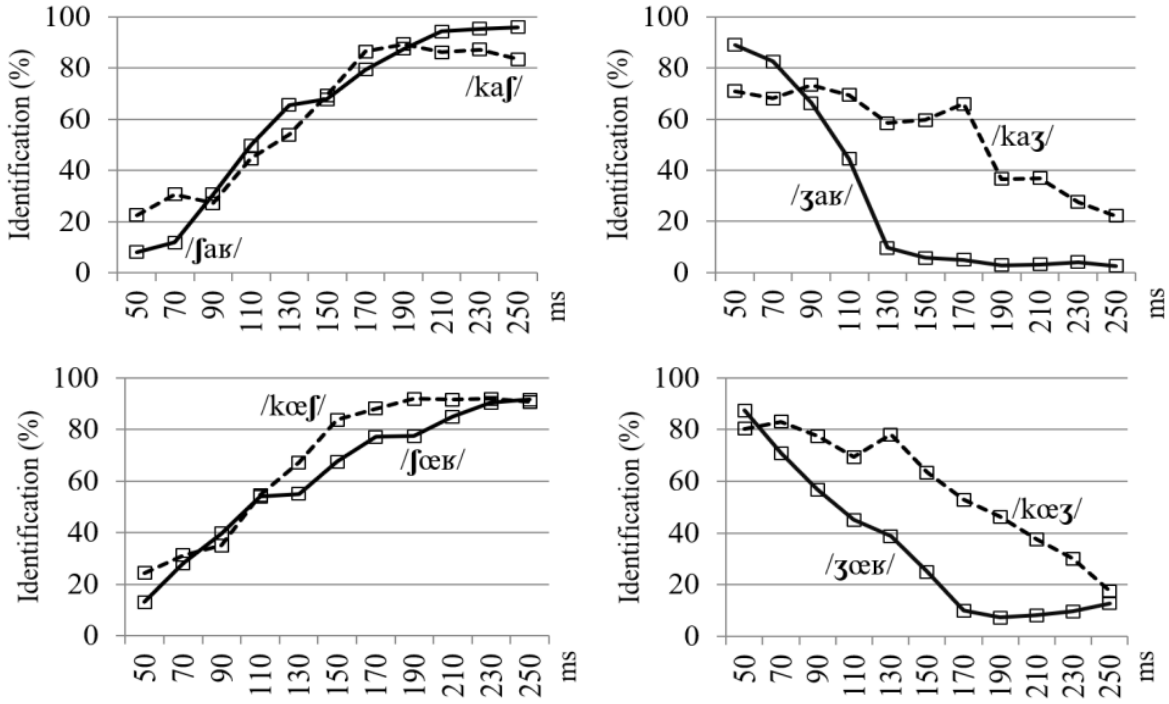


FIGURE 1 : Pourcentages moyens d'identification du trait de voisement en fonction de la durée de chaque fricative chuchotée selon sa position dans un mot (haut) ou un non-mot (bas)

3.2 Effet du voisement, de la position et de la lexicalité

Pour comprendre comment le trait de voisement, la position dans le (non)mot et la lexicalité du stimulus influencent l'identification du trait de voisement selon la durée de la fricative chuchotée, nous avons étudié les différences entre les frontières de perception catégorielle des différentes conditions. L'analyse statistique des interactions entre la durée de la fricative et les facteurs *voisement*, *position* et *lexicalité* a été effectuée sur les valeurs de durée des frontières perceptives catégorielles du trait [±voisé] des fricatives cibles. Pour chaque sujet (N = 64) dans chaque condition, à savoir pour chaque fricative (/ʃ/, /ʒ/) dans chaque position (initiale, finale) de chaque (non)mot (*ch*ar, *j*arre, ca*ch*e, cage, *ch*eur, *j*eur, queu*ch*e, queuge), la frontière catégorielle a été calculée à partir de la fonction spline cubique des pourcentages d'identification du voisement (à savoir les pourcentages moyens de réponses correctes) selon les 11 pas du continuum de durée de la fricative cible. Cette frontière correspond au point de la courbe où la perception du trait [±voisé] de la fricative bascule. Elle est déterminée par la détection de la durée en ms (x) qui conduit à l'intercept (y) entre la courbe interpolée et un taux de réponse correcte à 50%, à savoir à 50% d'identification (c'est-à-dire le seuil de hasard de réponses binaires). 11 sujets (6 pour les mots et 4 pour les non-mots) ont été exclus des analyses. Cette exclusion correspond à une absence de bascule catégorielle dans la perception du voisement de la fricative en fonction de sa durée dans au moins une des conditions du test. Soit le sujet à un comportement trop incohérent, soit ses pourcentages d'identification sont toujours inférieurs ou supérieurs au seuil de 50% quelle que soit la durée de la

fricative cible. Ainsi, 17,2% des données de frontière (soit 44 valeurs sur 256, à savoir 32 sujets x 8 (non)mots) ont été exclues.

Les valeurs moyennes de frontière (durée en ms) ont été analysées via une ANOVA à trois facteurs dont deux à mesures répétées (*voisement*, *position*) et un à mesures indépendantes (*lexicalité*). La triple interaction *voisement*position*lexicalité* est significative [F(1, 54) = 4,32) ; p < .05]. Afin d'étudier cette interaction, nous rapportons les effet simples de ces facteurs et les interaction à deux facteurs, ainsi que les comparaisons multiples deux à deux a posteriori entre les conditions expérimentales avec une correction de Bonferroni. Les comparaisons d'intérêt entre *voisement*, *position* et *lexicalité* sont rapportés dans la Table 2.

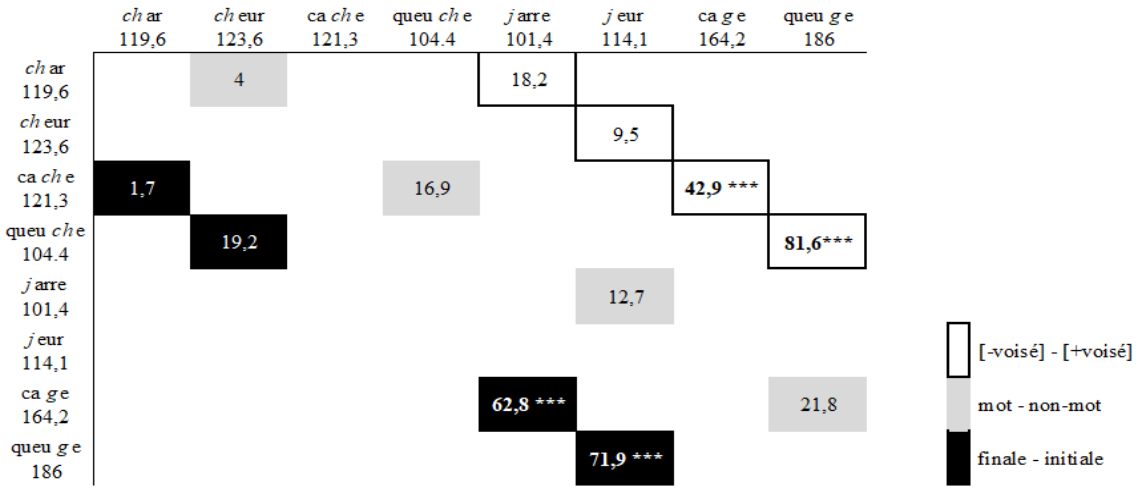


TABLE 2 : Valeurs (sous le nom du stimuli) et écarts de durée (en ms) des frontières de perception catégorielle du trait de voisement des fricatives chuchotées selon le voisement, la position dans le (non)mot et la lexicalité du stimulus. *** signale une différence significative avec p < .0001 ; l'absence d'astérisque indique une différence non significative (p > .05)

Toutes conditions confondues, seuls le *voisement* et la *position* de la fricative montrent un effet significatif sur les valeurs de durée des frontières catégorielles, respectivement [F(1, 54) = 52,10 ; p < .0001] et [F(1, 54) = 34,87 ; p < .0001]. Ces deux facteurs ont également une interaction significative [F(1, 54) = 117,60 ; p < .0001]. La Figure 2 montre ainsi que la frontière catégorielle de la fricative [+voisé] finale est spécifiquement plus haute que les autres. Elle atteint 164 ms pour le mot « cage » et 186 ms pour le non-mot « queuge », alors que toutes les autres se situent globalement autour de 100 à 120 ms. La Table 2 résume les tests post-hoc de ces facteurs. Elles montrent que seule la frontière catégorielle de la fricative [+voisé] finale se distingue significativement de celle de la fricative [-voisé] finale et de celle de la fricative [+voisé] initiale, pour les mots (respectivement [F(1, 52) = 38,01 ; p < .0001] et [F(1, 52) = 48,19 ; p < .0001]) comme pour les non-mots (respectivement [F(1, 52) = 144,46 ; p < .0001] et [F(1, 52) = 66,26 ; p < .0001]). Le déplacement vers une valeur plus élevée de la frontière perceptive de la fricative [+voisé] finale par rapport aux autres oscille entre 43 et 82 ms.

Enfin, une interaction significative entre le *voisement* et la *lexicalité* est également rapportée [F(1, 54) = 12,201 ; p < .001]. Elle indique que toutes positions confondues seules les frontières catégorielles de la fricative [+voisé] sont un peu plus hautes pour les non-mots (queuge, *j*eur) que pour les mots (cage, *j*arre). Cela est illustré par la Figure 2. L'écart de frontière est en moyenne de 22 ms entre « cage » et « queuge », et de 13 ms entre « *j*arre » et « *j*eur » (Table 2). Pour la fricative [-voisé], nous n'observons pas de régularité de ce point de vue. Reste que les tests post-hoc,

rapportés dans la Table 2, ne signalent aucun de ces écarts de frontière entre mot et non-mot comme étant significatif.

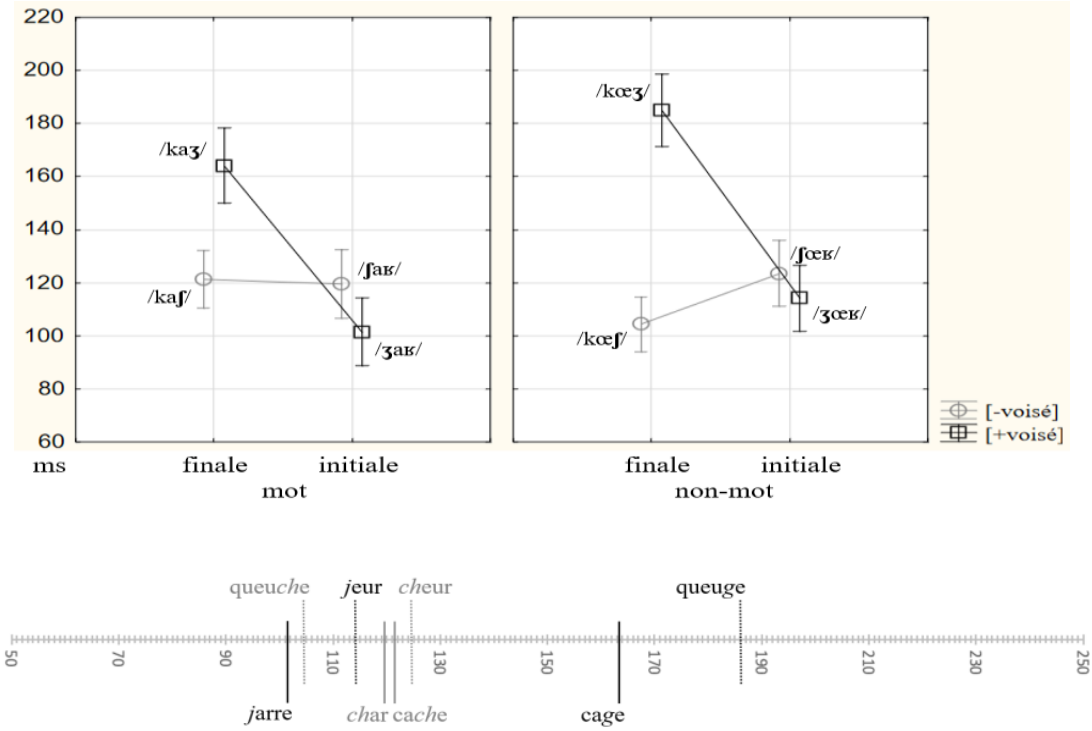


FIGURE 2 : Valeurs moyennes (en haut) et position sur le continuum (en bas) des durée des frontières de perception catégorielle du trait de voisement des fricatives chuchotées selon le voisement et la position pour les mots et les non-mots. Sur le graphique du haut, les barres représentent les intervalles de confiance à 95%

4 Discussion

Dans tous les cas, des durées plus longues favorisent toujours la perception du trait [-voisé], et plus courtes celle du trait [+voisé] (Figure 1). Les courbes d'identification moyenne du voisement montrent une influence plutôt graduelle que catégorielle. Elles ont également une forme particulière qui pourrait être une trace du biais perceptif en faveur du trait [-voisé] des obstruantes chuchotées noté dans la littérature (cf. Dufour & Meynadier 2019). En effet, elles dessinent un plateau relativement stable d'une perception en fricative [-voisé] au moins à partir de 170 ms, pour /ʃ/ mais également pour /ʒ/, sauf si ce dernier est final de (non)mot. De plus, toute fricative chuchotée, autre que /ʒ/ en position finale, à partir d'une durée de 110 ms, est majoritairement perçue en /ʃ/, et en deçà en /ʒ/. Les résultats montrent donc que la durée consonantique module la perception du trait de voisement des fricatives chuchotées. La reconnaissance de leur voisement sous-jacent s'appuierait bien sur le corrélat phonétique de durée consonantique tel qu'observé en production de la parole, à savoir sur la durée plus longue des obstruantes [-voisé] par rapport aux [+voisé]. L'hypothèse que le sujet est capable d'extraire de la parole chuchotée les informations acoustiques relatives à la durée des consonnes afin d'identifier perceptivement leur trait de voisement est donc bien confirmée. Cela plaide pour une reconstruction du trait à partir du traitement de détails phonétiques fins présents dans le signal acoustique.

Pour aller, plus loin nous avons aussi tester l'effet de la lexicalité des stimuli. Les analyses des frontières perceptives du trait de voisement selon la durée consonantique n'ont pas révélé de différences significatives entre les fricatives insérées dans des mots et celles insérées dans de non-

mot (Table 2). Ce résultat irait dans le sens d'une extraction du trait de voisement à partir des corrélats de durée segmentale présents dans le signal de parole, et donc de l'implication d'un traitement pré-lexical de ces indices acoustiques lors de la reconnaissance d'un trait, d'un phonème et/ou d'un mot dont la forme phonétique est altérée ou plus éloignée de sa forme phonologique.

Enfin, de manière intéressante, cette étude montre, pour la première fois, que l'influence de la durée consonantique n'est pas homogène, mais est fonction de l'interaction entre le trait de voisement de la fricative et sa position dans le (non)mot. En effet, la Figure 1 montre que pour une même durée et au moins à partir de 90 ms, la fricative [+voisé] est toujours mieux reconnue en finale qu'en initiale de (non)mot. Ainsi, seule la fricative [+voisé], et seulement si elle se trouve en position finale de (non)mot, a une frontière perceptive du voisement selon la durée significativement plus haute que celle des fricatives dans toute autre condition. Sa frontière apparaît donc décalée vers une durée consonantique beaucoup plus longue, de 164 (pour « cage ») à 186 ms (pour « queuge »), par rapport aux autres situées autour de 100 à 120 ms. Ainsi, par rapport une position en initiale, une fricative [+voisé] finale nécessite un allongement de 60-70 ms (Table 2) pour erronément être perçue [-voisé]. La perception du trait [+voisé] de la fricative chuchotée résisterait spécifiquement mieux à un allongement en position finale de mot.

Cette asymétrie de l'effet de durée en fonction de la position dans le mot pourrait s'expliquer par différents mécanismes, qui restent spéculatifs à ce stade de notre étude. En effet, le profil spectral de fricatives [+voisé] et [-voisé] de nos stimuli n'ont pas été neutralisé (par exemple par un morphing entre les deux spectres). La possibilité de différences spectrales corrélatives au voisement sous-jacent de la fricative présentes dans le signal et utiles à l'auditeur n'est donc pas exclue. Cette possibilité reste cependant une question ouverte sachant qu'à notre connaissance, aucune étude spectrale en parole n'a pu véritablement attester de différence systématiques liées au voisement dans le spectre de bruit des obstruantes chuchotées (cf. Meynadier & Dufour 2018). Si ces traces existent et sont exploitées par l'auditeur, il reste toujours à expliquer pourquoi seule la fricative [+voisé] et seulement en position finale montre une résistance spécifique au changement de durée, à savoir l'allongement. Une piste pourrait s'appuyer sur l'influence bien connue de la structure prosodique (dont syllabique et lexicale) sur la réalisation phonétique des phonèmes (connue sous la notion de *phonetic encoding of prosodic structure*, par exemple Cho 2016). Les segments montrent des variations acoustiques selon leur position et le poids de la frontière prosodique. L'une de ces variations concerne l'allongement en fin de constituant réalisé par une décélération du tempo de parole. Les segments finaux sont ainsi systématiquement plus longs en position finale qu'en position initiale, comme l'illustre, par exemple, les durées segmentales de nos stimuli présentées dans la Table 1. Or, des études récentes mettent en évidence que la perception d'une frontière prosodique, marquée par un allongement final, module la catégorisation des phonèmes en frontière (Steffman 2019, Mitterer et al. 2016 ; pour une revue et une discussion). Dans la lignée de ces travaux, on pourrait donc supposer ici que l'auditeur recrute ses connaissances prosodiques lors de la reconnaissance de traits segmentaux qui lui permettent d'opérer une compensation perceptive de l'allongement subi par la fricative [+voisé] chuchotée en position finale. Mais alors, reste à savoir pourquoi seule la fricative [+voisé], et non la [-voisé], montre cet effet de position ? L'absence d'effet de la position finale pour la fricative [-voisé] chuchotée pourrait s'expliquer par le fait qu'une fricative longue serait par défaut interprétée comme [-voisé] et que le biais perceptif vers le trait [-voisé] (relevé dans la littérature) jouerait pour une reconnaissance immédiate des fricatives [-voisé] allongées en finale de mot. Leur reconnaissance ne nécessiterait donc pas de traitement particulier dans cette position. Inversement, la reconnaissance du trait [+voisé] chuchoté engage un traitement spécifique et plus coûteux des détails phonétiques corrélatifs du voisement (Dufour & Meynadier 2019). La durée faisant partie de ces traces acoustiques du voisement, en position finale, du fait de l'allongement final prosodique, la fricative [+voisé] chuchotée serait potentiellement plus ambigüe et nécessiterait un traitement supplémentaire impliquant le recrutement d'informations prosodiques (top-down), permettant la compensation perceptive relative à sa durée segmentale.

Remerciements

Financement de l'ANR-16-CONV-0002 (ILCB), ANR-11-LABX_0036 (BLRI) et A*MIDEX. Et à Thierry Legou (LPL) pour son aide technique dans l'analyse des données.

Références

ANDRÉ C., GHIO A., CAVÉ C., TESTON B. (2003). PERCEVAL: a computer-driven system for experimentation on auditory and visual perception. *Proceedings of the 15th ICPhS*, 1421-1424.

BAAYEN R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

BAAYEN R. H., DAVIDSON D. J., BATES, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.

BARR D. J., LEVY R., SCHEEPERS C., TILY H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.

BATES D., BOLKER B., WALKER, S. (2016). Package 'lme4', version 1.1- 12, linear mixed effects models using 'Eigen' and S4. https://cran.r-project.org/web/packages/lme4/lme4.pdf.

BOERSMA P., WEENINK D. (2019). *Praat: doing phonetics by computer*. Version 6.0.52. http://www.fon.hum.ua.nl/praat.

CHO T. (2016). *Prosodic* boundary strengthening in the phonetics–Prosody interface. *Lang Ling Compass*, 10(3), 120-141.

CORRETGE R. (2019). Praat Vocal Toolkit. http://www.praatvocaltoolkit.com.

DANNENBRING G. L. (1980). Perceptual discrimination of whispered phoneme pairs. *Percept Motor Skill*, 51(3), 979-985.

DENES P. (1955). Effect of duration on the perception of voicing. *JASA*, 27(4), 761-764.

DUFOUR S., MEYNADIER Y. (2019). Temporary ambiguity in whispered word recognition: a semantic priming study. *J Cogn Psychol*, 31(2): 157-174. DOI: 10.1080/20445911.2019.1573243.

GILICHINSKAYA Y., STRANGE W. (2011). Perception of final consonant "voicing" in whispered speech. *JASA*, 129, 2420.

MEYNADIER Y., DUFOUR S. (2016). Accès lexical et reconnaissance du voisement en voix chuchotée. *Actes des XXXIè JEP,* 19-27. Paris.

MEYNADIER Y., DUFOUR S. (2018). Ambiguïté temporaire des obstruantes voisées en parole chuchotée. *Actes des XXXIIè JEP*, 125-133. Aix-en-Provence.

MEYNADIER Y., GAYDINA Y. (2013). Aerodynamic and durational cues of phonological voicing in whisper. *Proceedings of the14th Interspeech*, 335-339, Lyon.

MITTERER H., CHO T., KIM, S. (2016). How does prosody influence speech categorization? *J Phon*, 54, 68–79.

MUNRO M. J. (1990). Perception of "voicing" in whispered stops. *Phonetica*, 47, 173–181.

STEFFMAN J. (2019). Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing. *JASA*, 154, EL560.

TARTTER V. C. (1989). What's in a whisper? *JASA*, 86, 1678–1683.

VERCHERAND G. (2010). *Production et perception de la parole chuchotée en français : analyse segmentale et prosodique*. Thèse de doctorat. Université Paris VII.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo)

Alexis Michaud[1] Oliver Adams[2] Séverine Guillaume[1] Guillaume Wisniewski[3]

(1) Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle – INALCO, 7 rue Guy Môquet, 94800 Villejuif, France
(2) Miner & Kasch, 8174 Lark Brown Rd #101, Elkridge, MD 21075, Etats-Unis d'Amérique
(3) Laboratoire de Linguistique Formelle (LLF), CNRS – Université Paris-Diderot, Case 7031, 5 rue Thomas Mann, 75013 Paris, France

alexis.michaud@cnrs.fr, oliver.adams@gmail.com, severine.guillaume@cnrs.fr, guillaume.wisniewski@linguist.univ-paris-diderot.fr

RÉSUMÉ

Les systèmes de reconnaissance automatique de la parole atteignent désormais des degrés de précision élevés sur la base d'un corpus d'entraînement limité à deux ou trois heures d'enregistrements transcrits (pour un système mono-locuteur). Au-delà de l'intérêt pratique que présentent ces avancées technologiques pour les tâches de documentation de langues rares et en danger, se pose la question de leur apport pour la réflexion du phonéticien/phonologue. En effet, le modèle acoustique prend en entrée des transcriptions qui reposent sur un ensemble d'hypothèses plus ou moins explicites. Le modèle acoustique, décalqué (par des méthodes statistiques) de l'écrit du linguiste, peut-il être interrogé par ce dernier, en un jeu de miroir ? Notre étude s'appuie sur des exemples d'une langue « rare » de la famille sino-tibétaine, le na (mosuo), pour illustrer la façon dont l'analyse d'erreurs permet une confrontation renouvelée avec le signal acoustique.

ABSTRACT

Analyzing errors in automatic phonemic transcriptions of the Na (Mosuo) language (Sino-Tibetan family)

Automatic phonemic transcription tools now reach high levels of accuracy on a single speaker with relatively small amounts of training data: on the order two to three hours of transcribed speech. Beyond its practical usefulness for language documentation, use of automatic transcription also yields some insights for phoneticians/phonologists. Acoustic models are built on the basis of the linguist's transcriptions, and thus encapsulate linguistic hypotheses and assumptions. To what extent can the acoustic model be examined by the linguist? The present report explores this topic by going into qualitative error analysis on Yongning Na (Sino-Tibetan). Among other benefits, error analysis allows for a renewed exploration of phonetic detail: examining the output of phonemic transcription software compared with spectrographic and aural evidence.

MOTS-CLÉS : transcription phonologique, reconnaissance de la parole, apprentissage machine, analyse d'erreurs, interdisciplinarité, documentation linguistique assistée par ordinateur.

KEYWORDS: phonological transcription, speech recognition, machine learning, error analysis, interdisciplinarity, Computational Language Documentation.

Volume 1 : Journées d'Études sur la Parole, pages 451–462.
HAL : hal-02798572.

1 Introduction[1]

1.1 Phonétique et reconnaissance automatique de la parole

La reconnaissance vocale a connu d'importants progrès au cours des deux dernières décennies, mais les collaborations entre informaticiens et linguistes ont été moins intenses qu'on ne pourrait le souhaiter. Les gains de performance ont été principalement obtenus en tirant parti d'une puissance de calcul sans cesse accrue, ainsi que de nouveaux outils statistiques, dits d'*intelligence artificielle*. Dans ce contexte, il ne paraît pas inutile de souligner que le dialogue interdisciplinaire demeure aussi pertinent que jamais à l'ère de l'apprentissage machine. Informaticiens et linguistes ont tout intérêt à collaborer afin de concevoir et déployer des outils innovants, et d'en tirer parti pour la recherche (Adda et al., 2016; Neubig et al., 2020). En effet, les collaborations entre linguistes et spécialistes du Traitement Automatique des Langues Naturelles ne sont pas seulement utiles à une meilleure efficacité pratique des outils (l'apprentissage statistique supervisé, qui repose sur une annotation raisonnée, donne de meilleurs résultats que l'apprentissage non supervisé : Jimerson & Prud'hommeaux, 2018; Wu et al., 2018) : elles sont en outre prometteuses pour les sciences de la parole. Historiquement, des collaborations entre linguistes et chercheurs en intelligence artificielle se sont attachées à mieux cerner la nature du phonème (question-clef de la phonétique/phonologie : voir notamment Jones, 1950) et à examiner la possibilité de le représenter par des modèles acoustiques statistiques. Entre autres et nombreux travaux, on citera les études qui s'appuient sur deux grands corpus d'anglais américain, « Texas Instruments/Massachusetts Institute of Technology », TIMIT (Garofolo et al., 1993) et SWITCHBOARD (Godfrey et al., 1992), pour mener une analyse des erreurs de reconnaissance automatique par des systèmes utilisant des modèles acoustiques statistiques. Le premier constat est celui d'une distance considérable entre les réalisations canoniques des phonèmes et la parole spontanée : « De nombreux mots sont articulés d'une manière qui omet ou transforme profondément les propriétés phonétiques des phonèmes qui les constituent, ce qui entraîne une grande variabilité dans la prononciation d'un même mot. Souvent, on ne trouve d'un segment qu'un indice ténu, et ce bien que le signal soit tout à fait intelligible »[2] (Greenberg et al., 1996, p. 24). La présente communication se veut une (modeste) poursuite de ces réflexions. Elle repose sur l'utilisation d'un outil de transcription phonémique automatique, le logiciel `Persephone`. Le développement de cet outil est appelé à déboucher, à moyen terme (d'ici quelques années), sur un logiciel complet de reconnaissance de la parole, qui reconnaisse des mots entiers, puis des phrases entières (à titre d'exemple, voir les travaux de Hjortnaes et al., 2020 concernant le komi, langue ouralienne). Le stade actuel, celui d'une transcription phonémique, présente l'intérêt de demeurer au plus proche du signal acoustique, de sorte que l'analyse d'erreurs permet une confrontation détaillée avec le signal acoustique.

---

1 Le présent exposé reprend des résultats présentés dans une communication (en anglais) au XIX[e] Congrès des sciences phonétiques (Michaud et al., 2019). Cette communication portait non seulement sur la langue na, mais aussi sur le tsuut'ina, langue de la famille dene (athabasque), parlée dans l'ouest du Canada. Faute de place, les résultats sur les données tsuut'ina ne peuvent être exposés ici. Le lecteur intéressé est renvoyé au texte anglais, ainsi qu'à la présentation en vidéo du *plugin* créé par Christopher Cox pour faciliter l'emploi de `Persephone` par ses collaborateurs tsuut'ina (https://www.youtube.com/watch?v=-pDOEqRpZKs).

2 *Texte original :* Many words are articulated in such a fashion as to either omit or significantly transform the phonetic properties of phonemic constituents, thus resulting in wide variation of word pronunciations. Often, only the barest hint of a segment is realized phonetically, in spite of good intelligibility.

1.2 Genèse de l'outil `Persephone` et perspectives de développement

L'utilisation d'outils de transcription automatique constitue un enjeu considérable pour la documentation linguistique, dans un contexte d'urgence : il s'agit d'accélérer le travail de collecte et de description d'une diversité linguistique mondiale en déclin rapide (Littell et al., 2018; Thieberger, 2017; van Esch et al., 2019). L'outil logiciel `Persephone`, disponible en ligne (https://github.com/persephone-tools/persephone) sous licence libre, est issu de recherches exploratoires menées par Oliver Adams au fil de son travail de thèse (Adams, 2017). La façon dont notre collaboration s'est nouée est détaillée dans un article paru dans une revue spécialisée dans les questions de documentation et conservation de langues en danger (Michaud et al., 2018). Deux communications à des colloques de Traitement Automatique des Langues exposent le fonctionnement de l'outil (Adams et al., 2017, 2018).

Au plan du développement logiciel, notre projet actuel (2020-2023) consiste à contribuer à l'élaboration d'une interface utilisateur unique (Foley et al., 2019) qui permette au linguiste « de terrain » d'appliquer à ses données (corpus d'entraînement composé de fichiers audio transcrits, et corpus d'application composé de fichiers audio non transcrits) toute une gamme d'outils : `Persephone`, mais aussi `wav2letter++` (Pratap et al., 2018), `KALDI`, `ESPnet` (Watanabe et al., 2018)... En effet, au vu de l'ampleur des différences que présentent entre elles les langues naturelles au plan phonético-phonologique comme à d'autres niveaux (morphosyntaxe, structure de l'information...), il paraît vraisemblable que des outils logiciels différents soient plus ou moins performants selon la langue et le type de corpus : certains donneront de meilleurs résultats que d'autres pour le traitement des tons ou de l'accent, par exemple. Dans le présent travail, l'accent n'est pas mis sur ces questions, mais sur les possibilités qu'offre la transcription automatique pour la recherche phonétique.

2 Méthode

2.1 Corpus employé

Les données employées, intégralement disponible en ligne, sont celles d'un corpus réuni au fil d'enquêtes linguistiques sur le terrain (Michaud et al., 2012) au sujet du na, langue sino-tibétaine parlée à la frontière entre les provinces chinoises du Yunnan et du Sichuan (Lidz, 2010). La phonotactique de la langue na est relativement simple : chaque syllabe comporte une consonne initiale (l'inventaire consonantique est présenté dans le tableau 1), l'une des voyelles (rimes) suivantes : /i e æ ɑ ɯ u ɤ o ɣ ɻ wæ wɑ wɤ jæ jɤ jo/ (pour plus de précisions : Michaud, 2008, 2017, pp. 447–486), et un ton. Une caractéristique saillante de la langue na est le rôle de premier plan qu'y jouent les tons : rôle morpho-phonologique aussi bien que lexical (Michaud, 2017). Il existe cinq tons en « phonologie de surface » : Haut, Moyen, Bas, Bas-montant, et Moyen-montant (noté ˥, ˧, ˩, ˧˥, ˥˧).

	bilabiales	dentales	alvéolo-palatales	rétroflexes	vélaires	uvulaires	glottales
occlusives	pʰ p b	tʰ t d		ʈʰ ʈ ɖ	kʰ k g	qʰ q	ʔ
affriquées		tsʰ ts dz	tɕʰ tɕ dʑ	ʈʂʰ ʈʂ ɖʐ			
nasales	m	n	ɲ	ɳ	ŋ		
fricatives		s z	ɕ ʑ	ʂ ʐ		ʁ	h
latérales		ɬ l					
approximante				ɻ			

TABLEAU 1 : Les consonnes du na de Yongning.

2.2 Les algorithmes : principes de fonctionnement de l'outil `Persephone`

Le logiciel `Persephone` appartient à la génération des algorithmes qui recourent à la fonction objective dite de *classification temporelle connectionniste*, CTC (Graves et al., 2013 ; en français, on consultera notamment Tomashenko & Estève, 2018). Le signal audio est soumis à une décomposition fréquentielle par banc de filtres (ce qui revient, pour l'essentiel, à ce que livre une représentation spectrographique), par fenêtres de 10 ms (avec chevauchement). Les traits ainsi extraits sont fournis en entrée à un réseau multi-couche de neurones artificiels récurrents. Une caractéristique importante de cette approche est que le modèle ne contient pas d'hypothèses concernant l'alignement temporel des unités reconnues, de sorte que « l'alignement entre les éléments d'entrée et les étiquettes de sortie est inconnu » (Tomashenko & Estève, 2018, p. 561). Cette propriété du modèle permet de traiter, outre les phonèmes, l'information non segmentale, telle que les tons lexicaux (et tout autre type d'événement figuré dans la transcription fournie en entrée, par exemple un découpage en mots prosodiques).

Dans les expériences relatées ici, l'entraînement du modèle acoustique s'effectue à partir de zéro, sans recourir à des modèles déjà initialisés à partir de données multilingues. Le modèle acoustique est entraîné sur les données d'une unique locutrice. Cela limite d'emblée la portée des généralisations, du fait qu'il n'est pas possible de faire la part des habitudes spécifiques à la locutrice en question. Mais à l'inverse, le choix d'un corpus mono-locuteur permet d'exclure un important facteur de variabilité, et ainsi de pouvoir tirer des conclusions plus certaines, en attendant le stade (évidemment prévu pour la suite) d'une généralisation des observations réalisées. Au plan technique, une expérimentation systématique sur sept langues de la Collection Pangloss (Wisniewski et al., 2020) établit clairement que le passage d'un mode mono-locuteur à un mode multi-locuteurs demandera une amélioration des outils logiciels (selon les méthodes exposées par Tomashenko et al., 2020; Tomashenko & Estève, 2018).

Pour plus d'informations, on renverra à la documentation disponible en ligne[3]. On signalera également des exposés en vidéo (en anglais) au sujet de `Persephone`[4] et de son intégration dans le logiciel de documentation linguistique `ELAN`[5].

[3] https://persephone.readthedocs.io/en/stable/
[4] https://www.youtube.com/watch?v=IwWKqxQ7Qng
[5] https://www.youtube.com/watch?v=-pDOEqRpZKs

2.3 La validation croisée

La méthode employée est la *validation croisée*. L'un des ving-sept documents est retranché du corpus, et un modèle acoustique est entraîné sur le reste du corpus puis appliqué sur le texte qui avait été réservé à cet effet. Cette procédure est appliquée successivement à chacun des vingt-sept documents du corpus na. La validation croisée est une méthode pour éviter des biais d'apprentissage des modèles : il s'agit d'éviter que le modèle ne soit testé sur des données qui faisaient partie du corpus d'apprentissage (erreur classique *de débutant* dans le domaine de l'apprentissage machine). La validation croisée est particulièrement utile dans le cas où le corpus est petit, comme dans le scénario qui nous intéresse ici : cette méthode permet d'utiliser au mieux les données, en les employant comme données de test dans une des expériences, tout en les conservant parmi les données d'entraînement pour les autres.

2.4 Comparaison entre transcriptions générées automatiquement et transcriptions manuelles : le choix d'une analyse qualitative

L'évaluation d'un modèle acoustique s'effectue généralement en quantifiant le taux d'erreur par comparaison avec une transcription de référence produite (ou du moins vérifiée) par un annotateur humain. Dans le travail décrit ici, une évaluation globale a été réalisée, qui conclut à des taux d'erreur de l'ordre de 17% pour la langue na (Adams et al., 2018), en très net progrès par rapport à une étude-pilote réalisée sur les mêmes données au moyen de `CMU-Sphinx` (Do et al., 2014). Au-delà de ce résultat général encourageant, nous avons choisi d'examiner des exemples détaillés, l'un après l'autre, plutôt que d'aborder l'analyse d'erreurs au moyen d'outils statistiques. Des fichiers (au format PDF) ont été générés en mettant en valeur, pour chaque phrase (unité `<S>` du format de la Collection Pangloss : voir Michailovsky et al., 2014), les écarts entre la transcription de référence (manuelle) et la transcription générée automatiquement. Pour la langue na, les vingt-sept documents PDF sont disponibles en ligne[6]. Les documents du corpus d'entraînement peuvent être consultés dans la Collection Pangloss (Michaud et al., 2016)[7]. Nous sommes conscients du fait que les observations présentées ici ne constituent qu'une première approche, qu'il sera utile de poursuivre par une analyse statistique dans les règles de l'art.

3 Premières observations

3.1 La situation particulière des noms propres quadrisyllabiques

Un exemple de transcription automatique suivi de la transcription de référence (manuelle) en vis-à-vis est fourni ci-dessous. Les gloses figurent en exemple (1).

æ̃˩	ʈʂʰe˦	ɖɯ˦	mæ˩		ʈʂʰɯ˦	bi˦	mæ˩	pi˦	dʑo˩	*transcription automatique*
ɻ̩˩	ʈʂʰe˦	ɖɯ˩	mɑ˩	ɻ̩˩	ʈʂʰe˦	ɖɯ˩	mɑ˩	pi˦	dʑo˩	*transcription manuelle*

[6] https://github.com/alexis-michaud/na/tree/master/Persephone/2018_08_StoryFoldCrossValidation
[7] https://pangloss.cnrs.fr/corpus/list_rsc.php?lg=Na&name=na

(1) ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩ pi˧ dʑo˩
Erchei-Ddeema *(nom propre)* dire TOPICALISATEUR
Elle a crié : « Erchei-Ddeema ! Erchei-Ddeema ! » (Texte : *Enterrée vive*, phrase 13. DOI : 10.24397/pangloss-0004537#S13)

Dans cet énoncé, on relève des erreurs de transcription sur les deux occurrences du nom Erchei-Ddeema (nom d'un des principaux protagonistes). La forme phonémique de ce nom est /ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩/. Au vu du taux d'erreur globalement faible (de l'ordre de 17%), il est frappant d'observer neuf erreurs en l'espace d'à peine huit syllabes. L'examen des onze occurrences de ce nom propre dans le texte (reproduites ci-dessous) révèle qu'aucune n'est exempte d'erreurs.

p æ ˩ ʈʂʰ ɯ ˥ ɖ ɯ ˧ m ɣ˩	æ̃˩ ʈʂʰ e ˧ ɖ ɯ ˧ m æ ˩	∅ ʈʂʰ ɯ ˧ b i ˧ m æ ˩
ɑ ˩ ʈʂʰ e ˧ ɖ ɯ ˧ m ɤ ˩	∅ tʰ i ˥ ɖ ɯ ˧ m ɑ ˩	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧
ɻ˩ ʈʂʰ e ˧ ɖ ɯ ˧ m ɤ ˧	ɻ˧ ʈʂʰ ɯ ˧ dʑ ɯ ˧ m ɤ ˧	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧
ɻ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧	

TABLE 2 : Transcription automatique des onze occurrences du nom propre Erchei-Ddeema /ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩/. Le symbole de l'ensemble vide ∅ indique une syllabe manquante.

La première syllabe, l'approximante syllabique / ɻ/, est identifiée comme une voyelle dans six cas, et manque tout à fait dans deux cas. Ce qui la distingue d'une voyelle (dans les réalisations qu'on dira, selon ses préférences théoriques, *canoniques* ou *hyperarticulées*) est essentiellement la rétroflexion, laquelle se manifeste au plan acoustique par un abaissement du troisième formant, jusqu'à des valeurs de l'ordre de 2 000 Hz, nettement inférieures à toutes les autres rimes. Son identification comme une voyelle ouverte suggère que le degré phonétique de rétroflexion / rhotacisation est inférieur, dans ces exemples, à la moyenne statistique.

Le défaut de reconnaissance de ce segment, dans deux cas, tient vraisemblablement à sa coalescence phonétique avec une voyelle qui précède. La structure syllabique (C)V du na de Yongning place en hiatus le noyau de toute syllabe dépourvue de consonne initiale. Un exemple en est fourni en Figure 1. Il révèle un bref passage glottalisé, qui signale vraisemblablement le découpage en constituants (Dilley & Shattuck-Hufnagel, 1996 ; Kuang, 2017, p. 3218) et qui contribue peut-être à masquer la baisse du troisième formant qui, pour l'œil du phonéticien, signale un mouvement articulatoire que n'explique pas la coarticulation avec la consonne affriquée qui suit.

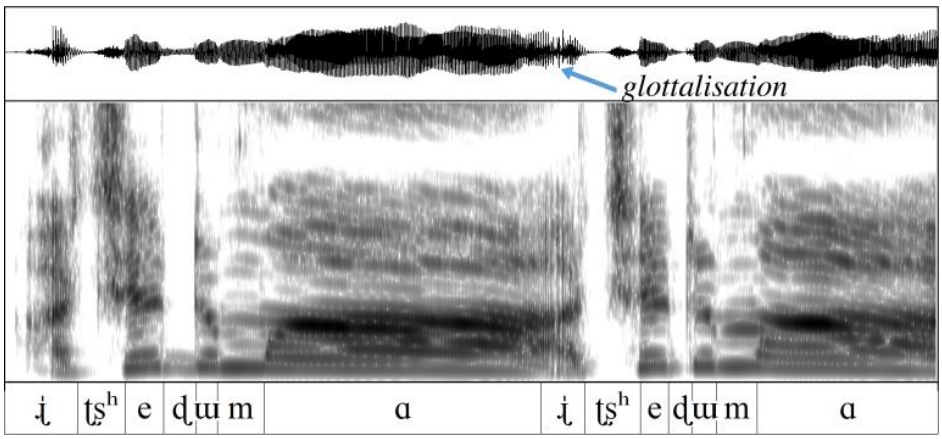


FIGURE 3 : Spectrogramme correspondant à l'exemple (1).

La voyelle de la seconde syllabe du nom /ɻ˩ ʈʂʰe˧ ɖɯ˩ mɑ˩/ est identifiée comme un /ɯ/ dans la majorité des cas. En langue na, la voyelle /ɯ/ possède un allophone apical après les fricatives rétroflexes et affriquées : ainsi, /ʈʂʰɯ/ est réalisé [ʈʂʰz̩] (ou, si l'on adopte les symboles proposés par Chao Yuen-ren, [ʈʂʰɻ̩]. (Au sujet de ces segments mi-voyelle mi-consonne, voir Shao & Ridouane, 2018 et références citées.) L'identification de la voyelle comme /ɯ/ plutôt que /e/ peut donc être interprétée comme la conséquence d'une hypo-articulation de la voyelle. Le mouvement de la langue en direction d'une cible [e] est moins ample que dans la configuration moyenne telle qu'elle est extraite du corpus d'entraînement par le logiciel de transcription automatique. La langue demeure proche de la configuration adoptée pour la consonne [ʈʂʰ].

La troisième des quatre syllabes du nom est, dans ces exemples, moins affectée que les autres, mais son ton est systématiquement identifié comme Moyen (˧) et non Bas (˩). Au plan acoustique, l'examen des données révèle que le schéma tonal /L.M.L.L/ du quadrisyllabe est réalisé avec des valeurs de f_0 plus élevées sur les deuxième et troisième syllabes que sur les première et dernière. Cette observation fait penser aux schémas observés au niveau du mot dans les langues polysyllabiques, et c'est sur cette similarité que nous allons nous appuyer pour proposer une interprétation.

En langue na, les racines lexicales sont monosyllabiques, du fait d'une érosion phonologique spectaculaire au fil de l'histoire de la langue (Jacques & Michaud, 2011). Ces racines monosyllabiques se recombinent en disyllabes par des processus morphologiques de composition et d'affixation, de sorte que les disyllabes sont largement attestés dans le lexique, en particulier parmi les noms (Michaud, 2012). Les disyllabes fournissent, à leur tour, une base pour la formation de mots plus longs. Les mots de quatre syllabes ou plus représentent environ 6% du lexique enregistré à ce jour (Michaud, 2015), et leur fréquence d'occurrence dans les vingt-sept textes transcrits est du même ordre (5,5%). Les quadrisyllabes sont donc marginaux en termes de distribution statistique. Il paraît donc vraisemblable que le modèle acoustique créé au moyen du logiciel `Persephone` fasse la part belle aux transitions acoustiques telles qu'elles sont réalisées sur les monosyllabes et les disyllabes. Le degré de précision avec lequel est réalisé chacun des phonèmes d'un mot court, donc pauvre en matériau phonologique, a toutes chances d'être plus élevé que pour des mots plus longs[8].

Il n'y a rien là de bien nouveau : cette tendance était déjà relevée par Marguerite Durand (1930), et les phonéticiens-phonologues qui s'intéressent à la typologie prosodique (tons et accents) ont maintes occasions de l'observer à l'œuvre. L'éclairage qu'apportent les résultats tirés d'expériences de transcription automatique n'en est pas moins intéressant : ces résultats ouvrent de nouvelles perspectives pour l'étude de la hiérarchie entre les multiples facteurs qui entrent en jeu dans les phénomènes de variation allophonique – laquelle recouvre, *in fine*, le domaine entier de la variation intonative (Vaissière, 2004). Par exemple, fréquence lexicale et nature grammaticale (« mot plein » par opposition à « mot outil », avec toutes les nuances intermédiaires des charges sémantiques et des degrés de grammaticalisation) constituent des facteurs de variation intonative (allophonique) d'importance variable d'une langue à l'autre (Brunelle et al., 2015) : une préposition vietnamienne homophone d'un verbe en diffère moins au plan phonétique que ne le laisserait attendre l'exemple des langues comme le français ou l'anglais. Les observations qualitatives réalisées au sujet des transcriptions automatiques de la langue na suggèrent que la différence entre mots pleins et mots grammaticaux n'est pas particulièrement saillante. Ces observations (qui restent à quantifier) amènent à formuler l'hypothèse selon laquelle la longueur d'un mot a une incidence plus forte sur la façon dont chacun de ses phonèmes est prononcé (dans le corpus considéré) que le statut

[8] Ce raisonnement ne vaut pas pour les mots courts fréquents et fortement prédictibles dans le discours (tels que les mots outils/grammaticaux), souvent hypo-articulés.

grammatical du mot (classe morphosyntaxique) et sa fréquence. On peut donc espérer que la poursuite de ces observations apporte une contribution de nature typologique aux questions de variation allophonique et de « prosodie articulatoire » (Fougeron, 1999, 2001).

3.2 Interprétation des observations

Les quadri-syllabes ne sont pas très fréquents en na (et de manière générale, peu fréquents dans les corpus oraux, dans de nombreuses langues). Les observations rapportées ci-dessus au sujet du nom propre /ɻ˩ ʈʂʰe˧ ɖɯ˩ mɑ˩/ éclairent une des limites du système automatique : le biais statistique qui conduit à accorder plus de poids aux phénomènes plus fréquent, avec pour résultat de moins bonnes performances pour une catégorie qui est marginale en terme de fréquence dans le corpus d'apprentissage.

4 Conclusion et perspectives

Les travaux présentés ici n'en sont qu'à leurs débuts, mais il paraît d'ores et déjà possible de conclure que l'emploi de techniques de Traitement Automatique des Langues Naturelles dans le contexte de la documentation linguistique (« linguistique de terrain ») livre des bénéfices dès les premières étapes de la collaboration entre informaticiens et linguistes. Entre autres perspectives pour la suite du travail, on mentionnera l'extraction d'information à partir des modèles acoustiques générés par apprentissage statistique. L'apprentissage machine suit des procédures qui ne sont pas celles des phonéticiens/phonologues, mais il ne paraît pas impossible de mettre en rapport les probabilités calculées par le modèle avec des variables qui soient interprétables. Il existe diverses méthodes pour explorer ce domaine (Hohman et al., 2019; Jiang et al., 2019; Lapuschkin et al., 2019; Montavon et al., 2017). Dans l'interprétation des résultats, il faut bien sûr savoir raison garder (Gomez-Marin, 2017), mais sans pour autant se priver de suivre les chercheurs en informatique dans leurs explorations en rapide renouvellement. L'étude des modèles acoustiques pourrait, en particulier, fournir un appui dans l'entreprise qui consiste à caractériser les phonèmes d'une langue en termes de propriétés acoustiques (Vaissière, 2011a, 2011b) et articulatoires (Stavness et al., 2012), et ainsi parvenir à un degré de précision nettement supérieur à celui que permet l'Alphabet Phonétique International.

Remerciements

Nos vifs remerciements aux locuteurs et amis na. Nous remercions vivement les deux relecteurs des *Journées d'Étude de la Parole*, ainsi que tous les collègues qui participent au développement et à l'utilisation d'outils de reconnaissance automatique pour langues peu dotées ; qu'ils nous pardonnent de ne pas nous livrer à l'exercice impossible qui consisterait à dresser une liste un tant soit peu complète.

Le logiciel `Persephone` a bénéficié en 2018-2019 du soutien de l'Université du Queensland et d'une bourse d'innovation transdisciplinaire du *Centre of Excellence for the Dynamics of Language* du Conseil australien de la recherche (ARC). Il bénéficie actuellement du soutien du projet franco-allemand « La documentation computationnelle des langues à l'horizon 2025 » (CLD 2025, ANR-19-CE38-0015-04) et du projet d'Institut des Langues Rares (ILARA) de l'École pratique des Hautes Études (dans le cadre du plan *Sciences humaines et sociales 2020* du Ministère de

l'enseignement supérieur, de la recherche et de l'innovation). Le présent travail s'inscrit en outre dans le cadre du Labex « Fondements empiriques de la linguistique » (EFL, ANR-10-LABX-0083).

Références

ADAMS, O. (2017). *Automatic understanding of unwritten languages* [Ph.D.]. The University of Melbourne.

ADAMS, O., COHN, T., NEUBIG, G., CRUZ, H., BIRD, S., & MICHAUD, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. HAL : https://halshs.archives-ouvertes.fr/halshs-01709648

ADAMS, O., COHN, T., NEUBIG, G., & MICHAUD, A. (2017). Phonemic transcription of low-resource tonal languages. *Proceedings of ALTA 2017 (Australasian Language Technology Association Workshop)*, 53–60. HAL : https://halshs.archives-ouvertes.fr/halshs-01656683

ADDA, G., STÜKER, S., ADDA-DECKER, M., AMBOUROUE, O., BESACIER, L., BLACHON, D., BONNEAU-MAYNARD, H., GODARD, P., HAMLAOUI, F., IDIATOV, D., KOUARATA, G.-N., LAMEL, L., MAKASSO, E.-M., RIALLAND, A., VAN DE VELDE, M., YVON, F., & ZERBIAN, S. (2016). Breaking the unwritten language barrier: The BULB Project. *SLTU-2016 5th Workshop on Spoken Language Technologies for Under-Resourced Languages 09-12 May 2016 Yogyakarta, Indonesia*, *81*(Supplement C), 8–14. DOI : 10.1016/j.procs.2016.04.023

BRUNELLE, M., CHOW, D., & NGUYỄN, T. N. U. (2015). Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. *Proceedings of ICPhS XVIII*. International Congress of the Phonetic Sciences XVIII, Glasgow.

DILLEY, L., & SHATTUCK-HUFNAGEL, S. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*, 423–444.

DO, T. N. D., MICHAUD, A., & CASTELLI, E. (2014). Towards the automatic processing of Yongning Na (Sino-Tibetan): Developing a "light" acoustic model of the target language and testing "heavyweight" models from five national languages. *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 153–160. HAL : http://halshs.archives-ouvertes.fr/halshs-00980431

DURAND, M. (1930). *Etude sur les phonèmes postérieurs dans une articulation parisienne*. Didier.

FOLEY, B., ARNOLD, J., COTO-SOLANO, R., DURANTIN, G., & ELLISON, T. M. (2018). Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, 200–204.

FOLEY, B., RAKHI, A., LAMBOURNE, N., BUCKERIDGE, N., & WILES, J. (2019). Elpis, an accessible speech-to-text tool. *Proceedings of Interspeech 2019*, 306–310.

FOUGERON, C. (1999). Prosodically conditioned articulatory variations: A review. *UCLA Working Papers in Phonetics*, *97*, 1–68.

FOUGERON, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135.

GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., & PALLETT, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report n*, *93*.

GODFREY, J. J., HOLLIMAN, E. C., & MCDANIEL, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 517–520.

GOMEZ-MARIN, A. (2017). Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient? In *Decoding neural circuit structure and function* (pp. 283–306). Springer.

GRAVES, A., MOHAMED, A., & HINTON, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.

GREENBERG, S., HOLLENBACK, J., & ELLIS, D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of the International Conference on Spoken Language Processing*, *96*, 24–27.

HJORTNAES, N., PARTANEN, N., RIEßLER, M., & TYERS, F. M. (2020). Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, 31–37.

HOHMAN, F., HEAD, A., CARUANA, R., DELINE, R., & DRUCKER, S. M. (2019). *Gamut: A design probe to understand how data scientists understand Machine Learning models*. ACM CHI Conference on Human Factors in Computing Systems, Glasgow.

JACQUES, G., & MICHAUD, A. (2011). Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze. *Diachronica*, *28*(4), 468–498.

JIANG, Z., XU, F. F., ARAKI, J., & NEUBIG, G. (2019). How can we know what language models know? *ArXiv:1911.12543*. http://arxiv.org/abs/1911.12543

JIMERSON, R., & PRUD'HOMMEAUX, E. (2018). ASR for documenting acutely under-resourced indigenous languages. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 4161–4166.

JONES, D. (1950). *The Phoneme, its Nature and Use*. Heffer.

KUANG, J. (2017). Creaky voice as a function of tonal categories and prosodic boundaries. *Proceedings of Interspeech 2017*, 3216–3220.

LAPUSCHKIN, S., WÄLDCHEN, S., BINDER, A., MONTAVON, G., SAMEK, W., & MÜLLER, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1096.

LIDZ, L. (2010). *A descriptive grammar of Yongning Na (Mosuo)* [Ph.D., University of Texas, Department of linguistics]. https://repositories.lib.utexas.edu/bitstream/handle/2152/ETD-UT-2010-12-2643/LIDZ-DISSERTATION.pdf

LITTELL, P., KAZANTSEVA, A., KUHN, R., PINE, A., ARPPE, A., COX, C., & JUNKER, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. *Proceedings of the 27th International Conference on Computational Linguistics*, 2620–2632.

MICHAILOVSKY, B., MAZAUDON, M., MICHAUD, A., GUILLAUME, S., FRANÇOIS, A., & ADAMOU, E. (2014). Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation*, *8*, 119–135. HAL : https://halshs.archives-ouvertes.fr/halshs-01003734

MICHAUD, A. (2008). Phonemic and tonal analysis of Yongning Na. *Cahiers de Linguistique - Asie Orientale*, *37*(2), 159–196. HAL : https://halshs.archives-ouvertes.fr/halshs-00358610

MICHAUD, A. (2012). Monosyllabicization: Patterns of evolution in Asian languages. In N. Nau, T. Stolz, & C. Stroh (Eds.), *Monosyllables: From phonology to typology* (pp. 115–130). Akademie Verlag. HAL : http://halshs.archives-ouvertes.fr/halshs-00436432

MICHAUD, A. (2015). *Dictionnaire na-chinois-français*. HAL : https://halshs.archives-ouvertes.fr/halshs-01204645

MICHAUD, A. (2017). *Tone in Yongning Na: Lexical tones and morphotonology*. Language Science Press. http://langsci-press.org/catalog/book/109

MICHAUD, A., ADAMS, O., COHN, T., NEUBIG, G., & GUILLAUME, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, *12*, 393–429. HAL : https://halshs.archives-ouvertes.fr/halshs-01841979

MICHAUD, A., ADAMS, O., COX, C., & GUILLAUME, S. (2019). Phonetic lessons from automatic phonemic transcription: Preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*. ICPhS XIX (19th International Congress of Phonetic Sciences), Melbourne. HAL : https://halshs.archives-ouvertes.fr/halshs-02059313

MICHAUD, A., GUILLAUME, S., JACQUES, G., MẠC, Đ.-K., JACOBSON, M., PHẠM, T. H., & DEO, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation: La Collection Pangloss et la Collection AuCo. *Actes de La Conférence Conjointe JEP-TALN-RECITAL 2016, Volume 1 : Journées d'Etude de La Parole*, *1*, 155–163. HAL : https://halshs.archives-ouvertes.fr/halshs-01341631

MICHAUD, A., HARDIE, A., GUILLAUME, S., & TODA, M. (2012). Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, E. Castelli, Dong Minghui, & Pham Thi Ngoc Yen, (Eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)* (pp. 169–172). MICA Institute, Hanoi University of Science and Technology. HAL : https://halshs.archives-ouvertes.fr/halshs-00731261

MONTAVON, G., SAMEK, W., & MÜLLER, K.-R. (2017). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

NEUBIG, G., RIJHWANI, S., PALMER, A., MACKENZIE, J., CRUZ, H., LI, X., LEE, M., CHAUDHARY, A., GESSLER, L., & ABNEY, S. (2020). A summary of the first Workshop on Language Technology for Language Documentation and Revitalization. *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-Resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*. Marseille, France. *ArXiv:2004.13203 [Cs]*. https://arxiv.org/abs/2004.13203

PRATAP, V., HANNUN, A., XU, Q., CAI, J., KAHN, J., SYNNAEVE, G., LIPTCHINSKY, V., & COLLOBERT, R. (2018). wav2letter++: The fastest open-source speech recognition system. *ArXiv:1812.07625 [Cs]*. http://arxiv.org/abs/1812.07625

SHAO, B., & RIDOUANE, R. (2018). La « voyelle apicale » en chinois de Jixi: Caractéristiques acoustiques et comportement phonologique. *XXXIIe Journées d'Études Sur La Parole*, 685–693. DOI : 10.21437/JEP.2018-78

STAVNESS, I., GICK, B., DERRICK, D., & FELS, S. (2012). Biomechanical modeling of English /r/ variants. *The Journal of the Acoustical Society of America*, *131*(5), EL355–EL360. DOI : 10.1121/1.3695407

THIEBERGER, N. (2017). LD&C possibilities for the next decade. *Language Documentation and Conservation*, *11*, 1–4.

TOMASHENKO, N., & ESTÈVE, Y. (2018). Impact des techniques d'adaptation au locuteur dans l'espace des paramètres pour des modèles acoustiques purement neuronaux. *XXXIIe Journées d'Études Sur La Parole*, 559–567. DOI : 10.21437/JEP.2018-64

TOMASHENKO, N., KHOKHLOV, Y., & ESTÈVE, Y. (2020). *Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models*. HAL : https://hal.archives-ouvertes.fr/hal-02551714

VAISSIÈRE, J. (2004). The perception of intonation. In D. B. PISONI & R. E. REMEZ (Eds.), *Handbook of Speech Perception* (pp. 236–263). Blackwell.

VAISSIÈRE, J. (2011a). On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. *Proceedings of ICPhS XVII*. ICPhS XVII, Hong Kong.

VAISSIÈRE, J. (2011b). Proposals for a representation of sounds based on their main acoustico-perceptual properties. In E. HUME, J. GOLDSMITH, & W. L. WETZELS (Eds.), *Tones and Features* (pp. 306–330). De Gruyter Mouton.

VAN ESCH, D., FOLEY, B., & SAN, N. (2019). Future directions in technological support for language documentation. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, *1*, 3. https://www.aclweb.org/anthology/W19-6003.pdf

WATANABE, S., HORI, T., KARITA, S., HAYASHI, T., NISHITOBA, J., UNNO, Y., SOPLIN, N. E. Y., HEYMANN, J., WIESNER, M., & CHEN, N. (2018). Espnet: End-to-end speech processing toolkit. ArXiv:1804.00015.

WISNIEWSKI, G., GUILLAUME, S., & MICHAUD, A. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-Resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*. 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop, Marseille, France. HAL : https://halshs.archives-ouvertes.fr/hal-02513914

WU, M., LIU, F., & COHN, T. (2018). Evaluating the utility of hand-crafted features in sequence labelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2850–2856. DOI : 10.18653/v1/D18-1310

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comment l'oreille de présentation affecte-t-elle la capacité des francophones à discriminer des contrastes accentuels natifs et non-natifs ?

Amandine Michelas[1] Sophie Dufour[1]
(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
michelas@lpl-aix.fr, sophie.dufour@lpl-aix.fr

RÉSUMÉ

Dans cette étude, nous avons examiné la capacité des auditeurs francophones natifs à percevoir la variation accentuelle en manipulant l'oreille de présentation des mots. Deux contrastes accentuels ont été testés : un contraste natif (/balɔ̃/-/baˈlɔ̃/) et un contraste non-natif (/ˈbalɔ̃/-/baˈlɔ̃/). Dans une tâche ABX, les participants entendaient trois mots produits par trois locuteurs différents et devaient déterminer si X était identique à A ou à B. Les stimuli A et B différaient sur l'accent (/ˈbalɔ̃/-/baˈlɔ̃/), sur un phonème (/baˈlɔ/-/baˈlɔ̃/) ou sur l'accent et un phonème (/ˈbalɔ/-/baˈlɔ̃/). Les résultats ont montré des difficultés persistantes pour le contraste non-natif quelle que soit l'oreille de présentation. Par contre, pour le contraste natif, des meilleures performances ont été observées lorsque les mots étaient présentés dans l'oreille gauche. D'une façon générale, notre étude montre que la variation accentuelle présente au niveau du mot est traitée par les auditeurs francophones natifs comme de la variation de surface.

ABSTRACT

How does the ear of presentation affect the ability of French listeners to discriminate native and non-native accentual contrasts?

In this study, we examined the ability of French listeners to perceive accentual variation by manipulating the ear of words presentation. Two accentual contrasts were tested : a native (/balɔ̃/-/baˈlɔ̃/ "ball") and a non-native (/ˈbalɔ̃/-/baˈlɔ̃/) contrasts. In an ABX task, in which participants heard three words produced by three different speakers, participants had to determine whether X was identical to A or to B. The stimuli and A and B varied in accent (/ˈbalɔ̃/-/baˈlɔ̃/), in one phoneme (/baˈlɔ/-/baˈlɔ̃/ " bundle-ball") or both in accent and in one phoneme (/ˈbalɔ/-/baˈlɔ̃/). The results showed persistent difficulty for the non-native contrast regardless of the ear of presentation. By contrast, for the native contrast, better performance was observed when words were presented in the left ear. Together, these results show that French native listeners process accentual variation that affects the word level as surface variation.

MOTS-CLÉS : Perception de la parole, variation de la parole, avantage hémisphérique droit, oreille de présentation, contraste accentuel, prosodie, français.
KEYWORDS: Speech perception, speech variation, right hemisphere advantage, ear of presentation, accentual contrast, prosody, French.

Volume 1 : Journées d'Études sur la Parole, pages 463–471.
HAL : hal-02798573.

1 Introduction

Une des caractéristiques les plus frappantes de la prosodie du français est que l'accent (c'est-à-dire la mise en valeur d'une syllabe d'un point de vue acoustique) est déterminé de façon post-lexicale et ne permet donc pas de créer deux mots de sens différents. A cause de cela, il est bien connu que les auditeurs francophones natifs ont des difficultés à discriminer deux items qui diffèrent du point de vue de la position de l'accent (ex. BEbe /ˈbebe/ « il/elle boit » et beBE /beˈbe/ « bébé » en espagnol; Dupoux, Pallier, Sebastian-Gallès & Mehler, 1997; Dupoux, Peperkamp & Sebastian-Gallés, 2001; Dupoux, Peperkamp & Sebastian-Gallés, 2010; Peperkamp, Vendelin & Dupoux, 2010; Rahmani, Rietveld & Gussenhoven, 2010). Le terme de « surdité à l'accent » a d'ailleurs été employé pour qualifier la faible sensibilité des auditeurs francophones natifs aux variations accentuelles qui affectent le mot. Cependant, des études récemment menées dans notre laboratoire (Michelas, Frauenfelder, Schön & Dufour, 2016; Michelas, Esteve-Gibert & Dufour, 2018) ont montré que cette faible sensibilité aux variations accentuelles n'est pas vraie pour tous les contrastes accentuels. En effet, nous avons montré que les auditeurs du français sont pleinement capables de différencier un mot accentué d'un mot non-accentué, c'est-à-dire de différentier les contrastes accentuels qu'ils produisent et perçoivent continuellement dans la vie de tous les jours. En nous basant sur des études précédentes, montrant que la variation dans le signal de parole est prioritairement prise en charge par l'hémisphère cérébral droit lorsqu'elle n'est pas distinctive au niveau du mot (Van Lancker & Canter, 1982 ; Von Kriegstein, Eger, Kleinschmidt & Giraud, 2003; Gonzalez & McLennan, 2007), la présente étude a pour objectif d'examiner le rôle de l'oreille de présentation dans la discrimination de contrastes accentuels par des auditeurs francophones natifs. Plus particulièrement, nous nous attendons à observer de meilleures performances dans la discrimination de contrastes qu'ils soient natifs ou non-natifs lorsque le traitement est contraint dans l'hémisphère droit, et donc en raison de projections controlatérales, lorsque les mots sont présentés dans l'oreille gauche.

Que savons-nous de la capacité des auditeurs francophones natifs à discriminer des contrastes accentuels ? Dans une étude princeps, Dupoux et ses collaborateurs (Dupoux et al., 1997) ont comparé la capacité des auditeurs francophones et hispanophones natifs à discriminer des contrastes accentuels qui existent en espagnol mais pas en français. Dans une tâche ABX, dans laquelle A, B et X étaient prononcés par trois locuteurs différents, et au cours de laquelle les participants devaient juger si X était identique à A ou à B, les auteurs ont montré que, comparé aux auditeurs hispanophones, les auditeurs francophones avaient plus de difficultés à distinguer deux non-mots qui différaient du point de vue accentuel (FIdape, fiDApe). De plus, contrairement aux auditeurs hispanophones, les auditeurs francophones avaient de moins bonnes performances lorsque les stimuli différaient d'un point de vue accentuel (FIdape, fiDApe) que lorsqu'ils différaient d'un point de vue phonémique (FIdape, LIdape). De manière générale, les résultats obtenus par Dupoux et al. (1997) suggèrent qu'une différence accentuelle est traitée avec plus de difficultés qu'une différence phonémique par les auditeurs francophones natifs. Cependant, cette difficulté des francophones à traiter les contrastes accentuels pourrait être liée au fait que Dupoux et al. (1997) ont utilisé des contrastes de position d'accent qui n'existent pas en français (BOpelo/boPElo et boPElo/bopeLO).

Contrairement à Dupoux et al. (1997), des études récentes (Michelas et al., 2016 ; Michelas et al., 2018) ont examiné la capacité des auditeurs francophones natifs à discriminer des contrastes accentuels qui sont fréquemment rencontrés dans leur langue maternelle. En utilisant le même paradigme expérimental, les auteurs ont répliqué les résultats de Dupoux et al. (1997) qui montrent

que les auditeurs francophones natifs ont plus de difficultés à discriminer deux mots qui diffèrent du point de vue de la position de l'accent (ex. JUry-juRY) que deux mots qui diffèrent d'un point de vue phonémique (ex. juRON-juRY). Cependant, aucune difficulté n'a été observée lorsque les stimuli différaient du point de vue de la présence ou non d'un accent (ex. jury-juRY), c'est-à-dire lorsqu'ils étaient soumis au contraste accentuel auquel ils sont régulièrement exposés dans leur langue native. Une telle observation suggère que les auditeurs francophones natifs ont des difficultés à traiter des contrastes accentuels non-natifs, mais que cette difficulté disparaît dès lors qu'ils ont à discriminer des contrastes accentuels qui existent dans leur langue maternelle.

Dans la mesure où les francophones se sont montrés capables de faire la différence entre un mot accentué et un mot non-accentué (Michelas et al., 2016, 2018), dans une étude récente (Michelas & Dufour, 2019), nous avons examiné si ces variations accentuelles sont susceptibles d'être stockées dans le lexique mental. Dans cette étude, les participants entendaient un premier bloc de stimuli (bloc amorce) puis un second bloc de stimuli (bloc cible) dans lequel des mots du bloc amorce étaient répétés soit avec la même accentuation (banDEAU – banDEAU) soit avec une accentuation différente (bandeau – banDEAU). En comparaison à une condition contrôle dans laquelle amorce et cible n'avaient aucun lien (marron – banDEAU), nous avons rapporté un effet d'amorçage de répétition d'amplitude similaire que les amorces et les cibles soit répétées dans la même accentuation ou non. Un tel résultat laisse donc suggérer que les variations accentuelles ne sont pas stockées dans le lexique mental même si les francophones natifs sont pleinement capables de distinguer une forme accentuée d'une forme non-accentuée.

L'ensemble des études conduites sur le français pourraient donc laisser supposer que la variation accentuelle au niveau du mot serait traitée comme de la variation de surface. Le terme «variation de surface» (souvent utilisé pour désigner la variation indexicale) fait référence à toute variation dans la forme du mot qui n'est pas lexicalement distinctive et qui n'a donc pas d'impact direct sur son activation et sa reconnaissance. Dans cette étude, nous avons testé l'hypothèse selon laquelle la variation accentuelle au niveau du mot en français serait traitée comme de la variation de surface en nous inspirant de l'étude de Gonzalez et McLennan (2007) qui ont montré que la variation liée à l'identité du locuteur est prioritairement prise en charge par l'hémisphère droit (cf. aussi Van Lancker & Canter, 1982; Von Kriegstein et al., 2003). Ces auteurs ont montré que ce type de variation influence le traitement des mots parlés uniquement lorsque les mots sont présentés dans l'oreille gauche et donc, prioritairement traités par l'hémisphère droit[1]. De manière plus précise, dans notre étude, nous nous sommes demandées si les auditeurs francophones ne seraient pas meilleurs pour discriminer des contrastes accentuels lorsque les stimuli sont présentés dans l'oreille gauche. Pour ce faire, nous avons repris la tâche ABX de Dupoux et al. (1997) dans laquelle des participants entendaient trois mots produits par trois locuteurs différents et devaient décider si X était similaire à A ou à B. A et B étaient prononcés par deux voix féminines différentes et X était toujours prononcé par un locuteur masculin. Deux types de contrastes accentuels ont été testés : 1) le contraste accentuel qui existe en français; c.-à-d. la différence entre un mot non-accentué et le même mot portant un accent primaire sur sa dernière syllabe : /balɔ̃/ vs. /baˈlɔ̃/ («contraste accentuel natif») et 2) un contraste de position d'accent qui n'existe pas en français, c.-à-d. la différence entre

[1] Même si de façon inévitable, les activations se dissipent d'un hémisphère cérébral à l'autre via les commissures inter-hémisphériques et le corps calleux, manipuler l'hémichamp de présentation s'est avéré particulièrement efficace aussi bien à l'oral (ex. Gonzalez et McLennan, 2007) qu'à l'écrit (ex. Marsolek, 2004) pour évaluer le rôle respectif de chacun des hémisphères dans le traitement de la variation au sein des mots.

un mot porteur d'un accent primaire sur sa syllabe initiale et le même mot porteur d'un accent primaire sur sa syllabe finale[2]: /ˈbalɔ̃/ et /baˈlɔ̃/ («contraste accentuel non-natif»). Au sein de chaque contraste accentuel, les stimuli A et B différaient soit du point de vue accentuel, soit du point de vue phonémique, soit des points de vue accentuel et phonémique. Les participants entendaient les mots soit dans l'oreille gauche soit dans l'oreille droite, tandis qu'un bruit leur était présenté dans l'oreille opposée. Nous nous attendions à ce qu'une différence accentuelle soit traitée avec la même facilité qu'une différence phonémique, que le contraste soit natif ou non-natif, lorsque les mots étaient présentés dans l'oreille gauche et donc prioritairement traités dans l'hémisphère droit.

Méthodologie

2.1 Participants

50 participants droitiers, francophones natifs, entre 18 et 37 ans, ont participé à l'expérience. Ils ont tous rapporté n'avoir aucun trouble auditif ou du langage. 25 participants ont entendu les stimuli dans l'oreille droite tandis que les 25 autres ont entendus les stimuli dans l'oreille gauche.

2.2 Matériel

Quatre paires de mots français bisyllabiques, composés de quatre phonèmes, qui différaient sur un phonème (ex. ballon /balɔ̃/ vs. ballot /balo/) ont été sélectionnés. Trois locuteurs de langue maternelle française (deux femmes et un homme) ont produit les huit mots cibles (en gras dans les exemples suivants) à l'intérieur de phrases porteuses. A l'intérieur de ces phrases porteuses, les mots cibles étaient soit accentués sur la syllabe finale soit non-accentués en fonction de leur position à l'intérieur du syntagme (ex. [On m'avait parLÉ] [d'un **ballon** maGIQUE] [qui ne se creVAIT] [presque jaMAIS] vs. [On m'avait parLÉ] [d'un petit **baLLON**] [qui ne se creVAIT] [presque jaMAIS]). Les mêmes phrases porteuses que pour la condition non-accentuée ont été utilisées pour obtenir les huit mots qui étaient accentués sur leur syllabe initiale ([On m'avait parLÉ] [d'un **BAllon** maGIQUE] [qui ne se creVAIT] [presque jaMAIS]. Parce qu'il est normalement impossible de trouver des mots qui portent un accent primaire sur leur syllabe initiale en français, nous avons explicitement demandé aux locuteurs de produire artificiellement un accent primaire sur cette syllabe en insistant dessus. Pour éviter les effets de coarticulation liés à la parole en contexte, chaque mot a ensuite était extrait de sa phrase porteuse. Les trois locuteurs entendaient leur propre prononciation des huit mots cibles en isolation et devaient reproduire les différentes versions des mots qu'ils entendaient (cf. Michelas et al., 2016, 2018, 2019 pour une méthodologie similaire). Les phrases produites et les mots présentés en isolation ont été enregistrés avec une fréquence d'échantillonnage de 44 100 Hz et un taux de quantification de 16 bits. Des analyses acoustiques ont ensuite été conduites au moyen du logiciel Praat (Boersma & Weenink, 2019) afin de s'assurer que les mots répétés ont été produits avec les patrons accentuels attendus. Ainsi pour chaque version (non-accentuée, accentuée sur la première syllabe, accentuée sur la dernière syllabe) de chacun des mots, nous avons mesuré la durée des deux syllabes ainsi que le mouvement de f0 associé à ces syllabes. Des analyses statistiques conduites sur les huit mots cibles, produits par les trois locuteurs, ont ensuite été réalisées sur la durée de la syllabe et la montée de f0. Ces analyses ont montré que pour les mots non-accentués, la durée et la montée de f0 associée à la syllabe initiale n'étaient pas

[2] Bien qu'il existe un accent secondaire qui affecte de manière optionnelle la syllabe initiale des mots en français, cet accent n'a pas les mêmes propriétés acoustiques que l'accent primaire sur lequel nous nous focalisons dans cette étude (voir par exemple Welby, 2006).

différentes de celles associées à la syllabe finale [durée de la syllabe : t(23)=0.77, p > .20 ; montée de f0 : t(23)=1.31, p> .20]. Au contraire, lorsque les mots étaient porteurs d'un accent primaire sur la syllabe initiale, les syllabes initiales étaient plus longues et associées à une montée de f0 plus importante que les syllabes finales [durée de la syllabe : t(23)=7.75, p < .0001 ; montée de f0 : t(23)=13.14, p < .0001]. Enfin, lorsque les mots étaient porteurs d'un accent primaire sur la syllabe finale, les syllabes finales étaient plus longues et associées à une montée de f0 plus importante que les syllabes initiales [durée de la syllabe : t(23)=8.05, p < .0001 ; montée de f0 : t(23)=30.19, p < .0001]. La valeur efficace des signaux (c'est-à-dire le RMS, *root-mean-square*) a été normalisée à 70 dB pour tous les mots cibles. Les propriétés acoustiques du mot ballon /balɔ̃/ produit dans sa version non-accentuée, avec un accent primaire sur sa dernière syllabe et avec un accent primaire sur sa syllabe initiale sont montrées en Figure 1.

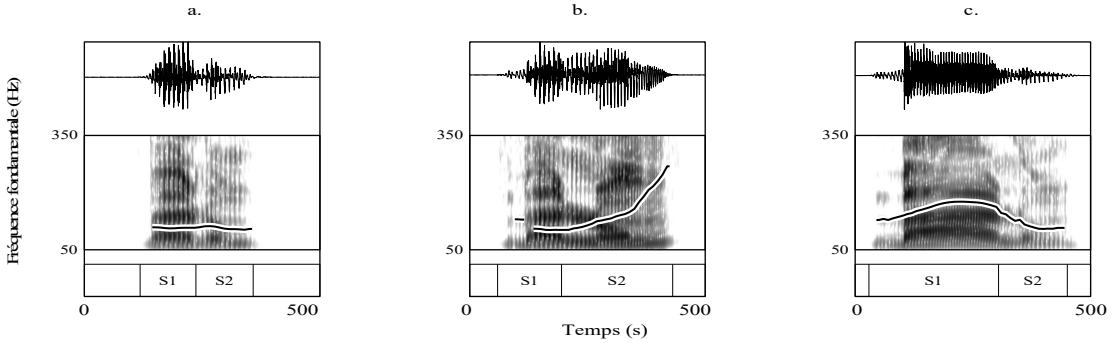


FIGURE 1: Profil phonémique et prosodique du mot ballon dans sa version non-accentuée (a), avec un accent primaire sur sa syllabe finale (b) et avec un accent primaire sur sa syllabe initiale (c).

De façon similaire à Gonzalez & Mc Lennan (2007), un bruit rose était également présenté aux participants dans l'oreille opposée à l'oreille de présentation des mots. Ce type de bruit a été utilisé étant donné qu'il partage certaines caractéristiques avec le signal de parole. En effet, tout comme ce dernier, la densité spectrale de puissance du bruit rose diminue à mesure que la fréquence augmente. Notre fichier contenant du bruit rose avait une durée de 612 ms correspondant à la durée du mot cible le plus long. La valeur efficace du signal contenant ce bruit a été normalisée à 50 dB.

Chaque essai était composé de trois stimuli : A, B et X, avec A et B prononcés par les deux voix féminines et X prononcé par la voix masculine. Pour les deux types de contrastes («contraste accentuel natif» et «contraste accentuel non-natif»), A et B différaient soit du point de vue accentuel, soit du point de vue phonémique, soit à la fois du point de vue accentuel et phonémique, donnant lieu à 6 conditions expérimentales. Les quatre paires de mots ont été utilisées dans chaque condition expérimentale. Pour chaque paire de mots et à l'intérieur de chaque condition expérimentale, 16 combinaisons ont été utilisées. Ces 16 combinaisons sont le résultat du croisement entre le patron accentuel (2 versions possibles à chaque fois : ex. ballon vs. baLLON), du contenu phonémique (2 mots possibles à chaque fois: ex. ballon vs. ballot), du type de réponse (2 types de réponse : A vs. B) et du type de voix (2 voix : voix féminine 1 vs. voix féminine 2). 384 essais ont ainsi été obtenus (6 conditions expérimentales x 4 paires de mots x 16 combinaisons).

2.3 Procédure

Les participants ont été testés dans une chambre sourde et les stimuli ont été présentés au moyen d'écouteurs à un niveau sonore confortable (60 dB) identique pour tous les participants. La moitié d'entre eux a entendu les stimuli dans l'oreille gauche alors que du bruit rose leur était présenté simultanément dans l'oreille opposée. L'autre moitié des participants a entendu les stimuli dans l'oreille droite alors que du bruit rose leur était simultanément présenté dans l'oreille opposée. Chaque essai expérimental était composé de trois stimuli A, B et X séparés par un intervalle de 500

ms. Les participants ont été informés au préalable qu'ils allaient entendre des mots dans une oreille seulement alors que du bruit leur serait simultanément diffusé dans l'oreille opposée. Ils ont a également été informé que les deux stimuli A et B présentaient des différences et que le troisième stimuli (X) était soit similaire au premier (A) soit au deuxième (B). La tâche des participants consistait donc à indiquer si X était similaire à A ou à B en appuyant sur un bouton situé sur leur droite ou sur leur gauche. L'ordre de présentation des essais était randomisé et différent pour chaque participant. 1000ms s'écoulaient entre la réponse du participant et le début de l'essai suivant. Les participants ont commencé l'expérience avec un entraînement de 12 essais. L'expérience durait environ 30 minutes.

3 Résultats

Le pourcentage de réponses correctes pour l'oreille droite et pour l'oreille gauche en fonction du type de différence (accentuelle, phonémique, accentuelle et phonémique) est présenté en Figure 2 pour le contraste natif et en Figure 3 pour le contraste non-natif.

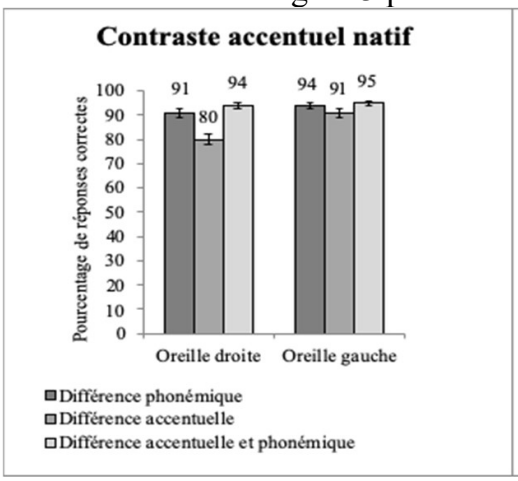


FIGURE 2: Pourcentage de réponses correctes pour le contraste accentuel natif. Les barres d'erreur représentent les erreurs standards.

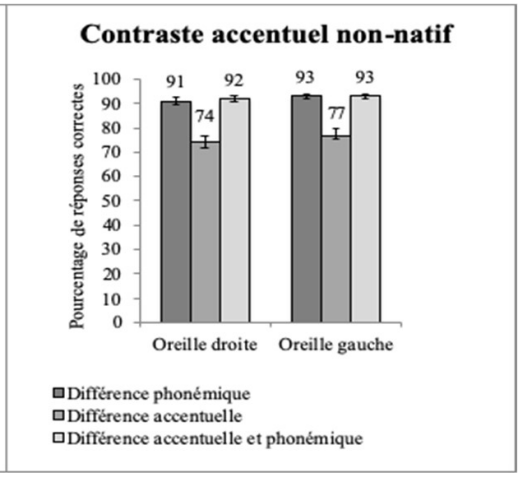


FIGURE 3: Pourcentage de réponses correctes pour le contraste accentuel non-natif. Les barres d'erreur représentent les erreurs standards.

Les réponses données par les participants ont été analysées grâce à un modèle de régressions logit à effets mixtes. Le modèle incluait le type de différence (phonémique, accentuelle, phonémique et accentuelle), le type de contraste accentuel (natif vs. non-natif) et l'oreille de présentation (droite vs. gauche) ainsi que leurs interactions comme effets fixes. Le modèle incluait également des intercepts aléatoires par participant et par item et des pentes aléatoires par participant et par item pour l'effet du type de différence et du type de contraste accentuel (Barr, Levy, Scheepers & Tily, 2013). Notre ensemble de données étant trop volumineux pour les paramètres par défaut de glmer.nb, l'option nAGQ=0 a été implémentée afin de permettre au modèle de converger (Bates, Bolker & Walker, 2016). Le modèle a révélé une interaction type de différence x type de contraste accentuel x oreille de présentation significative ($X^2 = 21.10$; $p < .0001$). Cette double interaction a été examinée dans deux modèles mixtes séparés pour le contraste accentuel natif d'une part et pour le contraste accentuel non-natif d'autre part. Les deux modèles incluaient le type de différence et l'oreille de présentation comme facteurs fixes, des intercepts aléatoires par participant et par item et des pentes aléatoires par participant et par item pour l'effet du type de différence. L'option nAGQ=0 a été ajoutée aux deux modèles. Les comparaisons multiples ont été obtenues grâce au package multcomp (Bretz, Hothorn & Westfall, 2011) avec une correction de Bonferroni.

Concernant le **contraste accentuel natif,** le modèle a révélé un effet significatif de l'oreille de présentation (X^2= 59.26, p < .0001) avec plus de réponses correctes lorsque les stimuli étaient présentés dans l'oreille gauche que dans l'oreille droite. L'effet du type de différence était également significatif (X^2= 34.32, p < .0001) avec plus de réponses correctes dans la condition de différence phonémique que dans la condition de différence accentuelle (z=-3. 27, p < .01) et plus de réponses correctes dans la condition de différence accentuelle et phonémique que dans la condition de différence accentuelle (z=-5. 25, p < .0001). **De manière cruciale, l'interaction entre l'oreille de présentation et le type de différence était significative (X^2= 22.09, p < .0001).** Lorsque les stimuli était présentés dans **l'oreille gauche**, les performances étaient similaires entre les conditions de différence (toutes les p.values >.20). Au contraire, lorsque les stimuli étaient présentés dans **l'oreille droite**, le modèle a révélé plus de réponses correctes dans la condition de différence phonémique que dans la condition de différence accentuelle (z=-4.71, p < .0001) et a également révélé plus de réponses correctes dans la condition de différence accentuelle et phonémique que dans la condition de différence accentuelle (z=-7.43, p < .0001).

Concernant le **contraste accentuel non-natif**, le modèle a révélé un effet significatif de l'oreille de présentation avec plus de réponses correctes lorsque les stimuli étaient présentés dans l'oreille gauche que dans l'oreille droite (X^2= 4.12, p < .05). L'effet du type de différence était également significatif (X^2= 94.79, p < .0001) avec plus de réponses correctes dans la condition de différence phonémique que dans la condition de différence accentuelle (z=-7.94, p < .0001) et plus de réponses correctes dans la condition de différence accentuelle et phonémique que dans la condition de différence accentuelle (z=-9.75, p < .0001). De façon inattendue, l'interaction entre l'oreille de présentation et le type de différence n'était pas significative (X^2= 2.62, p >.20).

4 Discussion

Le but de cette étude était d'examiner comment l'oreille de présentation affecte l'utilisation de la variation accentuelle dans la discrimination de mots chez des auditeurs francophones natifs. Nous somme parties de la présupposition que la variation accentuelle affectant le mot serait traitée par les auditeurs francophones comme de la variation de surface étant donné que celle-ci n'est pas pertinente pour distinguer les mots en français. Plus particulièrement, nous nous attendions à ce qu'une différence d'accent soit traitée avec un même niveau de performance qu'une différence de phonème que le contraste accentuel soit natif ou non quand les mots étaient présentés dans l'oreille gauche et donc prioritairement traités par l'hémisphère droit.

Les résultats observés pour le contraste accentuel natif étaient conformes à notre hypothèse. En effet, nous avons observé des performances similaires dans la discrimination d'un contraste accentuel et dans la discrimination d'un contraste phonémique à la condition que les stimuli soient présentés dans l'oreille gauche. Ainsi lorsque le traitement était contraint dans l'hémisphère droit, les participants traitaient une différence accentuelle avec autant de facilité qu'une différence phonémique. Ces résultats viennent confirmer ceux précédemment obtenus en écoute binaurale (Michelas et al., 2018) montrant que les francophones natifs sont tout à fait capables d'utiliser la présence/absence d'un accent pour discriminer des mots. Par contre, nous avons observé des difficultés dans le traitement de la présence/absence d'un accent lorsque les mots étaient présentés dans l'oreille droite, et donc lorsque le traitement était contraint dans l'hémisphère gauche. Ceci s'est traduit par de moins bonnes performances dans la condition de discrimination accentuelle que dans la condition de discrimination phonémique et ceci malgré le fait que ce contraste soit quotidiennement produit et perçu par les participants dans leur vie quotidienne.

De manière inattendue, aucun effet de l'oreille de présentation n'a été observé lorsque les participants devaient discriminer des contrastes accentuels qui n'existent pas en français. Aussi bien pour l'oreille droite que pour l'oreille gauche de présentation, nos participants francophones natifs avaient plus de difficultés à discriminer un contraste accentuel qu'un contraste phonémique. Un tel résultat vient répliquer les observations faites par Dupoux et al. (1997; voir aussi Dupoux et al., 2001; Dupoux et al., 2010; Peperkamp et al., 2010; Rahmani et al., 2015; Michelas et al., 2018). Ce résultat montre également qu'à partir du moment où un contraste accentuel n'est pas présent dans notre langue, il est extrêmement difficile de remédier aux difficultés de perception de ce contraste. Notre étude vient donc renforcer toutes les études qui ont montré que le système phonologique de la langue native impacte fortement la manière dont nous percevons les sons de parole (voir par ex. Best, McRoberts, Goodell, 2001). En effet, nos résultats montrent que le système prosodique de notre langue maternelle agit lui aussi comme un filtre rendant ainsi difficile la discrimination de contrastes accentuels qui ne sont pas présents dans notre langue maternelle.

Étant donné que l'information prosodique n'est pas pertinente pour distinguer les mots en français, nous avons mis en évidence un avantage de l'hémisphère droit pour le traitement de la prosodie de façon similaire à l'avantage de l'hémisphère droit qui a été observé par Gonzalez et Mc Lennan (2007) pour le traitement de la variation liée à l'identité du locuteur. Le fait que de la variation prosodique au niveau du mot soit traitée comme de la variation de surface par des auditeurs francophones natifs a également été observé dans une étude portant sur les tons (Hallé, Chang & Best, 2004). Dans cette étude, les auteurs ont comparé la manière dont des locuteurs natifs du français et des locuteurs natifs du chinois mandarin percevaient des continuums tonals créés de manière artificielle en manipulant à la fois la f0 et l'intensité associées aux syllabes porteuses de tons. Dans une tâche ABX où A et B constituaient les extrémités d'un continuum tonal et X variait d'une extrémité à l'autre selon huit étapes, les participants devaient juger si X ressemblait d'avantage à A ou à B. Les résultats montrent que les auditeurs sinophones avaient de meilleures performances lorsque X était proche des extrémités du continuum (étape 2 ou 7) que lorsque X était situé au milieu du continuum (étape 4 ou 5). Au contraire, les auditeurs francophones avaient des performances similaires, quelle que soit la position de X au sein du continuum. Ces résultats montrent donc que les auditeurs sinophones et francophones ne traitent pas l'information prosodique de la même manière. Parce que cette information permet de créer des distinctions de sens au niveau du mot en chinois, les auditeurs natifs de cette langue la perçoivent de façon catégorielle. Par contre, pour les auditeurs francophones natifs, comme cette information ne permet pas de créer des distinctions de sens au niveau de mot, elle est traitée comme de la variation de surface.

Les résultats de la présente étude corroborent également l'hypothèse de Van Lancker (1980) selon laquelle l'hémisphère cérébral prenant en charge l'information prosodique dépendrait du rôle de la prosodie dans la langue. Comme l'information prosodique n'est pas pertinente au niveau du mot en français, il est apparu que c'est l'hémisphère droit qui prend en charge le traitement de cette information pour les francophones natifs. De ce fait, notre étude montre qu'un contraste natif est perçu comme un contraste non-natif lorsque le traitement est contraint dans l'hémisphère non-dominant pour le traitement de la prosodie dans la langue. D'une façon cruciale, notre étude montre aussi qu'un contraste non-natif n'atteint jamais les performances d'un contraste natif et ceci quelle que soit l'oreille de présentation.

En conclusion, il apparaît que les locuteurs francophones natifs utilisent l'information accentuelle qu'ils produisent et perçoivent quotidiennement pour discriminer deux mots. De manière intéressante, notre étude montre qu'ils utilisent d'autant mieux cette information lorsqu'elle est prise en charge par l'hémisphère dominant pour le traitement de la prosodie dans cette langue, autrement dit lorsqu'elle est pris en charge par l'hémisphère droit.

Références

BARR, D. J., LEVY, R., SCHEEPERS, C. & TILY, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

BATES, D., B. BOLKER & WALKER, S. 2016. Package 'lme4', version 1.1- 12, linear mixed effects models using 'Eigen' and S4. Available at: https://cran.r-project.org/web/packages/lme4/lme4.pdf.

BOERSMA, P., & WEENINK, D. (2019). Praat: doing phonetics by computer (Version 6.0. 52)[Windows].

BEST, C. T., MCROBERTS, G. W., & GOODELL, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America,* 109(2), 775–794.

BRETZ, F., HOTHORN, T. & WESTFALL, P. (2016). *Multiple comparisons using R*. CRC Press.

DUPOUX, E., PALLIER, C., SEBASTIÁN-GALLÉS, N. & MEHLER, J. (1997). A destressing "deafness" in French?, *Journal of Memory and Language,* 36, 406-421.

DUPOUX, E., PEPERKAMP, S. & SEBASTIÁN-GALLÉS, N. (2001) A robust method to study stress "deafness". *The Journal of the Acoustical Society of America*, 110 (3), 1606-1618.

DUPOUX, E., PEPERKAMP, S., & SEBASTIÁN-GALLÉS, N. (2010). Limits on bilingualism revisited: Stress 'deafness' in simultaneous French–Spanish bilinguals. *Cognition*, *114*(2), 266-275.

GONZÁLEZ, J. & MCLENNAN, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 410.

HALLÉ, P. A., CHANG, Y. C. & BEST, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics*, *32*(3), 395-421.

MARSOLEK, C. J. (2004). Abstractionist versus exemplar-based theories of visual word priming: A subsystems resolution. *The Quarterly Journal of Experimental Psychology*, Section A, 57(7), 1233-1260.

MICHELAS, A., & DUFOUR, S. (2019). Are Prosodic Variants Stored in the French Mental Lexicon?. *Experimental Psychology*, 66 (6), 393-401.

MICHELAS, A., FRAUENFELDER, U. H., SCHÖN, D. & DUFOUR, S. (2016). How deaf are French speakers to stress?. *The Journal of the Acoustical Society of America*, *139*(3), 1333-1342.

MICHELAS, A., ESTEVE-GIBERT, N. & DUFOUR, S. (2018). On French listeners' ability to use stress during spoken word processing. *Journal of Cognitive Psychology*, *30*(2), 198-206.

PEPERKAMP, S., VENDELIN, I. & DUPOUX, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, *38*(3), 422-430.

RAHMANI, H., RIETVELD, T. & GUSSENHOVEN, C. (2015). Stress "deafness" reveals absence of lexical marking of stress or tone in the adult grammar. *PloS one*, *10*(12).

VAN LANCKER, D. (1980). Cerebral lateralization of pitch cues in the linguistic signal. *Research on Language & Social Interaction*, *13*(2), 201-277.

VAN LANCKER, D. R. & CANTER, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and cognition*, *1*(2), 185-195.

VON KRIEGSTEIN, K., EGER, E., KLEINSCHMIDT, A. & GIRAUD, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48-55.

WELBY, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, *34*(3), 343-37

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Beatboxer, est-ce parler ? Ce que nous en dit l'étude de la dynamique articulatoire d'un beatboxeur

Annalisa Paroni[1] Nathalie Henrich Bernardoni[1] Christophe Savariaux[1]
Pierre Baraduc[1] Hélène Lœvenbruck[2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA–lab, F–38000 Grenoble
(2) Univ. Grenoble Alpes, Univ. Savoie Mont–Blanc, CNRS, LPNC, F–38000 Grenoble
annalisa.paroni@gipsa-lab.fr nathalie.henrich@gipsa-lab.fr

RÉSUMÉ

Les consonnes plosives sont parmi les phonèmes les plus représentés dans l'inventaire phonologique des langues du monde. Outre leur rôle linguistique, elles remplissent également un rôle paralinguistique dans la pratique instrumentale et vocale, notamment au sein de la pratique vocale du Human Beatbox. Cet article apporte un éclairage sur les similitudes et différences dans la dynamique articulatoire de trois consonnes plosives du français et des sons percussifs correspondants du Human Beatbox. Si ces deux modes de production vocale ont une racine commune, une dynamique articulatoire différente est mise en évidence pour le Human Beatbox. Nous retrouvons des indices d'un mécanisme éjectif, qui a un impact sur la dynamique linguale.

ABSTRACT

Beatboxing, is it talking ? What the study of the articulatory dynamics of a beatboxer tells us

Plosive consonants are among the most commonly-found phonemes in the phonological inventory of the world's languages. In addition to their linguistic role, they also fulfil a paralinguistic role in instrumental and vocal practice, especially within the practice of Human Beatbox. This article sheds light on the similarities and differences in the articulatory dynamics of three plosive consonants of French and the corresponding percussive sounds of Human Beatbox. These two modes of vocal production share a common root. Yet, different articulatory dynamics are highlighted for Human Beatbox. Clues of an ejective mechanism are found, that impact the lingual dynamics.

MOTS-CLÉS : plosive, son percussif, beatboxer, human beatbox, dynamique articulatoire, articulographie électromagnétique.

KEYWORDS: plosive, percussive sound, beatboxing, human beatbox, articulatory dynamics, electromagnetic articulography.

1 Introduction

Les consonnes plosives /p, t, k/ font partie des phonèmes les plus représentés dans l'inventaire phonologique des langues du monde (Maddieson & Disner, 1984). Du point de vue articulatoire, les plosives sont produites grâce au relâchement plus ou moins soudain d'une occlusion du conduit vocal (CV). Cette occlusion se situe dans le conduit oral, au niveau des lèvres, du palais dur, du velum ou dans l'espace pharyngé. Ce point d'articulation a une influence sur les caractéristiques acoustiques du bruit de plosion (Forrest *et al.*, 1988) qui joue un rôle important dans la discrimination des consonnes

Volume 1 : Journées d'Études sur la Parole, pages 472–479.
HAL : hal-02798574.

plosives (CNET-ENST, 1989). En français les points d'articulation discriminatifs pour les plosives sont les lèvres (plosives bilabiales), les alvéoles (plosives alvéolaires) ou le vélum (plosives vélaires). A la marge de leur rôle linguistique, les plosives ont aussi un rôle paralinguistique. Les syllabes avec consonnes plosives sont utilisées dans la pratique instrumentale, pour le jeu des instruments à vent et des percussions. Elles soutiennent également la pratique vocale, comme par exemple dans le cas du scat ou du konnakol indien. L'origine linguistique de ces sons a un impact sur l'articulation de ces mêmes sons quand ils sont utilisés de façon paralinguistique. Certaines études se sont intéressées aux différences articulatoires des plosives utilisées dans la pratique instrumentale en fonction de la langue maternelle de l'instrumentiste (Lamkin, 2005; Heyne & Derrick, 2014, 2015; Heyne *et al.*, 2019).

Dans cette étude, nous nous intéressons à la pratique du Human Beatbox (HBB), un art vocal émergent qui s'appuie sur l'instrument vocal humain pour produire toutes sortes de son dans le but de faire de la musique. Les sons percussifs y sont très exploités. Bien souvent, l'apprentissage du HBB commence par un travail sur des plosives, des syllabes ou des phrases de la parole. Tel est le cas du kick, le son utilisé pour imiter la grosse caisse de la batterie, appris à partir d'un [p], du hi–hat, qui reproduit le son du charleston, travaillé à partir d'un [t] ou [ts], ou encore de la technique du rimshot, sur la base d'un [k]. Quelles sont les similitudes et les différences dans la dynamique articulatoire entre parole et HBB ? La base de données exploitée pour apporter une première réponse à cette question est présentée en Section 2. Nous présenterons les résultats en terme de trajectoires articulatoires en Section 3. Nous comparerons la production des plosives et des sons percussifs ayant même point d'articulation.

2 Matériel et méthodes

Nous présentons les résultats d'une étude pilote menée sur un sujet beatboxeur à partir de la technique expérimentale de l'articulographie électromagnétique (EMA) combinée à des mesures acoustiques, électroglottographiques et ventilatoires. Au sein de cet article, nous nous concentrons uniquement sur les mesures articulatoires. Le sujet est un beatboxeur de 28 ans ayant une pratique amateur de 9 ans en HBB et dont la langue maternelle est le français.

Nous nous sommes intéressés à la production en parole des plosives /p, b, t, d, k, g/ et de leur équivalent en HBB en termes de point d'articulation : le kick, le hi–hat et le rimshot. Trois phrases – *des p'tits cookies des gros cookies* (Cookies) ; *pâtes au pesto* (Pâtes) ; *boots and cats* (BootsAndCats) – ont été répétées plusieurs fois à partir d'un mode parlé jusqu'à un mode beatboxé, et inversement (Cookies, 8 répétitions parlées et 8 répétitions beatboxées ; Pâtes : 8 répétitions parlées et 17 répétitions beatboxées ; BootsAndCats : 16 répétitions parlées et 23 répétitions beatboxées). Ces phrases ont été choisies car elles sont assez couramment utilisées par les beatboxeurs dans l'apprentissage de leur art vocal.

Les enregistrements ont eu lieu dans la salle semi–anéchoique de la plateforme BEDEI du laboratoire GIPSA–lab de Grenoble, lieu de recherches biomédicales autorisé par l'ARS Auvergne–Rhône–Alpes depuis Mars 2016. Le protocole expérimental a reçu un avis favorable du Comité d'Ethique pour les Recherches Grenoble Alpes. Après signature du consentement éclairé, le sujet a été équipé d'un gilet pour la mesure pléthysmographique respiratoire à variation d'inductance (système VISURESP, RBI). Il a été installé dans la salle semi–anéchoique sur une chaise adaptée pour la stabilisation de la tête à l'intérieur du champ magnétique de l'articulographe électromagnétique 3D (EMA WAVE, NDI).

Pour le recueil des données articulatoires, 12 bobines ont été collées sur des points de chair d'intérêt :

— 3 bobines sur la langue dans le plan médio–sagittal : 1 dans la région apicale (TIP), 1 dans la région dorso–palatale (MID) et 1 dans la région dorso–vélaire (DORS) ;
— 1 bobine sur la mâchoire (JAW) rattachée à l'incisive inférieure ;
— 4 bobines sur les lèvres de façon médiane ou latérale : 2 sur la lèvre supérieure (ULL et ULM) et 2 sur la lèvre inférieure (LLL et LLM) ;
— 4 bobines de référence : 1 sur l'incisive supérieure, 1 derrière chaque oreille et 1 sur le front.

Le signal EMA a été enregistré à $400Hz$. Des électrodes de mesure du contact glottique par électroglottographie (EGG) ont été placées sur son cou au niveau du larynx. Un microphone AKG et un microphone omnidirectionnel pré-polarisé 1/2" (B&K 4189) connecté à un pré–amplificateur (B&K 2669C) et à un amplificateur de conditionnement NEXUS (B&K 2690) ont été positionnés à environ 20 cm de la bouche du beatboxeur pour mesurer le signal audio et son intensité. Une caméra a été placée face au sujet pour des enregistrements vidéo à 25 images par seconde.

Les données audio ont été segmentées manuellement et annotées sous Praat (Boersma, 2006). Les annotations phonétiques ont permis de repérer les instants où avaient lieu les bursts des plosives. Ces marqueurs temporels ont été sauvegardés dans un fichier TextGrid. Les données articulatoires ont ensuite été analysées sous MATLAB (MATLAB, 2018). Afin de visualiser la variabilité des mouvements des articulateurs, la trajectoire médiane et l'écart interquartile des bobines d'intérêt ont été calculés en prenant en compte toutes les occurrences du même son pour un même mode de production, en choisissant comme instant de référence le moment du burst détecté sur les données audio. Les données EMA sont mesurées en 3D, mais nous avons choisi de présenter ici l'analyse selon l'axe vertical (y) indiquant les mouvements bas–haut. La médiane de la vitesse tangentielle (3D) et l'écart interquartile ont été également calculés.

Notre étude de la dynamique articulatoire porte sur l'analyse de l'évolution dans le temps de la trajectoire d'une bobine donnée qui se déplace dans l'espace, ainsi que de sa vitesse. Nous menons cette analyse sur une fenêtre temporelle de $400ms$ pour la phrase Cookies et $800ms$ pour BootsAndCats, centrée sur le burst.

3 Résultats

L'analyse des données articulatoires (EMA et vidéo) montre que le lieu d'articulation de chacune des plosives reste sensiblement le même en parole et en HBB.

Les Figures 1 et 2 illustrent la dynamique articulatoire de la réalisation de la plosive /p/ et de son équivalent percussif, le kick, prononcée dans la phrase Cookies dans sa version parlée "*des **p**'tits cookies ...*" ou réalisée en HBB par la série "***p** t k t t **p** k t*". Ces sons sont obtenus par occlusion complète des lèvres. Dans les deux modes de production (parlé et beatboxé), cette occlusion précède le moment du burst d'environ $100ms$. En parole, un comportement articulatoire reproductible est observé pour les différentes répétitions de la phrase. Il est mis en évidence sur la Figure 1 par un écart interquartile faible. Un comportement plus variable est observé en HBB à proximité du burst pour ce même point d'articulation (milieu des lèvres). Il est intéressant de noter qu'un comportement homogène (écart interquartile faible) similaire à celui observé en parole se retrouve au niveau de la moité gauche des lèvres (bobines LL). Le relâchement de ce son percussif bilabial est donc latéralisé à gauche, ce que confirment les données vidéo.

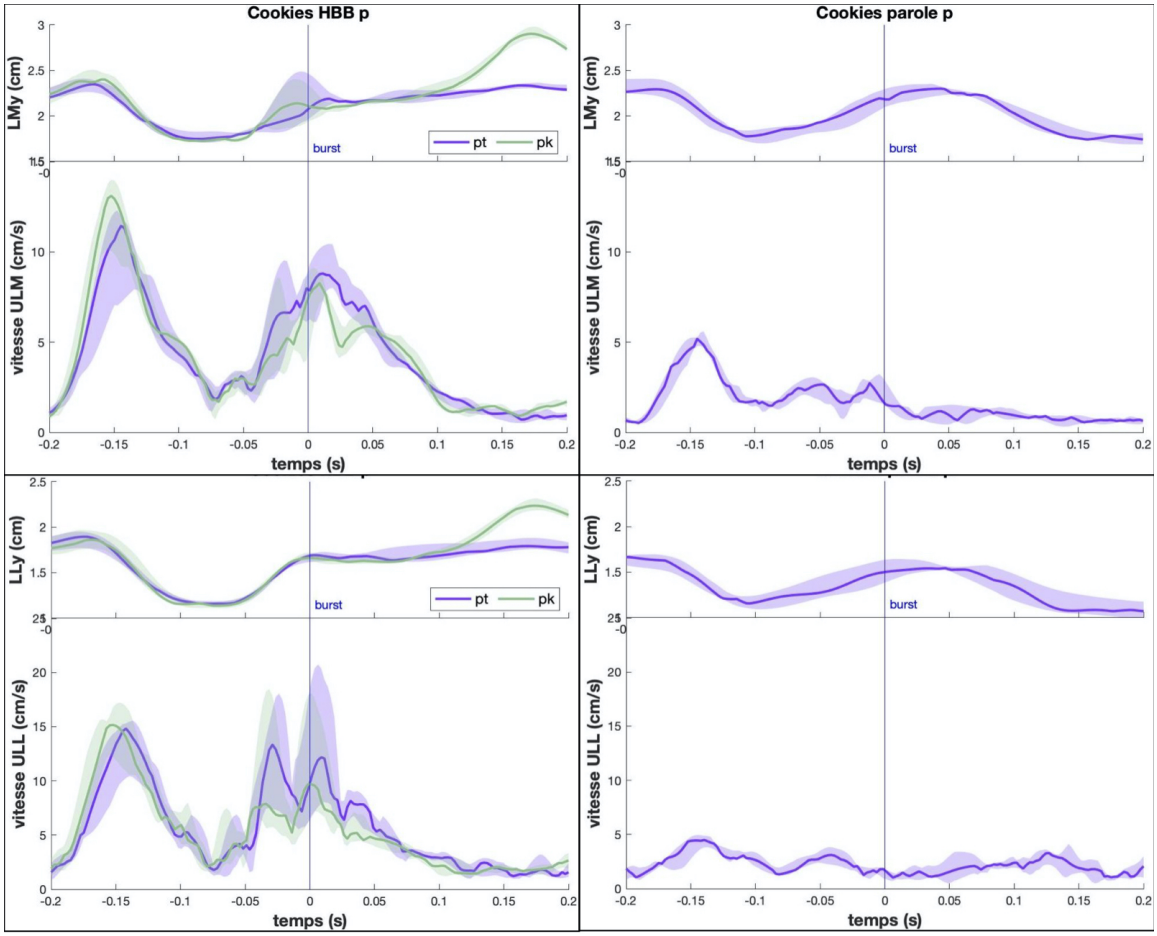


FIGURE 1 – Haut : médiane et écart interquartile de la distance interbobine du milieu des lèvre (LM), du côté gauche des lèvres (LL) au long de l'axe y, calculés à partir de toutes les occurences de la plosive bilabiale pour la phrase Cookies et sa contrepartie beatboxée. Bas : médiane et écart interquartile de la vitesse tangentielle (3D) de la bobine d'intérêt.

La vitesse de relâchement de l'occlusion bilabiale est nettement plus grande en HBB qu'en parole, avec une augmentation de la vitesse de la bobine rattachée au milieu de la lèvre du haut (ULM sur Fig. 1) qui peut atteindre 10 cm/s aux alentours du burst, contre 2,5 cm/s en parole. Quant à la bobine de la lèvre supérieure gauche (ULL sur Fig. 1), plus proche du lieu du relâchement de l'occlusion en HBB, elle peut atteindre 20 cm/s au moment du burst, alors qu'en parole sa vitesse est quasiment nulle. Une différence entre parole et HBB est également observée au niveau de la dynamique linguale, ainsi que l'illustre la Figure 2 pour la bobine posée sur le dos de la langue. Si le mouvement de cette bobine suit le burst en parole, il le précède en HBB. Nous pouvons noter un mouvement de déplacement vertical du dos de la langue marqué en HBB, s'accompagnant d'une variabilité très faible de la trajectoire au moment du burst, mais différent en fonction du son qui suit.

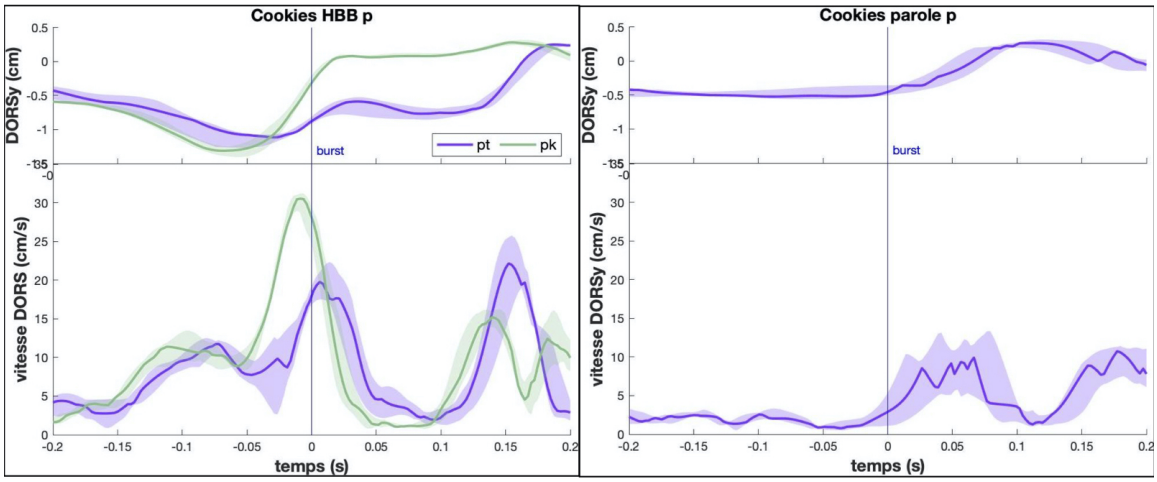


FIGURE 2 – Haut : médiane et écart interquartile de la bobine DORS au long de l'axe y, calculés à partir de toutes les occurences de la plosive bilabiale pour la phrase Cookies et sa contrepartie beatboxée. Bas : médiane et écart interquartile de la vitesse tangentielle (3D) de la bobine DORS.

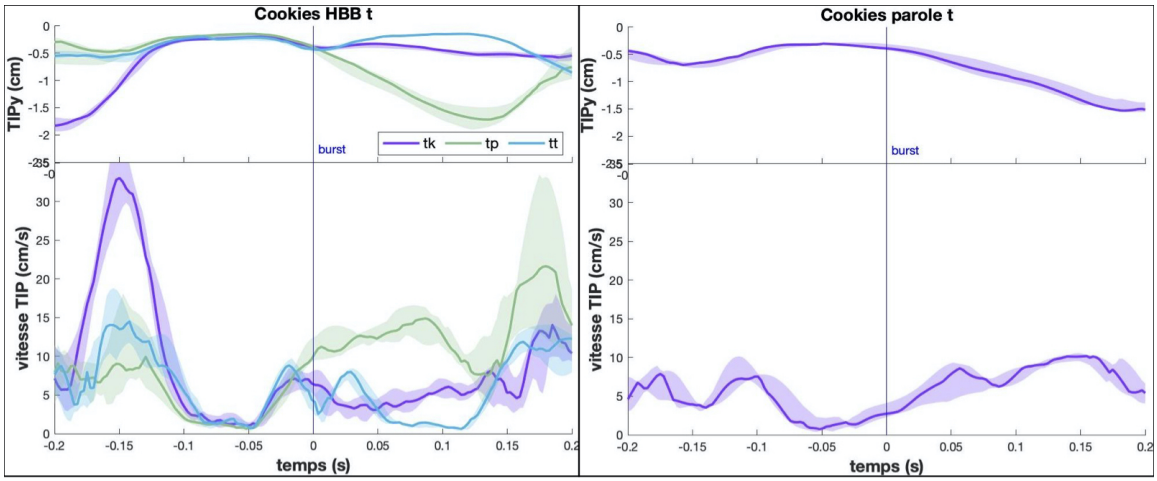


FIGURE 3 – Haut : médiane et écart interquartile de la bobine TIP au long de l'axe y, calculés à partir de toutes les occurences de la plosive alvéolaire pour la phrase Cookies et sa contrepartie beatboxée. Bas : médiane et écart interquartile de la vitesse tangentielle (3D) de la bobine TIP.

La Figure 3 illustre la dynamique articulatoire de la pointe de la langue pour la production de la plosive /t/ et de son équivalent en HBB, le hi–hat, dans le cas de la même phrase "*des p'**t**its cookies* ..." et de la série "*p **t** k **t** **t** p k **t***". Un plateau bien marqué est observé en HBB en correspondance duquel la vitesse de la bobine est presque nulle. Ceci suggère que l'occlusion du CV a bien lieu dans la région alvéolaire ou post–alvéolaire et qu'elle est maintenue un court instant. Ce plateau ne se retrouve pas en parole. Dans les deux modes de production, le relâchement de l'occlusion tel que détecté par la bobine TIP précède le burst détecté sur le signal audio. Dans les deux cas, la variabilité de la trajectoire est réduite. En particulier, l'écart interquartile est faible avant le burst. Bien que moins marquée que dans le cas du kick, la vitesse de déplacement de la bobine d'intérêt au moment du relâchement est plus grande en HBB qu'en parole.

Concernant la plosive /k/ et son équivalent rimshot en HBB, nous présentons le cas de la phrase BootsAndCats dans sa version parlée "*boots and **c**ats*", ou beatboxée comme suit : $p^{(n)}$ *ts* ***k*** *ts*.

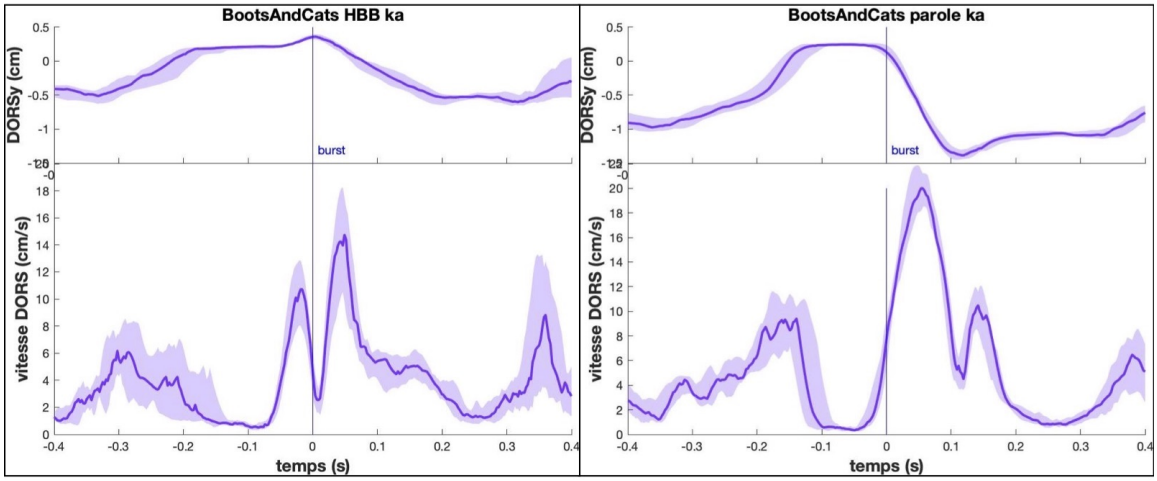


FIGURE 4 – Haut : médiane et écart interquartile de la bobine DORS au long de l'axe y, calculés à partir de toutes les occurences de la plosive vélaire pour la phrase BootsAndCats et sa contrepartie beatboxée. Bas : médiane et écart interquartile de la vitesse tangentielle (3D) de la bobine DORS.

En parole comme en HBB (Fig. 4), les trajectoires du dos de la langue (bobine DORS) montrent un plateau avant le burst, indiquant l'occlusion du CV au niveau de la région palato–vélaire. En parole, ce plateau se maintient jusqu'au burst, où la bobine montre un mouvement descendant vers la position de la voyelle suivante ([a]). En HBB, un mouvement ultérieur de la bobine DORS est observé avant le burst : vers la fin de l'occlusion, la bobine subit un déplacement vers le haut qui se termine en correspondance du burst. Ce mouvement n'est pas aléatoire, comme l'indique l'écart interquartile très faible. Il induit un pic de vitesse avant le burst (environ 10 cm/s), qui n'est pas observé en parole. En effet en parole, le seul pic de vitesse aux alentours du burst est celui lié au déplacement de la langue de l'occlusion vélaire vers la position beaucoup plus basse de la voyelle [a]. En HBB, deux pics de vitesse sont observés, l'un juste avant et l'autre juste après le burst, ce qui fait qu'au moment du burst la bobine DORS est en train de décélérer.

4 Discussion et conclusion

Nos résultats montrent que la production des sons percussifs du HBB se distingue clairement de celle des sons plosifs de la parole, même s'il est possible d'entrevoir une racine comune notamment au niveau du point d'articulation et de la trajectoire générale.

Les détails de l'articulation des sons percussifs du HBB revèlent une dynamique articulatoire propre au HBB et différente de celle de la parole. Dans le cas du kick comparé à une plosive sourde bilabiale, l'occlusion est latéralisée. Cette latéralisation du relâchement de l'occlusion bilabiale s'observe chez certains beatboxeurs, mais pas de façon systématique. Le type de latéralité (gauche ou droite) dépend du beatboxeur. Ce geste labial a probablement pour but de mieux contrôler la tension des lèvres au moment du relâchement. La langue est également très active même si l'occlusion s'effectue au niveau des lèvres. Les vitesses de déplacement des articulateurs sont souvent plus grandes en HBB qu'en parole, surtout avant le burst. La dynamique linguale des sons percussifs du HBB semble être cohérente avec l'utilisation d'un mécanisme éjectif. La dynamique linguale très active et reproductible durant l'articulation des sons percussifs bilabiaux du HBB ne semble pas pouvoir être complètement expliquée par la coarticulation, puisque à égalité d'environnement phonétique en parole (notamment /pt/), nous n'observons pas de mouvements comparables. En outre, toutes les trois bobines de la langue sont impliquées quasiment de la même manière, indiquant un déplacement global de la langue. Cela semble plutôt suggérer un mouvement de la langue en relation avec la remontée du larynx propre au mécanisme éjectif. En effet, plusieurs études ont attesté l'utilisation de ce mécanisme dans l'articulation des sons percussifs du HBB (Proctor *et al.*, 2013; De Torcy *et al.*, 2014; Sapthavee *et al.*, 2014; Blaylock *et al.*, 2017; Patil *et al.*, 2017; Dehais Underdown *et al.*, 2019). Il en va de même pour les mouvements de remontée de la langue à la fin de la phase d'occlusion des sons percussifs vélaires du HBB. Quant à la nature de ces mouvements linguaux, Proctor et collègues (Proctor *et al.*, 2013) font l'hypothèse que la langue soit utilisée avec le larynx pour produire une action de poussée plus efficace. En effet, un mécanisme éjectif optimalisé pourrait permettre d'accroître l'efficacité sonore du son produit, ce qui est très important en HBB. En revanche, nos données n'ont pas mis en évidence ce type de mouvements lors de l'articulation des plosives alvéolaires d'aucune des trois phrases beatboxées. Cela pourrait s'expliquer par le fait qu'en ces cas la remontée du larynx n'a pas d'influence sur la langue de part sa position articulatoire (apex plus elevé que le dos). Il se peut aussi que notre beatboxeur n'utilise pas de mécanisme éjectif pour la réalisation des alvéolaires. De manière générale et à notre connaissance, peu d'études ont exploré l'impact du mécanisme éjectif sur la dynamique linguale, en parole comme en HBB. Grâce à une dynamique articulatoire très active et à la large utilisation de ce mécanisme, l'étude de la production des sons percussifs du HBB permet de mettre en évidence des phénomènes articulatoires au niveau lingual.

Références

BLAYLOCK R., PATIL N., GREER T. & NARAYANAN S. S. (2017). Sounds of the human vocal tract. In *INTERSPEECH*, p. 2287–2291.

BOERSMA P. (2006). Praat : doing phonetics by computer. *http://www.praat.org/*.

CNET-ENST C. (1989). La parole et son traitement automatique.

DE TORCY T., CLOUET A., PILLOT-LOISEAU C., VAISSIERE J., BRASNU D. & CREVIER-BUCHMAN L. (2014). A video–fiberscopic study of laryngopharyngeal behaviour in the human beatbox. *Logopedics Phoniatrics Vocology*, **39**(1), 38–48.

DEHAIS UNDERDOWN A., BUCHMAN L. & DEMOLIN D. (2019). Acoustico-Physiological coordination in the Human Beatbox : A pilot study on the beatboxed Classic Kick Drum. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia. HAL : hal-02284132.

FORREST K., WEISMER G., MILENKOVIC P. & DOUGALL R. N. (1988). Statistical analysis of word-initial voiceless obstruents : preliminary data. *The Journal of the Acoustical Society of America*, **84**(1), 115–123.

HEYNE M. & DERRICK D. (2014). Some initial findings regarding first language influence on playing brass instruments. In *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*, p. 180–183.

HEYNE M. & DERRICK D. (2015). The influence of tongue position on trombone sound : A likely area of language influence. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)* : University of Canterbury. New Zealand Institute of Language, Brain & Behaviour.

HEYNE M., DERRICK D. & AL-TAMIMI J. (2019). Native language influence on brass instrument performance : An application of generalized additive mixed models (gamms) to midsagittal ultrasound images of the tongue. *Frontiers in Psychology*, **10**, 2597.

LAMKIN L. L. (2005). An examination of correlations between flutists' linguistic practices and their sound production on the flute. In *Proceedings of the Conference on Interdisciplinary Musicology*.

MADDIESON I. & DISNER S. F. (1984). *Patterns of sounds*. Cambridge university press.

MATLAB, 2018 (2018). Mathworks : Bioinformatics toolbox : User's guide (r2018b).

PATIL N., GREER T., BLAYLOCK R. & NARAYANAN S. S. (2017). Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging. In *Interspeech*, p. 2277–2281.

PROCTOR M., BRESCH E., BYRD D., NAYAK K. & NARAYANAN S. (2013). Paralinguistic mechanisms of production in human "beatboxing" : A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, **133**(2), 1043–1054.

SAPTHAVEE A., YI P. & SIMS H. S. (2014). Functional endoscopic analysis of beatbox performers. *Journal of Voice*, **28**(3), 328–331.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Différences acoustiques inter-genres chez des bilingues Anglais/Français : une étude des formants vocaliques et de la qualité de voix.

Erwan Pépiot[1], Aron Arnold[2]
(1) TransCrit (groupe LeCSeL), 2 rue de la liberté, 93526 Saint-Denis, France
(2) VALIBEL – Université catholique de Louvain, Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique
erwan.pepiot@free.fr, aron.arnold@uclouvain.be

RÉSUMÉ

Cette étude porte sur les productions de locutrices et locuteurs bilingues anglais/français lors d'une tâche de lecture. La fréquence des formants vocaliques (F1, F2, F3) et la différence d'intensité H1-H2 ont été mesurées dans les deux langues. Les résultats indiquent un effet significatif des facteurs *langue* et *genre* sur l'ensemble de ces paramètres. L'analyse des formants montre que les locutrices présentent globalement des valeurs plus élevées que les locuteurs, avec néanmoins des variations inter-langues. Aucune différence inter-genres significative n'a été trouvée sur le F2 du [u] en français, contrairement au [u:] anglais. La différence H1-H2 est significativement plus élevée chez les femmes dans les deux langues, indiquant l'utilisation d'une voix plus *breathy*. Les locutrices présentent une différence H1-H2 moins importante lors de l'emploi du français, quand l'inverse est observé chez les hommes. Ces données suggèrent l'existence de normes vocales dépendantes du genre et de la langue parlée, auxquelles les locuteur·rices bilingues semblent s'adapter.

ABSTRACT

A study of fundamental frequency in female and male English/French bilingual speakers.
The present study deals with the productions of English/French bilingual speakers in a reading task. Vowel formant frequencies (F1, F2 and F3) and H1-H2 difference were investigated in both languages. Results show a significant effect of gender and language on all these parameters. The analysis of vowel formants showed that overall, female speakers exhibited higher values than males. However, a significant cross-gender difference was found on the F2 of the back vowel [u:] in English, but not on the vowel [u] in French. H1-H2 was higher in female speakers in both languages, indicating a more breathy phonation type. Furthermore, female speakers tended to exhibit smaller H1-H2 in French, while the opposite was true in males. This resulted in a smaller cross-gender difference in French for this parameter. These data support the idea of language- and gender-specific vocal norms, to which bilingual speakers seem to adapt.

MOTS-CLÉS : voix et genre, différences inter-genres, formants vocaliques, qualité de voix, H1-H2, bilinguisme, variations inter-langues.

KEYWORDS: voice and gender, cross-gender differences, vowel formants, voice quality, H1-H2, bilingualism, cross-language variation.

Volume 1 : Journées d'Études sur la Parole, pages 480–488.
HAL : hal-02798575.

1 Introduction

Les différences vocales entre femmes et hommes ont fait l'objet d'un nombre important d'études au cours des dernières décennies. Ces recherches portent majoritairement sur la fréquence fondamentale (F0) et les fréquences de résonance (en particulier les formants vocaliques), souvent considérées comme les deux paramètres décisifs. Les fréquences de résonance des voyelles (formants vocaliques) sont généralement plus élevées chez les femmes (Pépiot, 2015) et la F0 des voix d'hommes se situe généralement dans des fréquences plus basses que celles des voix de femmes (Boë et al., 1975). Ces différences acoustiques seraient en partie liées aux différences sexuées qui émergent lors de la puberté sur les appareils phonatoires, (Kahane, 1978 ; Abitbol et al., 1999). En effet, les conduits vocaux sont en moyenne plus longs chez les hommes que chez les femmes adultes et des plis vocaux en moyenne plus longs et plus épais chez les sujets masculins (Kahane, 1978). Cela se traduirait par une F0 et des fréquences de résonance globalement plus élevées dans les productions vocales féminines.

Dans la voix, les facteurs anatomiques et sociaux sont cependant inextricables. Il est en effet établi que les pratiques vocales participent à la construction sociale des identités de genre (Arnold, 2015 ; Pépiot, 2014). Chaque locuteur·ice a un appareil phonatoire d'une forme donnée (qui influence la F0 et les fréquences de résonance produites), mais fait un usage spécifique de cet appareil en fonction de son genre. Une voix n'est ainsi jamais uniquement le reflet d'une anatomie, mais aussi le résultat d'une performance genrée : les femmes mobilisent certaines pratiques articulatoires afin de produire des voix relativement aigues et claires, et les hommes en utilisent d'autres afin de produire des voix relativement graves et sombres (Arnold, 2016).

D'autres paramètres acoustiques, tels que la qualité de voix (ou *type de phonation*) pourraient également être corrélés au genre des locuteur·ice·s. Les voix féminines sont généralement considérés comme plus *breathy* (i.e. ayant un plus grand quotient d'ouverture glottique ou *GOQ*) que les voix masculines (Klatt & Klatt, 1990; Henton & Bladon, 1985). Les voix d'hommes, du moins en anglais américain, ont longtemps été considérées plus *creaky* (i.e. ayant un très faible *GOQ*) que celles des femmes (Henton, 1989a). Cependant, ces résultats varient d'une étude à l'autre, d'une langue à l'autre (Pépiot, 2014), et dépendent du paramètre acoustique utilisé pour estimer la qualité de voix. Selon Gordon & Ladefoged (2001), la différence d'intensité entre le premier et le deuxième harmonique (H1-H2) peut être une mesure fiable, si elle est utilisée de manière adéquate (voir 2.4).

Qu'en est-il alors des locuteur·ice·s bilingues ? Comment s'adaptent-elles/ils aux normes genrées des différentes langues parlées ? Ces questions n'ont pour l'instant fait l'objet que de peu d'attention. S'il existe quelques études qui comparent les productions de locuteur·ice·s de différentes langues (Traunmüller, Eriksson 1995, Pépiot 2014), peu de chercheur·euse·s se sont penché·e·s sur les variations intra-individuelles des productions vocales des bilingues lors du passage d'une langue à l'autre et ont investigué l'influence éventuelle du facteur « genre » dans ces différences. Nous suggérons que l'étude de ces variations inter-langues dans les productions des bilingues permet non seulement de s'éloigner d'une vision statique des voix de femmes et d'hommes, dans laquelle celles-ci sont présentées comme une caractéristique essentielle des locuteur·ice·s et comme principalement due à leur anatomie, pour en adopter une vision dynamique qui permet d'intégrer des questions de différences culturelles, de normes de genre, et de performances du genre.

Quelques études menées sur des locuteur·ice·s bilingues ont montré qu'en fonction de la langue parlée, ces dernier·e·s allaient varier leur F0 moyenne (Altenberg, Ferrand, 2006 ; Lee, Van Lanker Sidtis, 2017). Par exemple, l'étude d'Altenberg et Ferrand (2006) montre que des locutrices bilingues

russes L1 / anglais L2 tendent à parler avec une F0 plus basse en anglais. Cette analyse a cependant été conduite uniquement sur des productions de locutrices – il est donc impossible de savoir si les variations observées relèvent d'une adaptation aux normes genrées des différentes langues, ou tout simplement de pratiques liées aux langues elles-mêmes, sans considération de genre. De plus, ces études ont été uniquement conduites sur la F0, les autres paramètres ayant pour l'heure été complètement négligés. Nous avons donc souhaité investiguer les pratiques de locutrices et de locuteurs bilingues anglais L1 / français L2, en mesurant leurs formants vocaliques et leur qualité de voix, avec l'hypothèse de départ suivante : *les locutrices et les locuteurs bilingues adaptent leurs pratiques vocales aux normes genrées de la langue employée*.

2 Méthode

2.1 Corpus

La présente étude se fonde sur l'analyse d'un corpus de mots dissyllabiques en français et en anglais collectés lors d'une tâche de lecture :

- Combinaisons /C (occlusive) – V – p – i/: /pipi/, /papi/, /pupi/, /bipi/, /bapi/, /bupi/, /tipi/, /tapi/, /tupi/, /dipi/, /dapi/, /dupi/, /kipi/, /kapi/, /kupi/, /gipi/, /gapi/, /gupi/ pour le corpus français, /pi:pi/, /pæpi/, /pu:pi/, /bi:pi/, /bæpi/, /bu:pi/, /ti:pi/, /tæpi/, /tu:pi/, /di:pi/, /dæpi/, /du:pi/, /ki:pi/, /kæpi/, /ku:pi/, /gi:pi/, /gæpi/, /gu:pi/ pour le corpus anglais.
- Combinaisons /C (fricative) – V – p – i/: /sipi/, /sapi/, /supi/, /zipi/, /zapi/, /zupi/, /ʃipi/, /ʃapi/, /ʃupi/, /ʒipi/, /ʒapi/, /ʒupi/ pour le corpus français, /si:pi/, /sæpi/, /su:pi/, /zi:pi/, /zæpi/, /zu:pi/, /i:pi/, /æpi/, /u:pi/, /i:pi/, /æpi/, /u:pi/ pour le corpus anglais.
- Combinaisons /V – p – i /: /ipi/, /api/, /upi/ pour le corpus français, /i:pi/, /æpi/, /u:pi/ pour le corpus anglais.

Il n'existe pas d'accent lexical en Français d'un point de vue phonologique (Hirst & al., 2001), mais au sein de la phrase porteuse utilisée ici (voir 2.3) les locuteur·ice·s ont naturellement accentué la première syllabe des mots du corpus.

2.2 Participant·e·s

Six locutrices et six locuteurs bilingues anglais L1 / français L2 ont pris part à cette étude. Les participant·e·s sont originaires du nord-est des États-Unis et vivent en région parisienne depuis plusieurs années. Tou·te·s font état d'une pratique quotidienne de la langue française et d'un niveau d'aisance dans cette langue supérieur ou égal à 3, sur une échelle allant de 0 à 5, via un questionnaire inspiré par celui de Grosjean (Grosjean & Li, 2013). Nous reprenons également à notre compte la définition du bilinguisme proposée par cet auteur.

Les participant·e·s étaient âgé·e·s de 29 à 54 ans (*SD*=7,6 ans) au moment des enregistrements, pour une moyenne d'âge de 41,8 ans chez les femmes et de 40 ans chez les hommes. Aucun·e n'était fumeur·euse ou ne présentait de troubles de la parole. Une clé USB a été offerte en échange de la participation à la présente étude.

2.3 Procédure d'enregistrement

Les enregistrements se sont déroulés dans une chambre anéchoïque, en utilisant un enregistreur numérique *Edirol R09-HR* de marque *Roland*. Les participant·e·s ont été invité·e·s à lire les mots

dissyllabiques listés dans la section 2.1. Ces derniers étaient présentés sous forme orthographique et placés dans une phrase porteuse afin d'obtenir des paramètres prosodiques constants : "He said '*WORD*' twice" pour le corpus anglais et "Il a dit '*MOT*' deux fois" pour le corpus français. Les locuteur·ice·s ont lu chaque phrase deux fois. Afin de neutraliser de possibles biais liés à l'ordre d'emploi des deux langues (voir Altenberg, Ferrand, 2006), la moitié des participant·e·s a commencé par effectuer ces tâches en langue française, et l'autre moitié en langue anglaise.

2.4 Analyse des données

L'analyse acoustique a été effectuée à l'aide du logiciel *Praat* (Boersma, 2018). Les mesures suivantes ont été réalisées sur les mots dissyllabiques (33 mots * 2 répétitions par locuteur·ice) :

- La fréquence de F1, F2 et F3 (en Hz) sur les voyelles en syllabe initiale. Ces valeurs ont été collectées manuellement en se basant sur l'onde sonore et le spectrogramme des voyelles. Les relevés ont été effectués sur une portion centrale et stable de chaque item.
- La différence d'intensité H1-H2 (en dB) sur les voyelles ouvertes.

Ce dernier paramètre donne une indication de la qualité de voix. L'intensité relative de H1 est corrélée au quotient d'ouverture glottique (QOG) : plus H1 présente une intensité élevée par rapport à H2, plus le QOG est élevé (Klatt & Klaat, 1990). Cependant, la mesure H1-H2 peut uniquement s'effectuer sur les voyelles ouvertes : sur les voyelles moyennes ou fermées, le premier formant (F1) viendrait en effet fausser les résultats (Klatt & Klatt, 1990). Ainsi, seules les voyelles [a] pour le corpus français et [æ] pour le corpus anglais ont été prises en compte. Une sélection d'environ cinq cycles a été effectuée sur une portion centrale de chaque voyelle ouverte. Le spectre correspondant était ensuite affiché et la différence d'intensité H1-H2 calculée manuellement. Les données ainsi recueillies ont ensuite fait l'objet de tests statistiques de type ANOVA, dans le but de tester l'influence des facteurs « *langue parlée* » et « *genre des locuteur·rice·s* ».

3 Résultats

3.1 Formants vocaliques

Les fréquences moyennes des formants des voyelles françaises (F1, F2, F3) produites par les locutrices et les locuteurs sont présentées dans le Tableau 1 ci-après (22 mesures par formant et par voyelle pour chaque locuteur·ice). Sans surprise, les valeurs formantiques sont en moyenne plus élevées chez les locutrices. Cependant, l'ampleur de cette différence inter-genre varie fortement en fonction de la voyelle et du formant considéré. Le ratio femme/homme est maximal sur le F1 de la voyelle [a] (+39% par rapport aux locuteurs), et sur le F2 de la voyelle [i] (+31%). A l'inverse, le deuxième formant de la voyelle [u] présente des valeurs très similaires dans les deux groupes (+5% chez les femmes).

Une ANOVA à deux facteurs (« *voyelle* » et « *genre des locuteur·ice·s* ») a été conduite sur le F1 des voyelles françaises. Il en ressort un effet global important et significatif du facteur *genre* : $F(1,786)=459,064$; $p<0,0001$. L'analyse détecte également une interaction significative entre les deux facteurs, montrant ainsi que l'influence du facteur *genre* dépend de la *voyelle* prise en considération : $F(2,786)=208,663$; $p<0,0001$. Pour chaque voyelle prise individuellement, l'effet du *genre* est significatif ($p<0,0001$ dans tous les cas). Des tests similaires ont été effectués pour F2. Un effet global important et significatif a également été observé pour le facteur *genre* ($F(1,786)=794,071$; $p<0,0001$) ainsi qu'une interaction avec le facteur *voyelle* ($F(2,786)=240,866$; $p<0,0001$). Dans le

détail, on constate que la différence inter-genres est significative sur [i] et [a] ($p<0,0001$) mais pas sur le F2 de la voyelle arrière [u] ($F(1,262)=3,333$; $p=0,069$).

FREQUENCE DES FORMANTS VOCALIQUES SUR LES VOYELLES FRANÇAISES (HZ)

	Loc.	F1	F2	F3	F4	F5	F6	Moy. F	H1	H2	H3	H4	H5	H6	Moy. H	Ratio F/H
	F1	1027	721	696	733	768	791	**789**	602	571	592	563	551	529	**568**	**1.39**
	σ	*47*	*52*	*77*	*36*	*62*	*53*	***124***	*48*	*51*	*72*	*40*	*43*	*36*	***55***	
[a]	**F2**	1885	1913	1797	1713	1814	1704	**1804**	1547	1584	1606	1569	1453	1581	**1557**	**1.16**
	σ	*134*	*109*	*78*	*91*	*127*	*89*	***131***	*62*	*84*	*71*	*89*	*56*	*97*	***91***	
	F3	2847	2684	2690	2865	2879	2781	**2791**	2248	2479	2521	2445	2563	2615	**2478**	**1.13**
	σ	*66*	*128*	*135*	*53*	*92*	*72*	***124***	*46*	*82*	*124*	*100*	*54*	*84*	***144***	
	F1	278	354	299	326	261	272	**298**	298	248	318	256	267	260	**274**	**1.09**
	σ	*8*	*37*	*16*	*15*	*9*	*16*	***38***	*18*	*6*	*20*	*12*	*14*	*7*	***29***	
[u]	**F2**	1085	997	1268	920	1250	862	**1064**	959	1021	1112	943	1074	987	**1016**	**1.05**
	σ	*139*	*136*	*361*	*49*	*231*	*71*	***246***	*135*	*172*	*159*	*117*	*218*	*158*	***171***	
	F3	2443	2343	2512	2638	2753	2821	**2585**	2109	2266	2137	1991	2280	2337	**2187**	**1.18**
	σ	*184*	*136*	*135*	*138*	*129*	*94*	***217***	*422*	*55*	*98*	*102*	*111*	*70*	***222***	
	F1	284	349	292	332	262	264	**297**	280	244	300	242	242	250	**260**	**1.14**
	σ	*18*	*48*	*12*	*15*	*13*	*14*	***40***	*11*	*5*	*19*	*8*	*24*	*13*	***27***	
[i]	**F2**	2783	2504	2572	2677	2619	2752	**2651**	2092	2117	1962	1960	2020	2021	**2029**	**1.31**
	σ	*45*	*127*	*110*	*109*	*46*	*65*	***132***	*53*	*75*	*49*	*57*	*77*	*73*	***87***	
	F3	3815	3733	3395	3593	3446	3537	**3587**	3176	3037	2940	3023	2836	2971	**2997**	**1.20**
	σ	*140*	*183*	*193*	*186*	*140*	*194*	***227***	*57*	*137*	*140*	*131*	*120*	*108*	***156***	

TABLE 1: *Fréquences formantiques (Hz) de F1, F2 et F3 sur les voyelles françaises [i] [a] et [u] produites par les locutrices (F) et les locuteurs (M). L'écart-type (SD) est mentionné en italique.*

Une analyse identique a été menée sur les valeurs de F3. Elle indique là encore un effet global très significatif du facteur *genre* ($F(1,786)=1072,078$; $p<0,0001$) et une interaction significative avec le facteur *voyelle* ($F(2,786)=38,164$; $p<0,0001$). La différence inter-genres est significative pour chaque voyelle prise séparément ($p<0,0001$). A l'instar des résultats obtenus sur le français, les valeurs formantiques sur les voyelles anglaises sont globalement plus élevées chez les locutrices, avec des variations notables en fonction la voyelle et du formant pris en considération (voir Tableau 2, ci-après). Le ratio femme/homme est maximal sur le F2 de la voyelle [æ] (fréquence 37% plus élevée chez les femmes) et sur le deuxième formant de la voyelle [i:] (+30%). Mais cette fois-ci, contrairement au français, on observe également une très large différence inter-genres sur le F2 de la voyelle arrière [u:] (+26% chez les femmes).

Une ANOVA à deux facteurs (« *genre des locuteur·ice·s* » et « *voyelle* ») a été conduite sur les valeurs de F1 pour ces voyelles anglaises. Globalement, l'effet du facteur *genre* est très significatif : $F(1,786)=451,643$; $p<0,0001$. Ce test met également en évidence une interaction entre les facteurs *genre* et *voyelle* : $F(2,786)=227,075$; $p<0,0001$. La différence inter-genres sur F1 est significative pour chaque voyelle prise individuellement (avec $p<0,0001$). La même analyse a été menée sur les valeurs de F2. Ici encore, un effet très fort et très significatif du facteur *genre* a été observé ($F(1,786)=942,947$; $p<0,0001$), ainsi qu'une interaction avec le facteur *voyelle* ($F(2,786)=240,866$; $p<0,0001$). Contrairement au français, la différence inter-genre est ici significative sur chaque voyelle prise séparément, y compris la voyelle arrière [u:] ($p<0,0001$). L'ANOVA à deux facteurs a également été conduite sur les valeurs de F3. L'effet du facteur *genre* est là encore très significatif ($F(1,786)=1406,561$; $p<0,0001$), et il existe également une interaction avec le facteur *voyelle* ($F(2,786)=39,889$; $p<0,0001$). En considérant individuellement chaque voyelle, on constate que la différence inter-genres est significative dans tous les cas ($p<0.0001$).

FREQUENCE DES FORMANTS VOCALIQUES SUR LES VOYELLES ANGLAISES (HZ)

	Loc.	F1	F2	F3	F4	F5	F6	Moy. F	H1	H2	H3	H4	H5	H6	Moy. H	Ratio F/H
[æ]	**F1**	1076	774	721	881	851	861	**861**	627	635	608	724	574	606	**629**	**1.37**
	σ	*51*	*61*	*75*	*66*	*43*	*78*	***128***	*23*	*41*	*47*	*38*	*70*	*33*	***64***	
	F2	1899	1924	1876	1690	1872	1725	**1831**	1561	1575	1538	1546	1554	1617	**1565**	**1.17**
	σ	*134*	*88*	*73*	*66*	*74*	*82*	***126***	*53*	*62*	*35*	*44*	*80*	*50*	***60***	
	F3	2841	2727	2686	2651	2944	2807	**2776**	2302	2388	2530	2471	2515	2547	**2459**	**1.13**
	σ	*80*	*96*	*146*	*121*	*97*	*85*	***145***	*58*	*68*	*67*	*93*	*57*	*69*	***111***	
[u:]	**F1**	313	334	314	343	291	307	**317**	302	257	347	276	279	293	**292**	**1.08**
	σ	*46*	*21*	*18*	*8*	*8*	*20*	***29***	*15*	*12*	*21*	*13*	*17*	*16*	***32***	
	F2	1218	1598	1495	1638	1601	1281	**1472**	995	1171	1391	987	1249	1234	**1171**	**1.26**
	σ	*187*	*252*	*330*	*279*	*265*	*230*	***304***	*191*	*215*	*208*	*134*	*227*	*228*	***246***	
	F3	2645	2576	2622	2737	2810	2779	**2695**	2280	2227	2198	2104	2259	2339	**2235**	**1.21**
	σ	*80*	*77*	*116*	*114*	*95*	*116*	***131***	*59*	*53*	*79*	*95*	*92*	*73*	***105***	
[i:]	**F1**	309	317	311	324	290	291	**307**	274	252	341	256	255	276	**276**	**1.11**
	σ	*23*	*35*	*21*	*7*	*10*	*15*	***24***	*11*	*23*	*22*	*26*	*16*	*10*	***36***	
	F2	2792	2675	2592	2873	2633	2804	**2728**	2165	2147	2029	2110	2048	2100	**2100**	**1.30**
	σ	*95*	*138*	*107*	*75*	*80*	*101*	***142***	*50*	*57*	*93*	*61*	*97*	*63*	***86***	
	F3	3704	3818	3256	3394	3245	3386	**3467**	3074	2998	2760	3028	2703	2757	**2887**	**1.20**
	σ	*164*	*225*	*103*	*116*	*96*	*251*	***275***	*84*	*146*	*123*	*135*	*111*	*95*	***189***	

TABLE 2 : *Fréquences formantiques (Hz) de F1, F2 et F3 sur les voyelles anglaise [i:] [æ] et [u:] produites par les locutrices (F) et les locuteurs (M). L'écart-type (SD) est mentionné en italique.*

Afin de rendre ces résultats plus lisibles, les valeurs de F1 et de F2 sont présentées sous la forme de triangles vocaliques (Figure 2) obtenus à l'aide du logiciel SaRP (Nikolov et al. 2011).

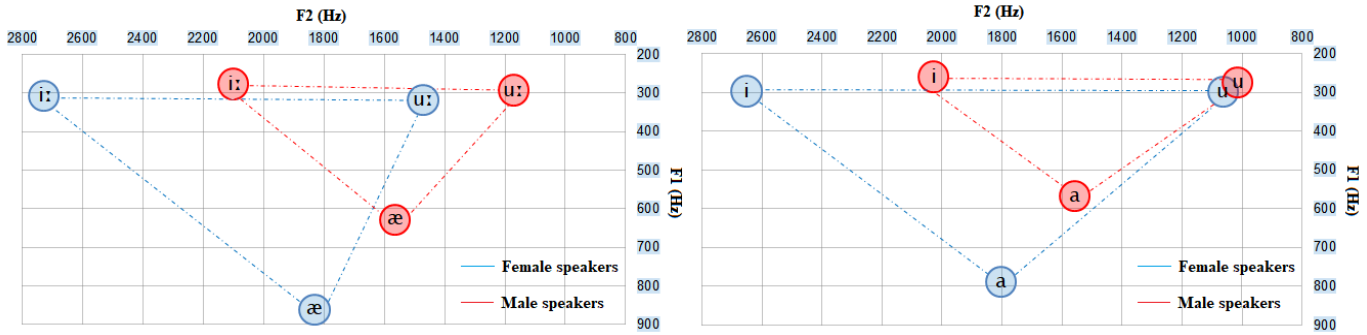


FIGURE 2 : *Triangles vocaliques représentant les valeurs formantiques (Hz) des voyelles anglaise [i:] [æ] et [u:] (à gauche) et françaises [i], [a], [u] (à droite) chez les locutrices (en bleu) et les locuteurs bilingues (en rouge).*

3.2 Qualité de voix

La différence d'intensité moyenne (dB) entre le premier (H1) et le deuxième harmonique (H2) en fonction du genre des locuteur·rice·s et de la langue parlée est présentée dans le Tableau 10. Elle a été mesurée sur les voyelles ouvertes, pour un total de 22 occurrences par locteur·rice dans chaque langue (11 mots avec [a] pour le français et 11 mots contenant la voyelle [æ] pour l'anglais, chaque mot étant prononcé deux fois). L'intensité relative de H1 est corrélée au Quotient d'Ouverture Glottique (QOG) : plus H1 est intense, plus le QOG est élevé. Ainsi, une valeur de H1-H2 élevée indique une voix *breathy*, alors qu'une valeur fortement négative est associée à une voix *creaky*.

	H1-H2 MOYEN SUR VOYELLES OUVERTES (dB)		
Loc.	**Anglais**	**Français**	**Diff. FR-AN**
F1	7,60	7,77	**0,16**
F2	2,75	1,58	**-1,17**
F3	7,48	5,62	**-1,86**
F4	4,50	4,30	**-0,20**
F5	10,13	9,66	**-0,47**
F6	4,71	4,66	**-0,05**
Moy. F.	**6,20**	**5,60**	**-0,60**
H1	-3,96	-2,52	**+1,45**
H2	-1,59	-0,08	**+1,51**
H3	1,52	2,42	**+0,90**
H4	1,54	2,07	**+0,54**
H5	6,85	9,09	**+2,25**
H6	0,53	0,75	**+0,22**
Moy. H.	**0,81**	**1,96**	**+1,14**

TABLE 3 : *Valeur moyenne de H1-H2 (dB) chez les locutrices (F) et les locuteurs (H) sur les voyelles ouvertes (11 x 2 occurrences) en fonction de la langue parlée. La différence entre les valeurs obtenues en français et en anglais est indiquée dans la colonne de droite.*

La différence d'intensité H1-H2 est plus élevée chez les locutrices dans les deux langues. On observe également que toutes les locutrices sauf une présentent une valeur H1-H2 plus faible en français, alors que l'inverse est vrai pour les locuteurs masculins. Ainsi, la différence inter-genres atteint 5,39 dB en anglais, mais n'est que de 3,64 dB en français. Une ANOVA à un facteur (« *genre* ») a été réalisée sur les données obtenues en français. L'analyse révèle un effet significatif de ce facteur : $F(1,262)=129,133$; $p<0,0001$. La même analyse conduite sur l'anglais met également en évidence un effet significatif du facteur *genre* : $F(1,262)=61,273$; $p<0,0001$. Ces résultats confirment que, dans les deux langues, la différence H1-H2 est significativement plus élevée chez les locutrices et suggèrent que les ces dernières ont utilisé un type de phonation plus *breathy* que les locuteurs.

Afin de vérifier si la variation inter-langues est significative, une ANOVA à deux facteurs (« *langue parlée* » et « *genre* ») a été menée sur l'ensemble des données (français et anglais). Cette analyse révèle bien une interaction significative entre les deux facteurs : $F(1,524)=6,871$; $p<0,01$. Cela confirme que l'adaptation de la qualité de voix lors du changement de langue parlée varie en fonction du genre des locuteur·rice·s. D'où l'abaissement des valeurs de H1-H2 en français chez les locutrices et leur augmentation chez les locuteurs, faisant apparaître une différence inter-genres moins prononcée en français sur ce paramètre.

4 Conclusion / discussion

Cette analyse multiparamétrique et inter-langues a permis d'obtenir de nombreux résultats intéressants, qu'il convient de mettre en perspective.

L'analyse des formants vocaliques (F1, F2, F3) a sans surprise mis en évidence des fréquences plus élevées chez les locutrices dans les deux langues. Sur l'anglais, cette différence inter-genres est maximale pour le F1 de la voyelle [æ] (+37%) et sur le F2 du [i :] (+30%). En français, les différences les plus marquées apparaissent sur le F1 de la voyelle [a] et sur le F2 du [i] (+31%). Cela pourrait laisser penser que les différences inter-genres sur ce paramètre acoustique sont relativement similaires dans les deux langues.

Cependant, pour ce qui concerne les voyelles arrières [u] (en français) et [u:] (en anglais), un phénomène particulièrement remarquable a été observé. Alors qu'en anglais, une très large différence inter-genres a été mesurée sur le F2 de [u:] (+26% chez les femmes), aucune différence significative n'est apparu sur le F2 de [u] (seulement +5% chez les femmes). Cela semble parfaitement cohérent avec les résultats obtenus par Pépiot (2015) sur des monolingues francophones et anglophones américain·e·s. L'explication physiologique traditionnellement avancée pour rendre compte des différences inter-genres sur les formants vocaliques (reposant sur la taille des conduits vocaux) s'avère donc inopérante. A l'inverse, des pratiques vocales genrées sont vraisemblablement à l'origine des différences observées ici. Les locutrices ont pu par exemple accentuer la protrusion des lèvres lors de la production de la voyelle [u] en français afin d'abaisser la valeur du F2 (Fant 1966), et ont probablement utilisé une position de la langue relativement centrale lors de l'articulation de la voyelle anglaise [u:], obtenant ainsi un F2 particulièrement élevé.

La mesure H1-H2 sur les voyelles ouvertes a fourni des indications précieuses sur le type de phonation utilisé par les locuteur·rice·s. Des différences inter-genres significatives ont été trouvées dans les deux langues : les locutrices présentent des valeurs plus élevées que les locuteurs, ce qui suggèrent qu'elles utilisent une qualité de voix plus *breathy*. Cependant, une interaction significative entre les facteurs « *langue parlée* » et « *genre des locuteur·rice·s* » a été observée, montrant ainsi que l'adaptation du type de phonation lors du passage d'une langue à l'autre est dépendante du genre. Les locuteurs présentent tous des valeurs de H1-H2 plus faibles en anglais, quand, à l'inverse, les locutrices ont eu tendance à présenter des valeurs plus élevées dans cette langue. La différence inter-genres est donc nettement plus prononcée en anglais qu'en français.

Ces résultats vont dans le sens des études de Henton (1989a) et Pépiot (2014) qui présentent les différences inter-genres sur le type de phonation comme un phénomène socio-phonétique et dépendant de la langue. Ils sont également cohérents avec les données recueillies dans des expériences perceptives : dans une étude inter-langues menée sur des auditeur·rice·s francophones et anglophones, Pépiot (2017) a montré que le type de phonation (mesuré via H1-H2) était un indice acoustique important pour l'identification du genre par la voix chez les anglophones, alors que ce paramètre n'était pas pertinent pour catégoriser les voix chez les auditeur·rice·s francophones.

L'analyse de ces différents paramètres acoustiques, ainsi que l'étude simultanée des facteurs *langue* et *genre* sur les productions de bilingues apporte ainsi de nouveaux éléments qui n'avaient pas pu être dégagés des études précédentes, comme celles d'Altenberg et Ferrand (2006), Mennen et al. (2012) ou Lee et Van Lanker Sidtis (2017). En effet, la présente étude montre clairement que la production d'une voix de femme ou d'homme mobilise des pratiques vocales différentes en fonction de la langue parlée. Elle confirme également que les fréquences de résonance des voyelles et la qualité de voix ne sont pas des caractéristiques essentielles et figées des locuteur·ice·s, liées uniquement à l'anatomie de leur appareil phonatoire, mais qu'elles résultent, pour une part importante, d'un apprentissage et d'une socialisation en tant que membre d'une catégorie de genre donnée. Plus largement, cela constitue un argument supplémentaire pour s'éloigner des conceptions simplistes sur les liens entre voix et anatomie, et pour une plus grande considération des facteurs sociaux dans les études phonétiques.

Parmi les limites de cette étude, on citera notamment le nombre relativement réduit de participant·e·s. Afin de confirmer nos résultats, cette expérience pourrait être répliquée avec un nombre plus important de locuteur·ice·s. D'autre part, nous avons ici uniquement analysé de la parole produite par des bilingues anglais L1 / français L2. Il serait intéressant de recouper ces résultats en étudiant également la parole de bilingues français L1 / anglais L2.

Références

ABITBOL J., ABITBOL P., ABITBOL B. (1999). Sex hormones and the female voice. *Journal of Voice,* 13, 424–446.

ALTENBERG E. P., FERRAND C. T. (2006). Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice,* 20, 89-96.

ARNOLD A. (2015). Voix et transidentité : changer de voix pour changer de genre ?. *Langage et société* 151(1), 87-105.

ARNOLD A. (2016). Voix. *Encyclopédie critique du genre*, 713-721. Paris : La Découverte.

BOERSMA P., WEENINK D. (2018). Praat: doing phonetics by computer [Logiciel]. Version 6.0.40, publiée le 11 mai 2018 sur le site www.praat.org

BOË L.-J., CONTINI M., RAKOTOFIRINGA H. (1975). Etude statistique de la fréquence laryngienne. *Phonetica,* 32(1), 1-23.

FANT, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scaling. *Speech Transmission Laboratory, Quarterly Progress and Status Report*, 81, 21-38.

GORDON, M. & LADEFOGED, P. (2001). Phonation types: A crosslinguistic overview. *Journal of Phonetics*, 29, 383-406.

GROSJEAN, F. & LI, P. (2013). *The Psycholinguistics of Bilingualism*. Oxford : Wiley-Blackwell.

HENTON, C. (1989a). Sociophonetic aspects of creaky voice, *Journal of the Acoustical Society of America*, 86, S26.

HENTON C. (1989b). Fact and fiction in the description of female and male pitch. *Language & Communication*, 9(4), 299-311.

HENTON, C. & BLADON, R. (1985). Breathiness in normal female speech: Inefficiency versus desirability, *Language and Communication*, 5, 221-227.

HIRST, D., DI CRISTO, A., NISHINUMA, Y. (2001). Prosodic parameters of French: A cross-language approach. In *Contrastive Studies of Japanese and Other Languages Series*, National Institute for Japanese Language, Tokyo, 7-20.

KAHANE J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy* 151, 11-19.

KLATT, D. H. & KLATT, L. C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820-857.

LEE B., VAN LANCKER SIDTIS D. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing* 20(3), 174-185.

NIKOLOV, R., DOMMERGUES, J-Y. & RYST, E. (2007). SaRP : outil polyvalent de représentations multi-points et multi-séries des formants vocaliques. *Paisii Hilendarski University of Plovdiv, Bulgaria – Research Papers*, 45, 88-95.

PÉPIOT E. (2014). Male and female speech: a study of mean F0, F0 range, phonation type and speech rate in Parisian French and American English speakers. *Proceedings of the 7th International Conference on Speech Prosody*, 305-309.

PÉPIOT, E. (2015). Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. *Corela*, HS-16.

PÉPIOT, E. (2017). Gender identification from speech in Parisian French and American English speakers. *Paisii Hilendarski University of Plovdiv, Bulgaria – Research Papers*, 55, 60-72.

TRAUNMÜLLER H., ERIKSSON A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Manuscrit : http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Corrélats acoustiques et perceptifs de la personnalité perçue à travers la voix dans une population de dysphoniques légères

Amelia Pettirossi[1] Nicolas Audibert[1] Lise Crevier-Buchman[1,2]
(1) Laboratoire de Phonétique et Phonologie : UMR 7018 CNRS / Sorbonne Nouvelle, 75005 Paris, France
(2) Hôpital Foch : Service de Laryngologie Phoniatre, 92150 Suresnes, France
prenom.nom@sorbonne-nouvelle.fr

RÉSUMÉ

Nous étudions les corrélats acoustiques et perceptifs de la personnalité à travers la voix dans une population de dysphoniques légères (G1 à G2) et de locutrices témoins (G0). 40 auditeurs naïfs ont évalué les voix de 61 femmes. Des échelles sémantiques différentielles ont été utilisées pour la cotation de la sévérité du trouble vocal et des traits de personnalité. Les 5 échelles sont : Joyeuse/Triste, Sympathique/Désagréable, Dynamique/Molle, Confiante/Hésitante, Aucun trouble vocal/Trouble vocal sévère. Le jugement de la pathologie vocale par les naïfs est principalement lié à l'évaluation experte du grade de dysphonie faite à partir du GRBAS. Des traits de personnalité plus négatifs sont attribués aux locutrices perçues comme plus dysphoniques. Certains facteurs acoustiques (f0, débit syllabique, HNR et ZCR) semblent influencer les auditeurs : les voix plus aigües, plus rauques et avec un débit rapide sont associées à un jugement plus positif.

ABSTRACT

Acoustical and perceptual correlates of perceived personality through voice in minor dysphonia
Our study deals with acoustical and perceptual correlates of perceived personality through voice in a population affected by minor dysphonia (G1 to G2) and controls (G0). 40 naive listeners assessed the voices of 61 women. Semantic differential scales were used to rate the severity of voice disorders and personality traits. The 5 scales are: Happy/Sad, Nice/Nasty, Dynamic/Apathetic, Confident/Uncertain, No vocal disorder at all/High vocal disorder. Vocal disorder ratings by naïve listeners are mainly related to the grade of dysphonia assessed by an expert on the GRBAS scale. Women perceived as more dysphonic are judged more harshly on all personality traits. Some acoustical measures (f0, syllabic speech rate, HNR and ZCR) seem to influence our listeners: higher-pitched, hoarser voices and higher speech rates are associated to a more positive judgment.

MOTS-CLÉS : voix, personnalité, dysphonie, acoustique, perception
KEYWORDS : voice, personality, dysphonia, acoustics, perception

Volume 1 : Journées d'Études sur la Parole, pages 489–497.
HAL : hal-02798576.

1 Introduction

1.1 Cadre théorique

La voix semble avoir un rôle important dans les représentations de la personnalité. Elle est porteuse de diverses informations et influe sur l'image que peut avoir un individu de son interlocuteur. Des auditeurs naïfs sont capables de reconnaitre avec précision des indices de personnalité dans des échantillons vocaux synthétisés sur ordinateur et donc non-naturels. La manipulation de mesures acoustiques comme la fréquence fondamentale ou l'intensité est perçue par les auditeurs et les amène à juger différemment les voix sur le degré supposé d'extraversion ou d'introversion du locuteur (Nass & al, 2001).

L'effet de la voix sur la personnalité perçue est directement ancré dans la culture du locuteur et de l'auditeur. Bien que certains indices puissent être partagé et même supposés universels (Montepare & al, 1987), certains autres amènent un jugement plus ou moins péjoratif selon les pays. C'est le cas d'un débit de parole lent, perçu négativement par les Américains et non par les Coréens (Peng & al, 1993). Aux Etats-Unis, un débit de parole élevé est associé à des traits de personnalité positifs : plus de dynamisme, de puissance et d'extraversion (Apple & al, 1979). Généralement, ces critères ne sont donc pas valables d'une culture à une autre, ni même dans le temps pour une société donnée.

Les questionnements sur le jugement de la personnalité à travers la voix ne peuvent pas être dissocié des influences culturelles et de la volonté de marquer une appartenance à un groupe. La théorie de l'accommodation communicative (Giles & al, 1991 cités par Barkat-Defradas & al, 2007) illustre le besoin du sujet parlant de se rapprocher d'une norme vocale avec ses interlocuteurs. Le locuteur naïf détecte des indices acoustiques chez l'autre et tente de les copier pour adapter sa voix. Cette convergence phonétique augmente au cours du discours sur certains facteurs comme la hauteur de voix (Pardo, 2006).

L'idée selon laquelle « *What sounds beautiful is good* » (Zuckerman & al, 1988) est largement mise en évidence dans les études sur le jugement vocal. La photo d'un même sujet peut être évaluée de manière significativement plus négative lorsque la voix accompagnant l'image est dysphonique que lorsqu'elle est saine (Blood & al, 1979). La photo associée à une voix non pathologique bénéficie de l'effet halo (Thorndike, 1920). Qu'elle soit négative ou positive la première impression donnée par une voix va influencer le jugement d'autres critères comme le physique, ou encore la personnalité. Un sujet ayant pour tâche de compléter le profil d'un personnage fictif va directement être influencé par les premiers traits de personnalité donnés par l'expérimentateur (Asch, 1946). Il sera donc plus enclin à lui construire un profil négatif si l'expérimentateur lui donne un trait péjoratif comme « froid » plutôt que la modalité inverse « chaleureux ».

De manière générale, les voix dysphoniques induisent un jugement de personnalité beaucoup plus sévère que les voix saines, et cela est encore plus visible chez les femmes (Amir & al, 2013). En revanche, tout ce qui est pathologique n'est pas nécessairement catégorisé comme mauvais. Il a déjà été démontré que dans certaines formes légères, une raucité chez l'homme peut être considérée comme séduisante pour des auditrices alors que cette même raucité est considérée comme pathologique par un expert (Barkat-Defradas & al, 2012).

1.2 Questions de recherche

Dans cet article nous confrontons à nos données les trois interrogations suivantes :

- Quelles dimensions du GRBAS mènent les naïfs à catégoriser les voix comme pathologiques ?
- Quelle est l'incidence du trouble dysphonique sur le jugement de personnalité ?
- Quels indices acoustiques les auditeurs exploitent-ils pour juger la personnalité ?

2 Méthodologie

2.1 Locutrices, corpus et prises de données

Nous avons recueilli nos données auprès de 61 femmes françaises, professeures des écoles (PE) en activité. La prévalence de dysphonies dans cette population, majoritairement féminine, est très élevée (INSERM, 2006). En France, d'après des questionnaires proposés par la MGEN en 2014, 59% des professeurs des écoles interrogés affirment avoir déjà eu des problèmes vocaux (Caetano & al., 2017). En 2017, dans le cadre de nos travaux (non publiés), nous avons observé que 80% d'une population de 709 femmes PE françaises auto-déclaraient des troubles vocaux.

Les participantes ont été enregistrées à partir de la station Computerized Speech Lab 4500 de KayPENTAX. Un micro-casque AKG C 410 a été positionné à environ 5 cm des lèvres de chaque locutrice. Le premier paragraphe de « La bise et le soleil » en condition de lecture « neutre » a été utilisé pour cette étude.

Une cotation par GRBAS (Hirano, 1981) est réalisée par un expert. Cet outil d'évaluation clinique permet de quantifier le grade de dysphonie (G), la raucité (R), le souffle (B), l'asthénie (A) et le serrage vocal (S). Chaque critère peut être noté de 0 (aucun trouble) à 3 (trouble sévère). Pour ce qui est de notre population, elle se compose de 2 locutrices de grade 2, 22 de grade 1 et 37 de grade 0.

2.2 Attribution de traits de personnalité

Une revue de littérature traitant de la personnalité à travers la voix (Kreiman et al, 2011), met en évidence l'utilisation, par la grande majorité des études, de deux méthodologies principales : les cinq facteurs du Big-5 (McCrae & al, 1985) et tout ou partie des 50 échelles sémantiques différentielles originellement proposées par Osgood (Osgood, 1952). Pour notre part, nous conservons, le principe de l'échelle sémantique différentielle, qui permet d'offrir une représentation simple et intuitive aux auditeurs (Aronovitch, 1976 ; Ruscello & al, 1988). Afin de proposer un choix d'échelles adapté aux spécificités de nos données, nous optons pour une méthodologie dans laquelle elles sont construites à partir d'un ensemble de traits de personnalité attribuées aux locutrices par un panel d'auditeurs.

2.2.1 Etude pilote 1 : Catégorisation libre

Notre première étude pilote consiste à mettre en place une catégorisation libre. Il s'agit de faire écouter tous les stimuli à des auditeurs naïfs et de les laisser choisir librement tous les adjectifs qu'ils souhaitent pour décrire la personnalité de la personne dont ils viennent d'entendre la voix. Sans être précisément décrite, cette étape de travail a déjà été utilisée dans l'étude de la personnalité à travers la voix dysphonique (Amir & al, 2013). Cette étude pilote a pour but de repérer les adjectifs les plus utilisés pour décrire les locutrices de notre panel afin de créer nos échelles sémantiques différentielles.

Ce test a été codé avec le logiciel PsychoPy (Peirce & al, 2019) puis mis en ligne sur la plateforme https://pavlovia.org/ afin de faciliter sa diffusion. L'interface est simple, l'auditeur entend une voix, puis, donne un ou plusieurs adjectifs de traits de personnalité qu'il pense adéquat pour décrire cette femme (et ainsi de suite pour les 61 locutrices) sans pouvoir réécouter le son. Dans chaque test les stimuli sont diffusés aléatoirement. Parmi les 61 stimuli, 2 ont été dupliqué afin de vérifier qu'il y ait un minimum de consistance dans le jugement des auditeurs.

29 auditeurs (9 hommes et 20 femmes d'une moyenne d'âge de 26.5 ans), tous locuteurs natifs du français ont participé. Nous n'avons finalement pas exclu de participants car même si les adjectifs n'étaient pas nécessairement identiques dans le test-retest, des antonymes n'ont jamais été utilisés pour décrire la même locutrice.

Nous obtenons donc 2557 adjectifs. Ainsi commence un pré-traitement manuel visant à uniformiser les réponses, sans qu'aucun choix subjectif ne soit encore fait, pour éliminer les fautes d'orthographe, de frappe ou encore les abréviations comme « sympa ». Nous avons donc 583 réponses uniques que nous appellerons « modalités ». Un deuxième travail de recodage est alors réalisé afin de regrouper dans une même modalité, différents synonymes. Pour illustration, nous obtenons des regroupements comme : « sympathique », « aimable » et « gentille » dans la modalité « sympathique » (cet adjectif donne son nom à la modalité car il est le plus fréquent).

2.2.2 Etude pilote 2 : Validation des termes proposés

Dans une seconde enquête en ligne, des locuteurs natifs du français sont invités à donner leur avis sur la synonymie ainsi que sur l'antonymie de certains adjectifs. Si trop de regroupements sont considérés comme abusifs, il est possible qu'une modalité ne fasse finalement pas l'objet d'une échelle sémantique différentielle pour l'expérimentation finale car elle ne sera plus assez fréquente. Dans le cas où les synonymes sont acceptés il nous faut trouver l'antonyme le plus correct pour les échelles finales. Les antonymes sont proposés par l'expérimentateur dans le cas où ils ne figurent pas déjà dans la liste des modalités les plus fréquentes. Si la paire est considérée comme non-antonyme, le participant peut donner l'adjectif de remplacement qui lui parait le plus correct.

120 natifs ont participé à notre enquête. Ils sont démarchés au hasard sur les réseaux sociaux, certains peuvent avoir des notions de linguistique, mais ce n'est majoritairement pas le cas. Bien que certaines synonymies ne soient pas validées par le panel (le seuil d'acceptabilité est fixé à 70%), les modalités sélectionnées restent celles les plus fréquemment utilisées par les auditeurs de la 1ère étude pilote. Pour ce qui est des antonymes, ils sont validés à plus de 84%, à l'exception d'un, qui est échangé pour celui le plus souvent proposé en remplacement par les sujets. Ainsi, la paire « confiante – anxieuse » devient « confiante – hésitante ». Nous obtenons alors les 4 échelles suivantes : « Joyeuse – Triste », « Sympathique – Désagréable », « Dynamique – Molle » et « Confiante – Hésitante ».

2.2.3 Expérimentation finale : Catégorisation des voix sur échelles sémantiques différentielles

L'expérimentation finale a été réalisée sur Praat (Boersma & al, 2019). L'auditeur écoute le même extrait de parole pour les 61 locutrices. Il est demandé de noter chaque voix sur des échelles à 5 niveaux non-continus, allant de 0 (le plus positif) à 4 (le plus négatif). En plus des quatre dimensions de la personnalité précédemment établie, une échelle visant à évaluer le trouble vocal de chaque locutrice a été ajoutée (« Aucun trouble vocal – Trouble vocal sévère »). Ce score supplémentaire permettra la comparaison de la perception du trouble par les auditeurs naïfs avec le GRBAS expert.

Les auditeurs ont la possibilité de réécouter le son jusqu'à 5 fois (soit une fois par échelle si nécessaire). Chaque test est généré aléatoirement pour que chaque auditeur soit confronté à un ordre de présentation des stimuli, des échelles, ainsi que du côté sur lequel est indiqué l'adjectif positif ou négatif différent. Ici encore, une duplication de deux des stimuli est mise en place afin de vérifier si nous observons bien une consistance dans le jugement des auditeurs.

40 auditeurs (18 hommes et 22 femmes d'une moyenne d'âge de 36 ans) ont passé l'expérimentation finale. 11 de ces 40 auditeurs sont également des participants de la catégorisation libre (étude pilote 1). Ainsi, nos analyses se basent sur 2520 notes (soit 63 réponses pour chacun des 40 auditeurs) pour chaque échelle à partir desquelles nous calculons un score moyen unique pour chaque trait de personnalité de toutes les locutrices. Pour ce qui est du test-retest, nos sujets semblent avoir un jugement assez constant car nous observons une différence de score moyen allant d'un minimum de 0.025 à un maximum de 0.325 sur 4. Les voix de ces deux locutrices ne sont donc pas jugées de manière foncièrement opposée lors des deux écoutes.

3 Résultats

Les analyses portent sur les résultats des auditeurs hommes et femmes car nous n'avons pas d'effet du sexe sur le jugement de personnalité des locutrices.

3.1 Quelles dimensions du GRBAS mènent les naïfs à catégoriser les voix comme pathologiques ?

Pour répondre à notre première question nous avons réalisé des corrélations de Spearman entre les cotations des différentes dimensions évaluées dans le GRBAS et les scores moyens de trouble vocal donnés par les auditeurs naïfs.

On observe une corrélation modérée entre le diagnostic expert du Grade de dysphonie de l'échelle GRBAS et le jugement du trouble vocal par les naïfs (voir Table 1). Nous pouvons également constater que certains autres critères du GRBAS ont un impact faible à modéré sur la catégorisation du trouble vocal par les auditeurs. Nous ne présentons pas ici les résultats de l'asthénie car toutes nos locutrices sont cotées 0 sur ce facteur.

Dimension du GRBAS	Coefficient ρ	P-value
G	0.630	<0.0001
R	0.579	<0.0001
B	0.391	0.002
S	0.304	0.02

TABLE 1 : Corrélations de Spearman réalisées entre les dimensions du GRBAS (0 à 3) et le jugement du trouble vocal par les auditeurs naïfs (0 à 4)

3.2 Quelle est l'incidence du trouble dysphonique sur le jugement de personnalité ?

Aux vues de nos données, notre hypothèse peut être validée. Le jugement est plus sévère sur chaque trait de personnalité pour les locutrices jugées comme ayant un trouble vocal plus lourd (voir Table 2) : Nous observons des corrélations significatives.

Echelle de personnalité	Coefficient r	P-value
Joyeuse - Triste	0.672	<0.0001
Sympathique - Désagréable	0.787	<0.0001
Dynamique - Molle	0.590	<0.0001
Confiante - Hésitante	0.719	<0.0001

TABLE 2 : Corrélations de Pearson réalisées entre les scores moyens de personnalité (0 à 4) et l'évaluation du trouble dysphonique (0 à 4) par les auditeurs naïfs

Nous obtenons également de fortes corrélations entre les jugements des différentes dimensions de personnalité entre elles. La plus forte étant de 0.842 entre l'évaluation d'une locutrice sur l'échelle « Joyeuse – Triste » et celle « Sympathique – Désagréable ». C'est pour les échelles « Dynamique – Molle » et « Sympathique – Désagréable » que nous trouvons la plus faible corrélation, soit 0.659.

3.3 Quels indices acoustiques les auditeurs exploitent-ils pour juger la personnalité ?

Nous pouvons observer de nombreux corrélats acoustiques et perceptifs de la personnalité à travers la voix (voir Figure 1).

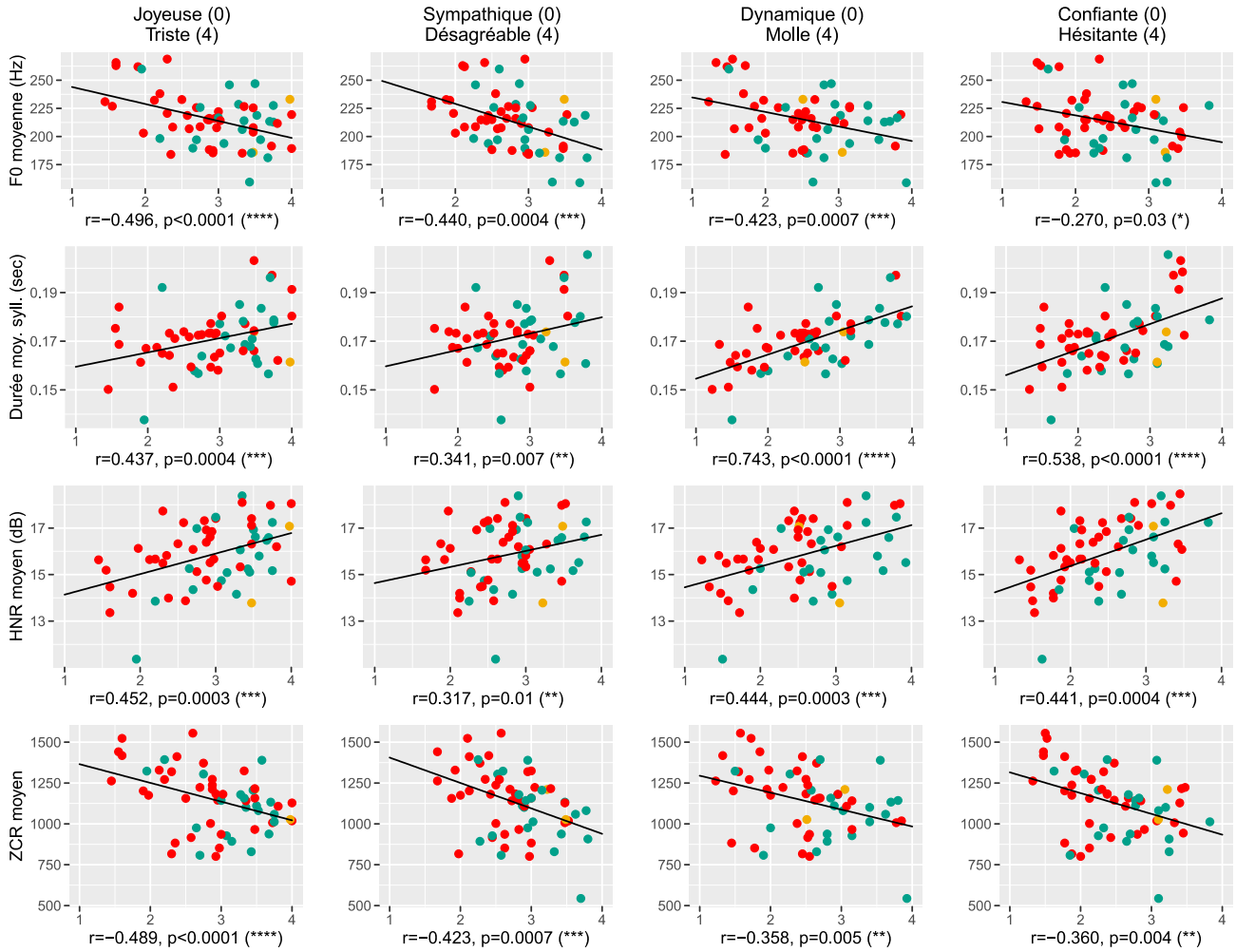


FIGURE 1 : Nuages de points des corrélations entre les scores moyens de personnalité (0 à 4, 0 étant le plus positif) et les différentes mesures acoustiques selon le Grade de dysphonie (G0 = rouge, G1 = vert, G2 = jaune) / * p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001

Parmi plusieurs mesures testées, nous utilisons celles pour lesquelles nous observons des corrélations intéressantes avec nos traits de personnalité. Nous nous intéresserons de plus près aux mesures suivantes, toutefois classiques dans les études sur la voix pathologique : F0 moyenne (Hz), Durée moyenne des syllabes (sec), Rapport harmonique sur bruit (dB), Taux de passage par zéro (ZCR).

De manière générale nous observons que les voix jugées les plus positivement, et donc très probablement considérées comme les plus attractives sont celles : avec une f0 élevée, un débit syllabique rapide, ainsi qu'un HNR bas et un ZCR élevé.

4 Discussion et conclusions

4.1 Discussion

- **Quelles dimensions du GRBAS mènent les naïfs à catégoriser les voix comme pathologiques ?**

Selon nos résultats, qui confirment la capacité d'auditeurs naïfs à percevoir les dysphonies légères (G1), les jugements de sévérité du trouble vocal seraient principalement liés à la perception du grade de dysphonie et de la raucité.

Cette corrélation modérée entre le grade de dysphonie attribué par les experts et le degré de trouble vocal moyen donné par les naïfs à partir des mêmes échantillons de parole peut s'expliquer de plusieurs façons. Elle pourrait être le fait de notre population dysphonique, qui est majoritairement composée de pathologies légères. Il a déjà été démontré que la capacité de naïfs à percevoir des troubles dysphoniques légers (G1 et G2) est moindre que pour les troubles sévères (Ghio & al., 2011). On peut également supposer que cela est le résultat d'une divergence de norme. Certains aspects vocaux peuvent en effet être jugés pathologiques par les experts, et non par les naïfs. Ainsi, dans une étude comparant les voix rauques de femmes anglophones en contexte de parole et de chant, les voix légèrement rauques ont été jugées attractives, cette dimension vocale pouvant être manipulée par la locutrice en voix chantée pour « sonner sexy » (Barkat-Defradas & al, 2013). Dans notre étude également nous observons que la raucité, facteur pourtant évalué comme indicateur du trouble dysphonique par les experts, induit des jugements plus positifs sur les traits de personnalité accordés par les naïfs à nos locutrices. Il est donc possible que cette corrélation modérée résulte d'une divergence d'interprétation sur ce qui caractérise une voix pathologique.

- **Quelle est l'incidence du trouble dysphonique sur le jugement de personnalité ?**

De manière générale nos résultats tendent à confirmer ceux des études présentées dans notre cadre théorique. Les voix qui sont identifiées comme étant les plus pathologiques sont effectivement jugées plus sévèrement sur les traits de personnalité. Ainsi, l'attribution de traits de personnalité par notre population d'auditeurs semble bien confirmer le biais perceptif lié à un effet halo négatif.

- **Quels indices acoustiques les auditeurs exploitent-ils pour juger la personnalité ?**

Les auditeurs semblent se baser sur certains indices acoustiques pour élaborer leur jugement : plus précisément le registre moyen de f0, le débit syllabique, le HNR et le ZCR.

Les multiples études s'intéressant à la personnalité à travers la voix ont différentes conclusions quant à la f0. La hauteur de voix la plus « appréciée » semble être une construction culturelle. Alors que de nombreuses études s'accordent sur le fait qu'une f0 élevée est très négative pour un homme, elle est associée à une voix dite « enfantine » pour les femmes, ce qui est parfaitement accepté et jugé positivement (Berry, 1990). A l'inverse on peut également observer que les Néerlandais préfèrent les

voix de femmes ayant une f0 basse (Van Bezooijen, 1995) contrairement aux japonais, qui plébiscitent les plus hautes fréquences. Nos données indiqueraient que cela est partagée par les français.

Les résultats indiquant qu'un HNR bas et un ZCR élevé conduisent à un jugement positif sont consistants avec la littérature. Nous savons qu'une raucité chez la femme anglophone est jugée comme attractive (Barkat-Defradas & al, 2013). Nous avons ici deux mesures acoustiques (HNR et ZCR) partiellement liées à la perception de la raucité, qui laisseraient supposer que cette tendance à apprécier les voix rauques est également vérifiée pour les voix féminines françaises. En revanche, cela ne semble pas être vrai pour tous les pays d'Europe. Une étude menée en Ecosse indique que des voix modifiées de manière à « lisser » les apériodicitées sont jugées comme plus attractives que celles avec un HNR naturel (Bruckert & al, 2010).

4.2 Conclusion générale

Les auditeurs naïfs ont une certaine capacité à déceler la dysphonie, même légère, bien que certains facteurs, comme la raucité, ne semblent pas les influencer de la même manière que les experts. En revanche, il est bien clair que les voix jugées comme pathologiques par les naïfs, que cela soit en accord ou non avec les critères experts, sont évaluées de manière beaucoup plus sévère sur les traits de personnalité que les voix catégorisées comme saines. Les croisements des scores de personnalité et des mesures acoustiques nous permettent de mettre en évidence un profil vocal qui semble être celui le plus plébiscité par les auditeurs naïfs pour les femmes française de notre panel : les voix plus aigües, plus rauques et avec un débit rapide sont associées à un jugement plus positif.

Remerciements

Ce travail est soutenu par le Labex EFL (ANR-10-LABX-0083).

Références

AMIR O., & LEVINE-YUNDOF R. (2013). Listeners' attitude toward people with dysphonia. *Journal of Voice*, *27*(4), 524-e1. DOI : 10.1016/j.jvoice.2013.01.015.

APPLE W., STREETER L. A., & KRAUSS, R. M. (1979). Effects of pitch and speech rate on personal attributions. *J. of personality and social psychology*, *37*(5), 715. DOI : 10.1037/0022-3514.37.5.715.

ARONOVITCH C. D., (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The J. of social psychology*, *99*(2), 207-220. DOI : 10.1080/00224545.1976.9924774.

ASCH S. E. (1946). Forming impressions of personality. *The J. of Abnormal and Social Psychology*, *41*(3), 258. DOI : 10.1037/h0055756.

BARKAT-DEFRADAS M., BUSSEUIL C., CHAUVY O., HIRSCH F., FAUTH C., REVIS J., & DE LA BRETEQUE B. A. (2012). Dimension esthétique des voix normales et pathologiques: approches perceptive et acoustique. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA), 28, 2-15. HAL : halshs-00778795.

BARKAT-DEFRADAS M., & DUFOUR F. (2007). La mimésis vocale : Un phénomène dialogique ? *Cahiers de praxématique*, (49), 57-78.

BARKAT-DEFRADAS M., FAUTH C., HIRSCH F., DE LA BRETEQUE B. A., SAUVAGE, J., & DODANE C. (2013, December). Rauque'n'Roll: La raucité, entre symptôme pathologique & expression artistique. In *5° Journées de Phonétique Clinique*. HAL : hal-00918332.

BERRY D. S. (1990). Vocal attractiveness and vocal babyishness: Effects on stranger, self, and friend

impressions. *J. of Nonverbal Behavior*, 14(3), 141-153. DOI : 10.1007/BF00996223.
BLOOD G. W., MAHAN B. W., & HYMAN M. (1979). Judging personality and appearance from voice disorders. *J. of Communication Disorders*, *12*(1), 63-67. DOI : 10.1016/0021-9924(79)90022-4.
BOERSMA P. & WEENINK D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.08, retrieved 5 December 2019 from http://www.praat.org/
BRUCKERT L., BESTELMEYER P., LATINUS M., ROUGER J., CHAREST I., ROUSSELET G. A., & BELIN P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116-120. DOI : 10.1016/j.cub.2009.11.034
CAETANO G., GILBERT F., LOIE C., LAPIE-LEGOUIS P., & GARSI J. P. (2017). Recherche interventionnelle sur les troubles vocaux chez les enseignants : vers une prévention collective ? *Archives des Maladies Professionnelles et de l'Environnement*, *78*(5), 421-436. DOI : 10.1016/j.admp.2017.03.003.
EXPERTISE COLLECTIVE. INSERM. (2006). La voix : ses troubles chez les enseignants. INSERM.
GHIO A., DUFOUR S., ROUAZE M., BOKANOWSKI V., POUCHOULIN G., REVIS J., & GIOVANNI A. (2011). Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle. *Rev. de laryngologie-otologie-rhinologie*, *132*(1), 1-9. HAL : hal-01491737.
GILES H., & COUPLAND N. (1991). Mapping social psychology. *Language: Contexts and consequences.* Thomson Brooks/Cole Publishing Co.
HIRANO M., Clinical examination of voice. Wien: Springer, 1981. DOI : 10.1121/1.393788.
KREIMAN J., & SIDTIS D. (2011). Foundations of voice studies: An interdisciplinary approach to voice production and perception. John Wiley & Sons. DOI : 10.1002/9781444395068.
MCCRAE R. R., & COSTA JR P. T. (1985). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and individual Differences*, *6*(5), 587-597. DOI : 10.1016/0191-8869(85)90008-X.
MONTEPARE J. M., & ZEBROWITZ-MCARTHUR L. (1987). Perceptions of adults with childlike voices in two cultures. *J. of Experimental Social Psychology*, *23*(4), 331-349. DOI : 10.1016/0022-1031(87)90045-X.
NASS C., & LEE K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *J. of experimental psychology: applied*, *7*(3). DOI : 10.1037/1076-898X.7.3.171.
OSGOOD C. E. (1952). The nature and measurement of meaning. Psychological bulletin, 49(3), 197. DOI : 10.1037/h0055737.
PARDO J. S. (2006). On phonetic convergence during conversational interaction. *The J. of the Acoustical Society of America*, *119*(4), 2382-2393. DOI : 10.1121/1.2178720.
PEIRCE J. W., GRAY J. R., SIMPSON S., MACASKILL M. R., HÖCHENBERGER R., SOGO H., KASTMAN E., LINDELØV J. (2019). PsychoPy2: experiments in behavior made easy. Behavior Research Methods. 10.3758/s13428-018-01193-y.
PENG Y., ZEBROWITZ L. A., & LEE H. K. (1993). The impact of cultural background and cross-cultural experience on impressions of American and Korean male speakers. *J. of Cross-Cultural Psychology*, *24*(2), 203-220. DOI : 10.1177/0022022193242005.
RUSCELLO D. M., LASS N. J., & PODBESEK J. (1988). Listeners' perceptions of normal and voice-disordered children. *Folia phoniatrica*. DOI : 10.1159/000265922.
THORNDIKE E. L. (1920). A constant error in psychological ratings. *J. of applied psychology*, *4*(1), 25-29. DOI : 10.1037/h0071663.
VAN BEZOOIJEN R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and speech*, 38(3), 253-265. DOI : 10.1177/002383099503800303.
ZUCKERMAN M., & DRIVER R. E. (1988). What sounds beautiful is good : The vocal attractiveness stereotype. *J. of Nonverbal Behavior*, *13*(2), 67-82. DOI : 10.1177/0265407515612445.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Émergence du contraste entre les fricatives sibilantes /s/ - /ʃ/ du français en contexte d'acquisition bilingue

Marie Philippart de Foy[1, 2], Véronique Delvaux[1, 2], Kathy Huet[1], Morgane Monnier[1]
Myriam Piccaluga[1] & Bernard Harmegnies[1]
(1) Institut de Recherche en Sciences et Technologies du Langage, Service de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fonds National de la Recherche Scientifique, Belgique
marie.philippartdefoy@umons.ac.be ; bernard.harmegnies@umons.ac.be

RÉSUMÉ

Cette contribution vise à observer l'émergence du contraste de lieu d'articulation entre les fricatives sibilantes sourdes /s/ - /ʃ/ en français chez des bilingues simultanés d'âge préscolaire exposés à l'une des deux combinaisons linguistiques suivantes : français-italien et français-arabe. Les productions orales des enfants ont été recueillies longitudinalement via une tâche de dénomination originale en français. Les deux fricatives ont fait l'objet d'analyses basées sur des données acoustiques, et plus précisément les premier et troisième moments spectraux, et sur les transcriptions phonétiques des productions de parole. L'impact du développement lexical sur la production des deux fricatives a été investigué. Les résultats suggèrent, d'une part, un contraste émergeant plus précocement chez les bilingues français-arabe et, d'autre part, une acquisition plus précoce du /s/ pour l'ensemble des participants.

ABSTRACT

Emergence of the contrast between the French sibilant fricatives /s/ - /ʃ/ in bilingual acquisition.
This contribution aims at examining the emergence of the place-of-articulation contrast between the voiceless sibilant fricatives /s/ - /ʃ/ in simultaneous bilingual toddlers exposed to one the two following language pairs: French-Italian and French-Arabic. The children's speech productions in French have been longitudinally collected via an original picture-naming task. The two fricatives have been subjected to analyses based on acoustic data, and more precisely the first and third spectral moments, and on phonetic transcriptions of the speech productions. The impact of the lexical development on each fricative's production has been investigated. Results suggest, on the one hand, an earlier emergence of the contrast in French-Arabic bilingual children and, on the other hand, an earlier acquisition of /s/ for all participants.

MOTS-CLÉS : acquisition phonologique, bilinguisme, production de parole, fricatives sibilantes.
KEYWORDS: phonological acquisition, bilingualism, speech production, sibilant fricatives.

1 Introduction

La mise en place du système phonologique implique pour l'enfant d'acquérir les sons de parole de sa langue native et les oppositions existant entre ceux-ci, autrement dit les contrastes vocaliques et consonantiques caractérisant sa langue maternelle. Si le rythme d'acquisition des sons de parole diffère d'une langue à l'autre, on observe toutefois des tendances similaires au niveau de l'acquisition des inventaires consonantiques. Des études menées dans différentes langues indiquent que les occlusives, les nasales et les glides sont généralement acquises précocement alors que les fricatives, les affriquées et les liquides tendent à être acquises plus tardivement (Dinnsen, 1992).

Volume 1 : Journées d'Études sur la Parole, pages 498–506.
HAL : hal-02798577.

Initialement prédominantes dans les productions des enfants (Boysson-Bardies & Vihman, 1991), les occlusives (labiales) impliquent des patrons articulatoires simples tandis que la production des fricatives requiert un contrôle moteur et articulatoire plus sophistiqué (Menn & Vihman, 2003). Plus précisément pour le français, les résultats d'une étude de cohorte menée par MacLeod et collaborateurs (2011) auprès de 156 enfants canadiens francophones âgés de 20 à 53 mois indiquent que les fricatives sibilantes /s, ʒ, ʃ/ feraient partie des consonnes acquises le plus tardivement. Dans le cas d'une acquisition bilingue du langage, les enfants sont confrontés au challenge de développer simultanément deux systèmes phonologiques. Un nombre restreint d'études ont porté sur l'acquisition phonologique du français en contexte bilingue (dont Brulard & Carr, 2003 ; Kehoe & Havy, 2019) et rares sont celles qui ont ciblé la production des fricatives. De fait, si les recherches existantes se sont majoritairement focalisées sur l'acquisition consonantique, elles ont plutôt évalué la précision de réalisation des consonnes globalement et/ou les processus phonologiques affectant les segments consonantiques dans leur ensemble. En outre, ces travaux ont principalement étudié le bilinguisme français-anglais. A notre connaissance, seule l'étude de Kehoe et Havy (2019) a investigué spécifiquement la production des fricatives palatales du français chez des enfants bilingues préscolaires, par ailleurs exposés à plusieurs combinaisons linguistiques associant le français à différentes langues. Néanmoins, les productions de fricatives n'ont pas été observées longitudinalement et ont été évaluées perceptivement, à partir des transcriptions phonétiques des productions orales, sans faire l'objet d'analyses acoustiques permettant une description objectivée des sons de parole. Il n'y a donc actuellement encore aucune étude qui ait impliqué l'analyse des caractéristiques acoustiques des fricatives du français produites par des enfants bilingues.

Similairement, peu de recherches sur l'acquisition monolingue du français se sont penchées sur les fricatives. Il convient de citer l'étude de Grandon (2016) qui a évalué la production des fricatives /f/ /s/ et /ʃ/ par le biais de mesures acoustiques chez des enfants francophones sourds et normo-entendants âgés entre environ 5 et 10 ans, et plus précisément via l'analyse des moments spectraux permettant de caractériser le spectre des fricatives sur base d'indices de statistique descriptive destinés à quantifier divers aspects de la forme des spectres (centre de gravité, variance, coefficients de dissymétrie et d'aplatissement). Les résultats de cette étude révèlent, entre autres, des centres de gravité distincts pour les fricatives alvéolaires /s/ et post-alvéolaires /ʃ/ chez les enfants normo-entendants, suggérant une émergence du contraste de lieu d'articulation entre les deux fricatives sibilantes sourdes avant l'âge de 5 ans. Ces résultats vont dans le sens de ceux de travaux ayant pareillement investigué les caractéristiques spectrales de fricatives produites par des enfants anglophones et japanophones (Nissen & Fox, 2005 ; Li et al., 2009) indiquant une maîtrise tardive des fricatives ainsi qu'une période d'acquisition prolongée pour le contraste entre les sibilantes /s/ et /ʃ/ à partir d'environ 4 ans. Par ailleurs, des différences inter-linguistiques dans l'acquisition des fricatives sont également relevées dans l'étude de Li et collaborateurs (2009) impliquant à la fois des analyses acoustiques et perceptives des segments.

L'étude exploratoire que nous présentons ici s'inscrit dans une démarche empirique et a pour objectif d'étudier l'évolution de la production du contraste de lieu d'articulation entre les fricatives sibilantes sourdes /s/ - /ʃ/ du français chez l'enfant en situation de bilinguisme simultané. Cette étude présente l'originalité d'adopter une approche comparative en observant des enfants exposés à deux combinaisons linguistiques : (1) français-italien et (2), français-arabe. L'objectif est d'observer si l'acquisition du contraste entre les deux fricatives du français pourrait être influencée par la deuxième langue auquel l'enfant est exposé, et plus particulièrement par la richesse de son inventaire consonantique et/ou par la complexité articulatoire des sons qu'il comporte. De fait, l'arabe, d'une part, comporte bien plus de fricatives que le français – 14 fricatives en arabe standard (/f, θ, ð, s, z, ʃ, ʒ χ, ʁ, ħ, ʕ, h, ðˤ, ʂˤ/ d'après Benamrane, 2013) pour 7 en français (/f, v, s, z, ʃ, ʒ, ʁ/) – réparties sur davantage de lieux d'articulation dont des lieux postérieurs non présents en français et, d'autre part, inclut des sons de parole présentant une grande complexité articulatoire telles que les consonnes emphatiques dont la production implique un déplacement de la zone d'articulation principale vers le

palais mou (Benamrane, 2013). L'inventaire consonantique de l'italien comporte quant à lui 5 fricatives (/f, v, s, z, ʃ/) comprises dans l'inventaire du français (Rogers & d'Arcangeli, 2004). Dans la mesure où l'enfant bilingue est simultanément exposé à deux systèmes linguistiques, chacun de ceux-ci est susceptible d'exercer des effets spécifiques séparément ou en interaction (Paradis & Genesee, 1996). Parallèlement, l'impact du développement lexical des enfants sur la précision de réalisation des deux fricatives est également investigué. Deux autres aspects innovants de l'étude résident dans l'élaboration d'un protocole original pour le recueil des productions orales des enfants et dans l'utilisation d'analyses complémentaires impliquant des données acoustiques et des mesures effectuées sur base des transcriptions phonétiques des productions de parole.

2 Méthode

2.1 Participants

L'échantillon de participants consiste en un groupe de 16 enfants bilingues d'âge préscolaire initialement âgés entre 21 et 36 mois et exposés à l'une des deux combinaisons linguistiques suivantes : français-italien et français-arabe. Plus précisément, les enfants ont été exposés au français de Belgique, à l'italien standard et à l'arabe marocain et soudanais. La répartition des participants dans chaque groupe linguistique est la suivante : 11 bilingues français-italien (5 filles et 6 garçons, âge moyen global = 34 mois, E.T. = 7 mois) et 5 bilingues français-arabe (2 filles et 3 garçons, âge moyen global = 34 mois, E.T. = 8 mois). Exposés à leurs deux langues depuis la naissance (ou dès les premiers mois), les participants sont des bilingues simultanés présentant des degrés d'exposition importants aux deux langues (au minimum trois jours entiers par semaine).

2.2 Recueil des données

Nous avons mis au point un paradigme expérimental spécifique afin de recueillir longitudinalement les productions orales en français des enfants ainsi que des données complémentaires hétéro-rapportées obtenues via des questionnaires parentaux et actualisées lors de chaque recueil de données. Plus précisément, les parents ont rempli des adaptations des *MacArthur-Bates Communicative Development Inventories* (Fenson et *al.,* 1993) dans chaque langue de l'enfant (Kern & Gayraud, 2010 ; Caselli & Casadio, 1995[1]) afin d'évaluer le développement lexical en français et globalement, dans les deux langues. Les productions de parole en français ont été recueillies lors de quatre sessions (ci-après S1-S2-S3-S4) planifiées à intervalles réguliers de 4 mois et tous les enregistrements ont été réalisés au domicile des enfants au moyen d'un enregistreur audio-portable *Zoom H5* avec une fréquence d'échantillonnage de 44 100 Hz. Les productions orales ont été recueillies via une tâche de dénomination de mots insérée au sein d'un jeu avec un livre imagier afin de cibler des structures phonologiques spécifiques dans un contexte interactif. Les items à faire produire ont été sélectionnés sur base de critères psycholinguistiques et phonologiques listés par ordre d'importance : (1) l'âge d'acquisition (ci-après AoA) sur base des normes de Chalard et *al.* (2003) et des rapports parentaux de Kern et Gayraud (2010) ; (2) l'imageabilité des mots ; la présence dans le corpus total de (3) tous les phonèmes du français, (4) toutes les consonnes du français en position initiale/médiane/finale dans le mot, (5) de groupes consonantiques dans différentes positions dans le mot et (6), de différentes structures syllabiques et longueurs de mots. Le corpus final inclut 3 items d'entrainement et 48 items test. Les items ont été organisés dans un ordre

[1] Pour l'arabe (marocain et standard), nous avons utilisé des listes de vocabulaire qui nous ont été transmises par l'équipe du Babylab de Plymouth travaillant actuellement sur des adaptations des MBCDI dans différents dialectes de l'arabe.

spécifique, en 8 séries de 6 items, par AoA et complexité phonologique croissants. La complexité phonologique a été évaluée à partir de critères précis situés à différents niveaux d'organisation phonologique (syllabique, segmental et inter-segmental) afin de générer un classement de complexité des mots. La Figure 1 représente les différentes séries de mots dans un graphe cartésien où l'axe des X correspond à l'ordre de présentation des mots et l'axe des Y au degré de complexité phonologique. Chacune des 8 séries apparait dans une couleur spécifique et 3 items d'AoA et de niveau de complexité différents sont mis en évidence, à titre illustratif.

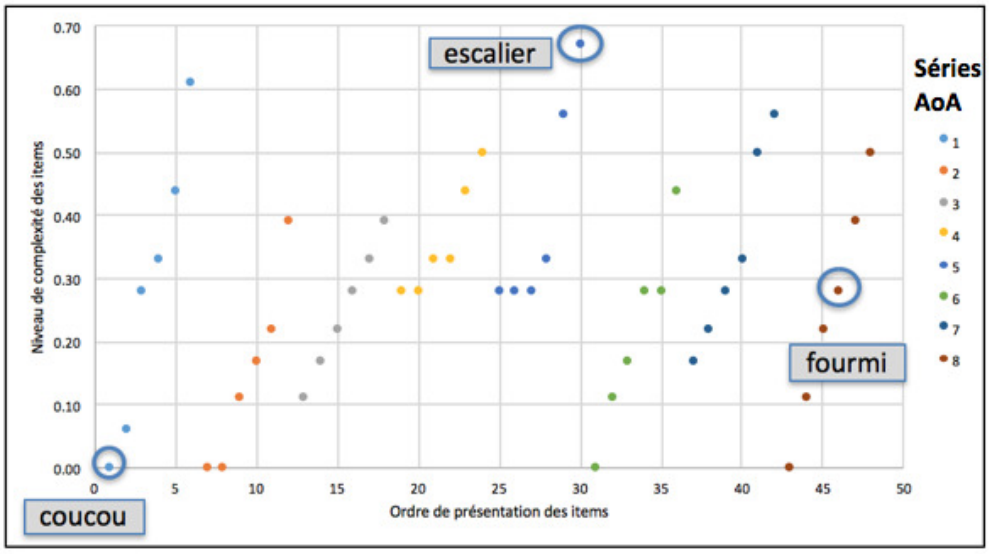


FIGURE 1 : Ordre de présentation des items en fonction de l'AoA (croissant avec les séries successives de 1 à 8) et de la complexité phonologique (croissante au sein de chaque série).

2.3 Traitement et sélection des données

Les données obtenues via les adaptations des MBCDI ont permis de calculer, pour chaque participant, deux scores de vocabulaire : un score pour le français et un score global pour les deux langues combinées. Les scores ont été codés de manière à créer des variables catégorielles à cinq niveaux[2]. Les enregistrements ont été annotés via PRAAT (Boersma & Weenink, 2015) dans le format Textgrid sur six couches d'annotation : (1) le locuteur, (2) l'item-cible, (3) la transcription phonétique, (4) la technique d'élicitation (dénomination/répétition), (5) la séquence de segments (issues de 3) alignés sur le signal acoustique et (6), les éventuels commentaires sur les productions. Toutes les annotations ont été faites manuellement excepté pour la couche 5 pour laquelle la segmentation et l'alignement automatique ont été réalisés via le logiciel SPPAS (Bigi, 2015) à partir des transcriptions de la troisième couche d'annotation. Toutes les transcriptions phonétiques ont été réalisées par l'expérimentatrice, linguiste entrainée et locutrice native du français et ont fait l'objet d'une vérification et d'un réajustement si nécessaire.

2.4 Analyses

Notre protocole nous a permis de récolter longitudinalement des productions des deux fricatives en position initiale/médiale/finale dans le mot (voir Table 1), lors des quatre sessions de chaque participant. Du fait des contraintes de corpus, nous n'avons pas pu contrôler l'environnement

[2]Niveaux des scores de vocabulaire français : (1) 27-155 mots, (2) 156-284 mots, (3) 285-413 mots, (4) 414-542 mots, (5) 543-670 mots. Niveaux des scores de vocabulaire total : (1) 54-272 mots, (2) 273-491 mots, (3) 492-710 mots, (4) 711-929 mots, (5) 930-1150 mots.

vocalique des fricatives. Si le corpus implique plus d'occurrences de la fricative /ʃ/ (7 au total et 4 pour la fricative /s/), aucun contexte vocalique n'a été favorisé pour aucune des deux fricatives.

Position / Fricative	**Initiale**	**Médiale**	**Finale**
/s/	***s**ouris*	*poi**ss**on, chau**ss**ure*	*gla**c**e*
/ʃ/	***ch**eveux, **ch**aussure, **ch**aise, **ch**ien, **ch**ampignon*	*éc**h**arpe*	*clo**ch**e*

TABLE 1 : Occurrences des fricatives dans le corpus de la tâche de dénomination.

Les fricatives ont fait l'objet d'analyses basées sur des mesures acoustiques et sur les transcriptions phonétiques des mots produits par les enfants. La 5ème couche d'annotation PRAAT impliquant les phonèmes segmentés a servi de base pour les analyses acoustiques et les valeurs des quatre premiers moments spectraux des fricatives (centre de gravité, variance, coefficients de dissymétrie et d'aplatissement) ont été automatiquement extraites via un script PRAAT personnalisé, élaboré à partir du script *Time averaging for fricatives.praat* développé par Christian Di Canio (2013, Haskins Laboratories). Les analyses ont été focalisées sur les premier et troisième moments spectraux, respectivement le centre de gravité et le coefficient de dissymétrie (ci-après, CoG et skewness pour des raisons de concision), étant donné que ces deux indices permettent le mieux de différencier les deux sibilantes /s/ et /ʃ/. En effet, le CoG correspond à la zone de fréquence principalement excitée durant la production de la fricative, qui devrait se localiser dans une zone de fréquence plus haute pour la fricative alvéolaire /s/ que pour la post-alvéolaire /ʃ/. Le skweness quantifie quant à lui la dissymétrie de la distribution de l'énergie autour de la moyenne. Un skewness positif correspondrait à une distribution biaisée à droite et donc à un maximum énergétique opposé, soit dans des fréquences basses, et à l'inverse, un skweness négatif indiquerait une concentration d'énergie dans des fréquences hautes. Dès lors, /s/ devrait présenter des valeurs de skewness plus basses que /ʃ/. Sur base des transcriptions phonétiques des productions orales, les taux de réalisation correcte *vs.* de substitution ainsi que le Pourcentage de Consonnes Correctes (ci-après, PCC) ont été calculés pour chaque fricative via le logiciel PHON (Rose et *al.*, 2006).

3 Résultats

La Table 2 présente les valeurs moyennes des premiers et troisièmes moments spectraux des deux fricatives pour chaque groupe linguistique. Nous avons exclu les fricatives inintelligibles, ainsi que celles substituées par des consonnes ayant un mode d'articulation différent, de nos analyses basées sur un total de 227 productions pour la fricative alvéolaire /s/ et de 371 productions pour la fricative post-alvéolaire /ʃ/. Les résultats de tests non-paramétriques (U de Mann-Whitney) révèlent des différences statistiquement significatives entre les deux groupes linguistiques (toutes sessions confondues) pour les valeurs de CoG ($p < .001$) et les valeurs de skewness ($p = .004$) de la fricative post-alvéolaire /ʃ/.

Groupe linguistique	Fricative	Nombre total de fricatives	Nombre moyen de fricatives par enfant par session	CoG moyen en Hz	Skewness moyen
Français-italien	s	145	3	7410 (2938)	0.16 (1.01)
	ʃ	251	6	7221 (2974)	0.37 (1.18)
Français-arabe	s	82	4	7811(2518)	0.04 (1.22)
	ʃ	120	6	6534 (2119)	0.68 (1.21)

TABLE 2 : Valeurs moyennes (avec E.T.) des 1er et 3ème moments spectraux des fricatives /s/ et /ʃ/ pour les deux groupes linguistiques.

Les graphes de la Figure 2 représentent l'évolution des valeurs de CoG (à gauche) et de skewness (à droite) des deux fricatives au cours de quatre sessions pour les deux groupes linguistiques. On peut observer que les courbes des valeurs de CoG des deux fricatives sont très proches l'une de l'autre, se confondant presque, pour le groupe de bilingues français-italien. Ceci indique que les enfants ne produisent pas encore les deux fricatives de manière distincte et ce, lors des quatre sessions. Qui plus est, les valeurs de CoG des deux fricatives sont concentrées dans des zones de hautes fréquence (au-dessus de 6000 Hz), correspondant davantage à des valeurs de CoG généralement attendues pour la fricative alvéolaire /s/. Toujours pour les bilingues français-italien, on observe des valeurs moyennes de skewness légèrement plus basses pour /s/ par rapport à /ʃ/ et ce, de manière plus prononcée lors des deux premières sessions. De manière contrastée, les courbes des valeurs de CoG des deux fricatives sont plus distinctes les unes des autres chez les participants français-arabe, avec des valeurs de CoG plus hautes pour /s/ que pour /ʃ/. La distance entre les deux courbes atteint son maximum à la S3 pour ensuite se réduire fortement à la S4. Cette réduction de la distance entre les valeurs moyennes de CoG des deux fricatives à la S4 semble être due davantage à une diminution du CoG moyen du /s/ qu'à une augmentation de celui du /ʃ/. Ensuite, on observe chez les bilingues français-arabe des valeurs de skewness plus hautes pour le /ʃ/ et plus basses pour le /s/ que chez les bilingues français-italien lors des deux dernières sessions. Similairement aux valeurs de CoG, les valeurs moyennes de skewness des deux fricatives se rapprochent lors de la S4.

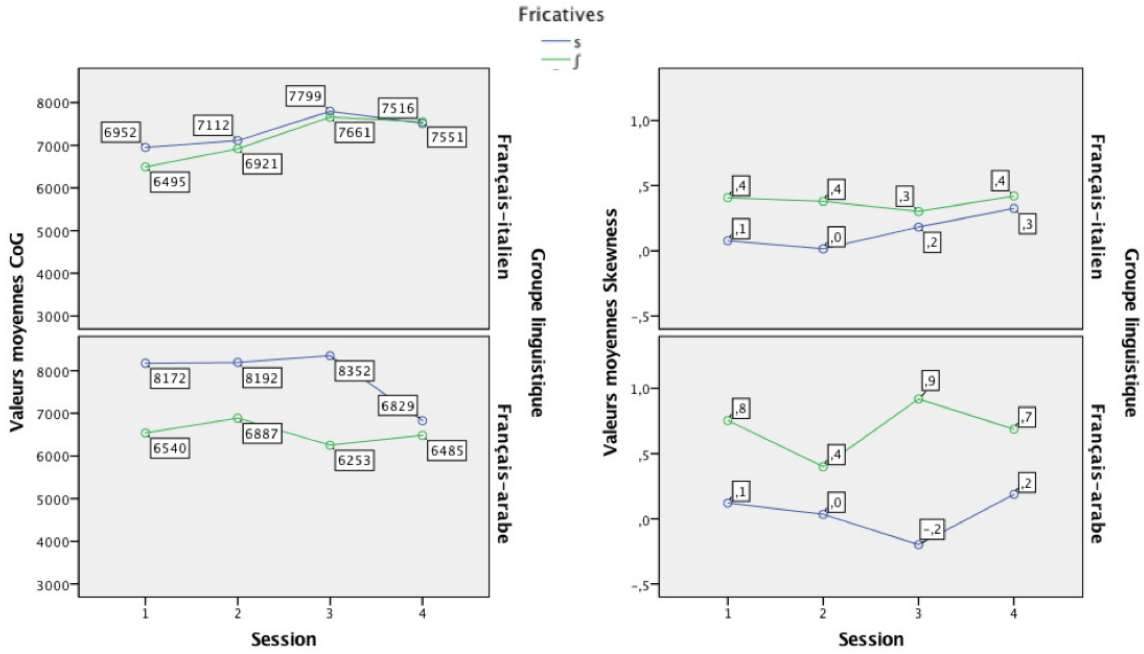


FIGURE 2 : Valeurs moyennes de CoG (en Hz, à gauche) et de skweness (à droite) des fricatives /s/ et /ʃ/ au cours de quatre sessions pour les deux groupes linguistiques.

Sur base des transcriptions phonétiques des productions de parole, les processus phonologiques affectant spécifiquement les deux fricatives ont également été investigués. Nous avons plus particulièrement examiné dans quelle proportion chaque fricative a été soit produite correctement soit substituée par l'autre fricative afin d'observer si l'une, l'autre, ou les deux consonne(s) étai(en)t plus ou moins bien réalisée(s) par les enfants et, dans le premier cas, si la consonne mieux réalisée aurait tendance à être produite à la place de l'autre. De fait, les résultats des analyses des moments spectraux suggèrent, entre autres, que les bilingues français-italien pourraient produire un son similaire pour les deux consonnes qui, d'après les valeurs moyennes de CoG se rapprocherait davantage d'un [s]. La Figure 3 représente les taux de réalisation correcte (c'est-à-dire, le son cible /s/ produit [s] et le son cible /ʃ/ produit [ʃ]) pour les deux fricatives ainsi que les taux de substitution mutuelle (c'est-à-dire, le son cible /s/ produit [ʃ] et le son cible /ʃ/ produit [s]). Comme le montre la

Figure 3, les bilingues français-italien produisent la fricative /s/ bien plus correctement que la fricative /ʃ/ et qui plus est, tendent à produire le /ʃ/ comme un [s]. Les résultats des analyses perceptives basées sur les transcriptions sont donc corrélés avec ceux des analyses acoustiques. Il apparait que le taux de réalisation correcte de /ʃ/ augmente progressivement de la S2 à la S4 ; toutefois, le taux de substitution de /ʃ/ par [s] diminue à peine. Les bilingues français-arabe présentent quant à eux des taux de substitution de /ʃ/ par [s] significativement plus bas que ceux des bilingues français-italien ($p = .006$). S'ils produisent initialement la fricative /s/ plus correctement (comme un [s]) que la fricative /ʃ/, les taux de réalisation correcte des deux consonnes deviennent très proches lors de la dernière session. Par ailleurs, les taux les plus hauts de réalisations correctes pour les deux fricatives sont observés chez les bilingues français-arabe – les différences entre les deux groupes sont significatives pour les deux fricatives (/s/ : $p = .005$ et /ʃ/ : $p < .001$) – ce qui corrobore également les résultats des analyses acoustiques. Néanmoins, l'apparente régression dans la production du contraste entre les deux fricatives observée à la S4 pour les deux moments spectraux (voir Figure 2) n'est pas reflétée par les analyses perceptives, étant donné que les taux de réalisation correcte dépassent les 80 % pour les deux fricatives. Finalement, il apparait que, pour les deux groupes de bilingues, l'alvéolaire /s/ est très peu substituée par la post-alvéolaire [ʃ].

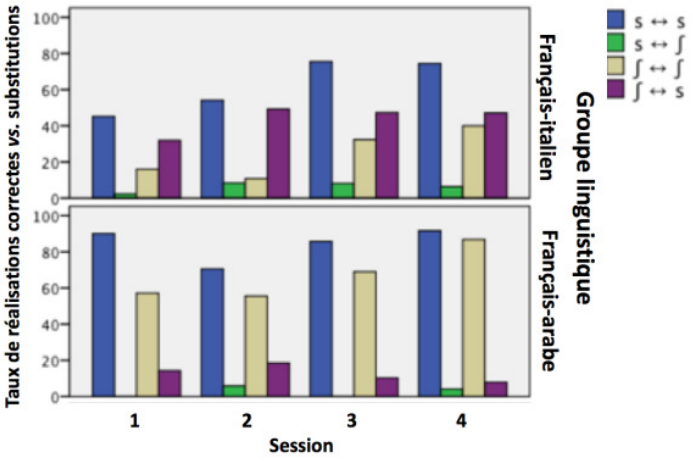


FIGURE 3 : Taux de réalisation correcte *vs.* substitution des fricatives /s/ - /ʃ/ au cours des quatre sessions pour les deux groupes linguistiques.

Nous avons également examiné l'impact du développement lexical en français et global sur la précision de réalisation des deux fricatives dans les deux groupes de bilingues et plus particulièrement, sur les valeurs moyennes de PCC de chaque fricative. Les résultats de tests non-paramétriques (Kruskall-Wallis) révèlent un effet statistiquement significatif des deux scores de vocabulaire sur la production de la post-alvéolaire /ʃ/ chez les bilingues français-italien (score de vocabulaire français : $X^2 (4) = 38.4$, $p < .001$ - score de vocabulaire global : $X^2 (4) = 18.08$, $p < .001$). Pour les deux scores de vocabulaire, l'effet va dans la même direction : les valeurs de PCC de la fricative augmentent en parallèle avec les scores de vocabulaire.

4 Discussion

L'analyse des moments spectraux suggère que le contraste de lieu d'articulation entre les fricatives sibilantes sourdes /s/ - /ʃ/ ne semble pas encore être acquis, voire n'aurait pas encore émergé, chez les participants français-italien qui tendent à produire un son similaire pour les deux consonnes ayant davantage les caractéristiques spectrales d'un [s]. Autrement dit, il est probable que lorsqu'ils tentent de produire un /ʃ/, les bilingues français-italien auraient tendance à produire le son [s]. Ensuite, les valeurs distinctes de moments spectraux des deux fricatives observées chez les bilingues

français–arabe lors des trois premières sessions semblent indiquer l'existence d'un contraste entre les deux consonnes pour ce groupe. Toutefois, les valeurs moyennes des deux moments spectraux à la S4 suggèrent une régression dans les patrons de production des enfants. Il pourrait s'agir d'une régression temporaire ; néanmoins, deux sessions supplémentaires seraient nécessaires pour pouvoir affirmer si le contraste est effectivement en voie d'être acquis. Les analyses basées sur les transcriptions indiquent que les participants français-italien produisent le son /s/ bien plus correctement que le son /ʃ/ et tendent à initialement réaliser la fricative /ʃ/ comme un [s]. De manière contrastée, les bilingues français-arabe présentent des taux de substitution de /ʃ/ par /s/ beaucoup plus bas et présentent globalement les taux de réalisation correcte les plus hauts pour les deux fricatives. Les résultats issus des deux types d'analyses, acoustiques et perceptives, corrèlent donc globalement pour les deux groupes, si ce n'est pour la régression à la S4 observée via les analyses acoustiques chez les bilingues français-arabe. Il est probable, d'une part, que les analyses des moments spectraux aient permis de mettre en évidence des phénomènes plus difficilement détectables via l'analyse perceptive des productions orales et, d'autre part, que la variabilité propre à la parole enfantine ait été propice à déclencher chez la transcriptrice des réflexes de reconstruction et d'interprétation plutôt que d'analyse (Haidar, 2018).

Par ailleurs, ces résultats sont cohérents avec ceux d'études mentionnées dans l'introduction ayant utilisé l'analyses des moments spectraux, d'après lesquels le contraste entre les deux sibilantes émergerait entre 4 et 5 ans et mettrait un certain temps avant d'être acquis (Nissen & Fox, 2005 ; Grandon, 2016). Aussi, pour les deux groupes de bilingues, la substitution de la fricative alvéolaire /s/ par la post-alvéolaire [ʃ] n'est que marginalement observée et la consonne /s/ présente les taux de réalisation correcte les plus élevés. La substitution ne se fait donc pas dans les deux directions et la consonne /s/ est produite de manière prédominante, ce qui indique une acquisition plus précoce du /s/ pour les enfants exposés aux deux combinaisons linguistiques. Les résultats suggèrent une acquisition des fricatives plus rapide chez les bilingues français-arabe. Ces différents patrons développementaux dans les deux groupes pourraient être attribués à une potentielle influence inter-linguistique entre les deux systèmes phonologiques en contact. Plus précisément, l'exposition à l'arabe pourrait avoir un effet facilitateur pour l'acquisition des fricatives en français potentiellement dû à des propriétés quantitatives et/ou qualitatives de l'inventaire consonantique de l'arabe comportant davantage de fricatives, qui plus est réparties sur des lieux d'articulation plus diversifiés, ainsi que des consonnes emphatiques présentant un mode d'articulation particulier. Cette complexité articulatoire pourrait donc permettre aux bilingues français-arabe d'acquérir plus précocement un plus grand contrôle moteur et articulatoire en comparaison avec les bilingues français-italien, ce qui pourrait les avantager pour la production des fricatives en français.

Enfin, le développement lexical n'a significativement impacté que la production de la fricative post-alvéolaire /ʃ/ chez les bilingues français-italien, avec une augmentation parallèle de la précision de réalisation de la consonne et de la compétence lexicale en français et dans les deux langues. Toutefois, il est difficile de complètement dissocier l'impact du développement lexical de celui de l'âge chronologique ; il pourrait donc être intéressant de comparer des enfants d'âge chronologique différents ayant un niveau de développement lexical similaire. Par ailleurs, on pourrait présumer que l'absence d'effet du développement lexical sur la précision de réalisation du /s/ pour les bilingues français-italien et, globalement, sur la précision de réalisation des deux fricatives chez les bilingues français-arabe serait probablement dû à l'acquisition plus précoce de la fricative alvéolaire et à un niveau de développement consonantique plus avancé et, conséquemment, à des patrons de production plus stables chez les bilingues français-arabe.

Références

BENAMRANE, A. (2013). *Etude acoustique des fricatives de l'arabe standard (locuteurs algériens)*. Doctoral dissertation, Université de Strasbourg.

BIGI, B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician-International Society of Phonetic Sciences*, (111-112), 54-69.
BOERSMA, P. & WEENINK, D. (2015). Praat, vers. 5.4. 01.
BRULARD, I., & CARR, P. (2003). French-English bilingual acquisition of phonology: One production system or two?. *International Journal of Bilingualism*, *7*(2), 177-202.
CASELLI, M. C. & CASADIO, P. (1995). *Il primo vocabolario del bambino: guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita* (Vol. 5). FrancoAngeli.
CHALARD, M., BONIN, P., MEOT, A., BOYER, B. & FAYOL, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French - relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. *European Journal of Cognitive Psychology*, 15(2), 209-245.
de BOYSSON-BARDIES, B., & VIHMAN, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67*(2), 297-319.
DI CANIO, C. (2013). Time-Averaging for Fricatives. Praat script. http://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives.praat
DINNSEN, D. A. (1992). Variation in developing and fully developed phonetic inventories. *Phonological development: Models, research, implications*, 191-210.
FENSON, L., DALE, P. S., REZNICK, J. S., THAL, D., BATES, E., HARTUNG, J. P., PETHICK, S. & REILLY, J. S. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego: CA Singular Publishing Group.
GRANDON, B. (2016). *Développement typique et atypique de la production de parole : caractéristiques segmentales et intelligibilité de la parole d'enfants porteurs d'un implant cochléaire et d'enfants normo-entendants de 5 à 11 ans* (Doctoral dissertation).
HAIDAR, L. A. (2018). *De la linguistique à la didactique, regards croisés en phonétique. Oralité-Variabilité-Corpu*. Mémoire d'Habilitation à diriger des recherches.
KEHOE, M., & HAVY, M. (2019). Bilingual phonological acquisition: the influence of language-internal, language-external, and lexical factors. *Journal of child language*, *46*(2), 292-333.
KERN, S., & GAYRAUD, F. (2010). *Inventaire Français du Développement Communicatif: 8/30 mois*. Les Editions de la Cigale.
LI, F., EDWARDS, J., & BECKMAN, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, *37*(1), 111-124.
MACLEOD, A. A., SUTTON, A., TRUDEAU, N., & THORDADOTTIR, E. (2011). The acquisition of consonants in Québécois French: A cross-sectional study of pre-school aged children. *International Journal of Speech-Language Pathology*, *13*(2), 93-109.
MENN, L., & VIHMAN, M. M. (2003). *Acquisition of Language: Phonology*. Oxford International Encyclopedia of Linguistics (2nd edition). Oxford: Oxford University Press.
NISSEN, S. L., & FOX, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *The Journal of the Acoustical Society of America*, *118*(4), 2570-2578.
PARADIS, J., & GENESSE, F. (1996). Syntactic acquisition in bilingual children: Autonomous or interdependent? Studies in Second Language Acquisition,18, 1-25.
ROGERS, D., & d'ARCANGELI, L. (2004). Italian. *Journal of the International Phonetic Association*, *34*(1), 117-121.
ROSE, Y., MACWHINNEY, B., BYRNE, R., HEDLUND, G., MADDOCKS, K., O'BRIEN, P. & WAREHEM, T. (2006). Introducing Phon: A software solution for the study of phonological acquisition. In *Proceedings of the 30th Annual Boston University Conference on Language Development*, 489-500. Somerville.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Apport des comptines pour la prononciation du /y/ français chez des enfants italophones : une étude perceptive pilote

Claire Pillot-Loiseau[1] Martina Grando[2]
(1) Laboratoire de Phonétique et Phonologie (LPP) UMR 7018, Université Paris 3 Sorbonne Nouvelle, CNRS, 19 rue des Bernardins, 75005 Paris, France
(2) New English Teaching School, Via Marconi, Res. Ripa 102, Milano 20080 Basiglio, Italie
claire.pillot@sorbonne-nouvelle.fr, grandomartina@gmail.com

RÉSUMÉ

Dans l'apprentissage de /y/ français par des enfants italophones débutants de 6 ans, les comptines sont-elles plus efficaces qu'un apprentissage phonétique les utilisant sans leur rythme et leur mélodies propres ? Deux classes de première année élémentaire d'une école publique milanaise ont suivi douze séances d'apprentissage de la prononciation du Français Langue Etrangère, avec comptines originales pour le Groupe Expérimental (GE) mais juste parlées pour le Groupe Contrôle (GC). L'apprentissage de /y/, durant 4 séances, s'est déroulé avec des tâches de perception, productions corporelle et verbale. Après la troisième séance, 7 enfants du GE et 7 du GC ont été enregistrés sur une comptine, perceptivement évaluée par 4 auditeurs experts et 4 auditeurs étudiants, français natifs : plus d'occurrences sont perçues comme correctes pour le GE chez les auditeurs experts. Pour le GE, /y/ non correctement produit était principalement remplacé par un phonème également antérieur (/i/), contrairement au GC (/u/).

ABSTRACT

Contribution of nursery rhymes for the pronunciation of French /y/ in Italian-speaking children: a perceptive pilot study.
In the learning of French /y/ by beginner Italian children aged 6, are nursery rhymes more effective than phonetic training using them without their own rhythm and melody? Two elementary first year classes at a public school in Milan attended twelve French as a Foreign Language's phonetics learning sessions, with original nursery rhymes for the Experimental Group (EG), and with just spoken nursery rhymes for the Control Group (CG). The learning of /y/, during four sessions, took place with tasks of perception, bodily and verbal productions. After the third session, 7 children from the EG and 7 from the CG were recorded on a nursery rhyme, perceptively assessed by 4 expert and 4 students, French native listeners: more occurrences are perceived as correct for the EG among the expert listeners. For the EG, incorrect /y/ was replaced by an equally anterior phoneme (/i/), unlike the CG (/u/).

MOTS-CLÉS : Français Langue Etrangère, comptines, prononciation de /y/, enfants italophones.
KEYWORDS: French as a Foreign Language, nursery rhymes, /y/ pronunciation, Italian children.

Volume 1 : Journées d'Études sur la Parole, pages 507–515.
HAL : hal-02798579.

1 Introduction et état de l'art

Si de nombreuses études portent sur le rôle de l'âge et des apprentissages explicites lors de l'acquisition d'une langue étrangère par des enfants (Watorek & Wauquier-Gravelines, 2016), peu de recherches ont expérimentalement objectivé l'impact d'enseignements spécifiques de leur prononciation du Français Langue Etrangère (FLE). En outre, il a déjà été démontré que les chansons et comptines améliorent l'apprentissage d'une langue : leur utilisation « pour jouer avec les formes sonores des mots » (Delasalle, 2005 :109) est fortement recommandée. Les comptines permettent en effet une approche plurisensorielle (visuelle, auditive, kinesthésique en particulier) respectant la nature multiple de l'intelligence enfantine (Gardner, 2001 ; Vanthier, 2009) : approche ludique, discursive (car « les mots de la langue ne prennent sens qu'en contexte, dans le cadre de discours et de textes », Vanthier, 2009 : 47), interculturelle et interdisciplinaire.

L'enfant « développe très tôt des capacités à repérer les phénomènes de rythme, d'accentuation et d'intonation [...] bien avant de comprendre le sens de chaque mot particulier » (Delasalle, 2005 : 104), c'est pourquoi tenir compte des aspects suprasegmentaux dans l'apprentissage de la prononciation d'une langue chez l'enfant est fondamental. Musique et langage partagent des paramètres et fonctions communs dans leur composition (fréquence, intensité, durée ; communication esthétique, structures profonde et de surface). Même si le langage seul possède une fonction communicative linguistique au sens strict du terme, ses nombreux points communs avec la musique expliquent qu'une pédagogie fondée sur l'apport de comptines est efficace pour améliorer la prononciation d'une langue étrangère (Calvet, 1980 ; Cornaz & Caussade, 2014).

A l'origine « formulette enfantine parlée ou chantée servant à départager ou à désigner celui à qui sera attribué un rôle particulier dans un jeu » (Vanthier, 2009 : 82), la comptine est aujourd'hui le genre littéraire le plus riche en rythme, capable d'associer rythme et parole (parlée ou chantée) avec régularité et simplicité. Étant « très facilement perçue, mémorisée et reproduite », elle suscite une « immense fascination [...] sur les enfants dans le domaine de l'apprentissage de la langue maternelle, dans celui d'une langue étrangère ou encore dans celui de la rééducation des troubles de la parole. » (Roberge, 2003 : 118). « Apprendre des comptines en langue étrangère ouvre à une autre culture où le découpage du réel à travers les mots est à la fois nouveau et différent » (Vanthier, 2009 : 82) ; les comptines permettent d' « affiner la perception auditive » (Vanthier, 2009 : 82) par des rythmes inconnus et étrangers qui suscitent l'intérêt des enfants.

L'italien standard ne partage comme voyelles communes avec le français que les voyelles antérieures étirées et les voyelles postérieures (Ferrero, 1972). Les oppositions /e/-/ɛ/ et /o/-/ɔ/, très variables au niveau régional, persistent en italien standard seulement dans les syllabes accentuées (Rogers & d'Arcangeli, 2004). Il est donc prévisible qu'un italophone aie des difficultés à prononcer les voyelles nasales et antérieures arrondies /y, ø, œ/, sauf s'il parle le piémontais qui possède cette série vocalique. Notre choix s'est donc porté sur l'acquisition de /y/ par des enfants italophones à Milan : il est attendu des écarts de prononciation de type [u] pour cette voyelle.

Cette étude a donc pour but de savoir si l'utilisation auprès d'enfants italophones de comptines chantées en classe de langue pour des finalités phonétiques, est perçue comme efficace pour l'acquisition du /y/ français, par des auditeurs francophones natifs experts et étudiants en phonétique et en didactique de la prononciation du FLE. Il est attendu que des enfants ayant travaillé les comptines chantées et rythmées originales, prononcent mieux le /y/ français que des enfants qui ont travaillé avec la version parlée (sans rythme et sans mélodie) de ces comptines.

2 Méthode

2.1 Population et travail effectué auprès des enfants

L'expérimentation s'est effectuée dans une école publique de Milan, actuellement considérée comme le pôle de standardisation du pays au niveau linguistique, où la langue nationale s'est diffusée et affirmée de façon plus définie (Berruto, 1987). Deux classes de 23 élèves débutants de FLE (âge moyen : 6 ans, autant de garçons que de filles) de l'équivalent du cours préparatoire français, ont suivi 12 séances conduites par la deuxième auteure durant six semaines (2 x 1h hebdomadaires sur deux jours différents). Quatre séances d'1h chacune constituaient une séquence didactique, visant notamment un travail ciblé sur un phonème du FLE à l'aide de l'apprentissage de deux comptines par phonème étudié (séances 1 à 4 : /y/ - /u/ ; séances 5 à 8 : /ʒ/ - /ʃ/, non développé ici; séances 9 à 12 : /ʁ/, non développé ici). Une classe était le groupe expérimental (GE) auquel les comptines ont été perçues et produites dans leur version originelle chantée, l'autre le groupe contrôle (GC) qui a travaillé sur les mêmes comptines dépourvues de rythme et de mélodie.

Les séances 1 à 4 (/y/-/u/) étaient composées de : 1. jeux d'*échauffement* (corps, articulateurs, onomatopées avec les sons-cible, 10 à 15mn), 2. l'*appropriation* et la *mémorisation de 2 comptines* (découverte de *Une poule sur un mur* séance 1 et de *Lulu la tortue* séance 3 ; explication à l'aide d'images et de gestes[1] durant 5mn, réécoute régulière d'un support sonore enregistré chanté pour le GE (dans Samson, 2010) et parlé sans rythme par la première auteure pour le GC ; exécution collective puis tour à tour, phrase par phrase, 10mn) ; 3. *la découverte du phonème-cible* (association geste-phonème ; caractéristiques perceptives de /y/ et /u/ puis des mots les contenant dans les 2 comptines, via cris, bruits et couleurs, 10mn) ; 4. l'*appropriation du phonème* (jeux phonologiques de perception – discrimination et identification à l'aide de cartes couleurs, images ou gestes – et de production – répétition, production de mots des comptines avec le phonème-cible en le transmettant à un autre élève jusqu'à l'enseignant, dénomination, 15mn) ; 5. la *consolidation et l'évaluation* (production des 2 comptines collectivement et chacun son tour, 10mn).

Après dépouillement d'un questionnaire proposé aux parents et enseignants pour cerner au mieux le profil linguistique de chaque enfant, 14 enfants (GE : 4 filles et 3 garçons italiens de 6 ans 7 en moyenne, écart-type : 3,5 mois ; GC : même nombre de garçons et de filles de 6 ans 5 en moyenne, écart-type 3,6 mois) ont été choisis, sans troubles du langage et ayant séjourné en France au plus 3 semaines et ne suivant par ailleurs aucun cours de français mais d'anglais depuis 4 mois. Une mère du GE et un père du GC sont bilingues français-italien, et ils parlent régulièrement en français à leurs enfants (90% pour l'enfant du GE et 100% pour l'enfant du GC). Nous n'avons malheureusement pas pu obtenir un enregistrement de leurs productions de /y/. La mère de l'enfant du GE est bilingue tardive, tandis que le père de l'enfant du GC est un bilingue précoce. Tous les enfants suivent des cours de musique hebdomadaires à l'école en chantant différents genres musicaux. En contexte scolaire, les enfants n'apprennent pas à jouer des instruments ; hors contexte scolaire, un seul locuteur du GC suit des cours de piano depuis trois ans. Tous avaient une attitude respectueuse envers l'enseignant et une disponibilité à apprendre en classe.

[1] La partition et la nature des gestes inventés par la deuxième auteure pour les mots contenant /y/ de la comptine *Une poule sur un mur* peuvent être visualisées au lien suivant : https://drive.google.com/file/d/1CFM8TxH9wp-ujMWAlqnJWGOCJaMC5_MY/view?usp=sharing

2.2 Corpus et modalités d'enregistrement

Une poule sur un mur et *Lulu la tortue* (Martin & Trésallet, 1998 : fiche 6) ont fait l'objet d'un apprentissage auprès des enfants, débuté respectivement aux première et troisième séance. Seule la première phrase de la comptine *Une poule sur un mur* (Samson, 2010) concerne la présente étude perceptive. Cette célèbre comptine française est courte dans ses groupes rythmiques mais aussi dans ses mots dont la plupart son monosyllabiques, et de rythme lent et régulier. Sa mélodie descend à la fin de chaque groupe rythmique dont les deux premiers se terminent sur une valeur longue correspondant à la production de /y/ (*mur* et *dur*) ; /y/ se trouve également en initiale de groupe (*une*, *sur*, *du*), marqué par une note plus élevée que les autres[1]. /u/ apparaît également une fois (*poule*). En outre, l'occurrence de /u/ et les 5 occurrences de /y/ se trouvent dans la plupart des cas dans des contextes favorisant son émission (Callamand, 1981), notamment devant /n/ pour *Une*, après /s/ pour *sur*, après /d/ pour *dur* concernant le phonème /y/. Enfin, du point de vue lexical, la présence d'onomatopées ressemblant au cri de l'animal mentionné et l'introduction d'un nom d'animal (« poule ») rentrent dans une sphère sémantique proche de celle des enfants.

Pour n'engendrer aucune frustration, tous les élèves ont été enregistrés dans un lieu calme de leur école avec un enregistreur *Fostex* et un microphone externe *Beyerdynamic M201TG* à 2 cm de la bouche, après la troisième séance. Trois comptines les plus faciles parmi les 6 apprises en 12 séances, ont été enregistrées, répétées deux fois. Seul le test de perception issu de l'enregistrement de la comptine comportant les occurrences de /y/ (*Une poule sur un mur*, séance 3) est ici présenté.

2.3 Analyse perceptive

Les données obtenues des 14 enfants ont été découpées puis segmentées par groupe rythmique en notations phonétique et orthographique sous Praat (Boersma et Weenink, 2018). Le premier groupe rythmique de ***U**ne **pou**le **su**r un **mu**r* a été sélectionné pour le test perceptif et les stimuli obtenus normalisés en intensité puis regroupés selon l'ordre suivant : un premier stimulus d'exemple, puis deux stimuli d'entraînement, puis 28 stimuli (2 par enfant des deux groupes, 14 stimuli extraits des productions du GE puis les 14 autres issus de celles du GC). Chaque stimulus a été répété deux fois avec une pause inter stimuli constante (respectivement 2s entre deux répétitions et 9s pour laisser aux auditeurs le temps de répondre aux différentes questions). Les occurrences avec les phonèmes cible ont été numérotées (exemple : ***u(1)**ne poule s**u(2)**r un m**u(3)**r*).

Après chaque stimulus, l'auditeur devait coter chacune des trois occurrences par stimuli (soit 84 occurrences en tout par auditeur : 28 stimuli x 3) en choisissant entre 0 (« confusion avec un autre phonème » comme indiqué aux auditeurs sur le formulaire de recueil des réponses), 1 (absence de confusion phonologique mais « faute de réalisation phonétique » ; par exemple : /y/ non réalisé comme un autre phonème de type /u/ mais pas assez arrondi ou trop ouvert) ou 2 (« réalisation correcte »). Ce type de cotation de 0 à 2 est inspiré de Di Cristo (1975). S'il notait 0 ou 1, l'auditeur devait préciser, pour chaque occurrence, quel phonème (score 0) ou vocoïde (score 1) remplaçant le phonème cible il avait perçu. Au moyen du logiciel VassarStats, un test de Mann-Whitney a permis de comparer l'ensemble des scores par auditeur entre le GE et le GC, et un test de Wilcoxon a comparé le nombre de réponses avec les scores 0, 1 et 2 entre les deux groupes d'auditeurs.

4 auditeurs francophones experts (3 femmes dont 2 musiciennes, 1 homme musicien, professeurs de FLE et linguistique, âge moyen 48 ans, écart-type ET 8,6) et 4 auditeurs étudiants en M2 de linguistique dont 2 musiciennes (3 femmes, 1 homme, français natifs, âge moyen 24 ans, ET : 0,9) ont passé ces tests en ligne avec un casque, sauf pour un auditeur étudiant et un auditeur expert.

3 Résultats

3.1 Réponses de l'ensemble des auditeurs

Le test de perception sur *Une poule sur un mur* (phonème /y/) a été jugé difficile à faire par nos auditeurs, notamment pour faire abstraction des autres écarts de prononciation que ceux concernant /y/. 672 réponses (=28 stimuli × 3 occurrences × 8 auditeurs), soit 336 par groupe, ont été obtenues.

La figure 1 (gauche) montre le nombre de réponses de tous les auditeurs concernant /y/, en fonction du score attribué aux trois occurrences confondues, et elle montre au milieu et à droite le détail des scores de perception en fonction de l'occurrence : si le GE présente un nombre légèrement plus important de réalisations perçues comme correctes (score 2) que le GC, le nombre de confusions avec un autre phonème perçu (score 0) est notablement inférieur pour le GE, parce-que ce groupe comporte plus d'écarts uniquement phonétiques (score 1). La figure 1 (au milieu et à droite) montre le nombre de réponses de tous les auditeurs en fonction de l'occurrence et du groupes d'enfants : *sur* (contexte favorisant avant /y/) est plus favorisé que *une* (en début de groupe rythmique) et *mur* (contexte défavorisant) pour les GE et GC, le GE révélant moins de réponses avec un score 0 que 1.

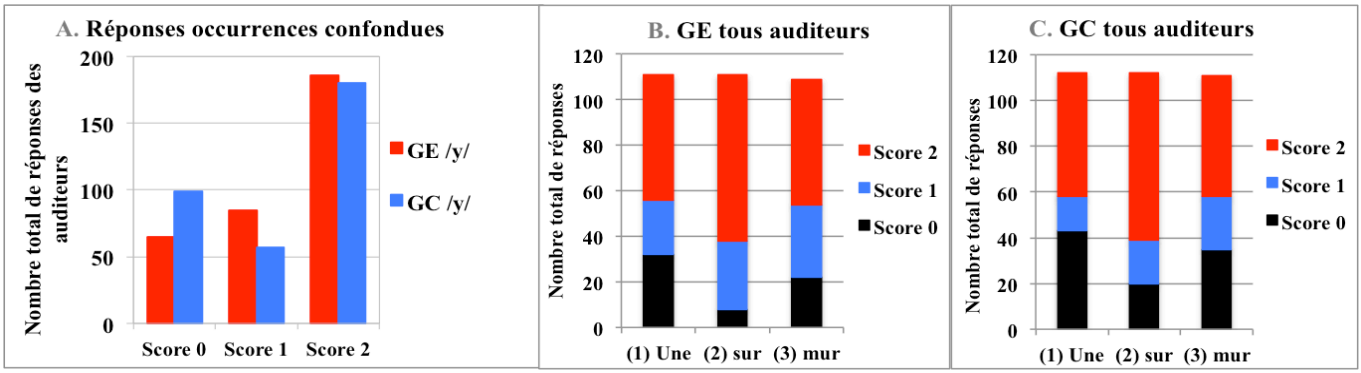


FIGURE 1: gauche : nombre de réponses de tous les auditeurs pour /y/ en fonction du score attribué aux trois occurrences confondues. Droite : nombre de réponses en fonction de l'occurrence. Score 0 : confusion avec un autre phonème, Score 1 : faute de réalisation phonétique, Score 2 : correct.

3.2 Différences entre les réponses des auditeurs experts et étudiants

La figure 2 montre les nombres de réponses concernant /y/ (A et B) ainsi que les moyennes et écart-types des scores auditeur par auditeur (C) pour les auditeurs étudiants et experts en fonction du groupe (GE et GC) et de l'occurrence (« une », « sur », « mur », figures D1 à D4) : même si le score 2 est majoritaire pour les deux groupes et pour les deux catégories d'auditeurs, on observe plus de réponses avec le score 0 pour le GC chez les experts, et plus de score 2 (correct) pour le GE chez ces mêmes experts, par rapport aux étudiants : un écart d'ordre phonologique est donc plus perçu chez les professeurs que chez les étudiants qui détectent plus d'altérations phonétiques. En outre, il existe moins de réponses avec le score 1 par les experts pour les GC et GE (W_6=-21, p<0,05). Les moyennes et écart-types des scores auditeur par auditeur (Figure 2C) montrent des scores supérieurs pour le GE chez tous les auditeurs experts (dont les auditeurs A1 et A2, musiciens amateurs de bon niveau) et deux auditeurs étudiants (contrairement à l'auditeur étudiant A2, musicien et choriste amateur, et A3, non musicien). Excepté chez l'auditeur 2 expert ($U_{(42,\ 42)}$=631,5 ; p=0,012), ces tendances ne sont cependant pas significatives. Il est à noter que les productions de /y/ des deux enfants dont les parents sont bilingues français-italien sont perçus par tous les auditeurs avec des scores moyens globaux de 1,7 contre 1,3 pour l'ensemble des autres enfants.

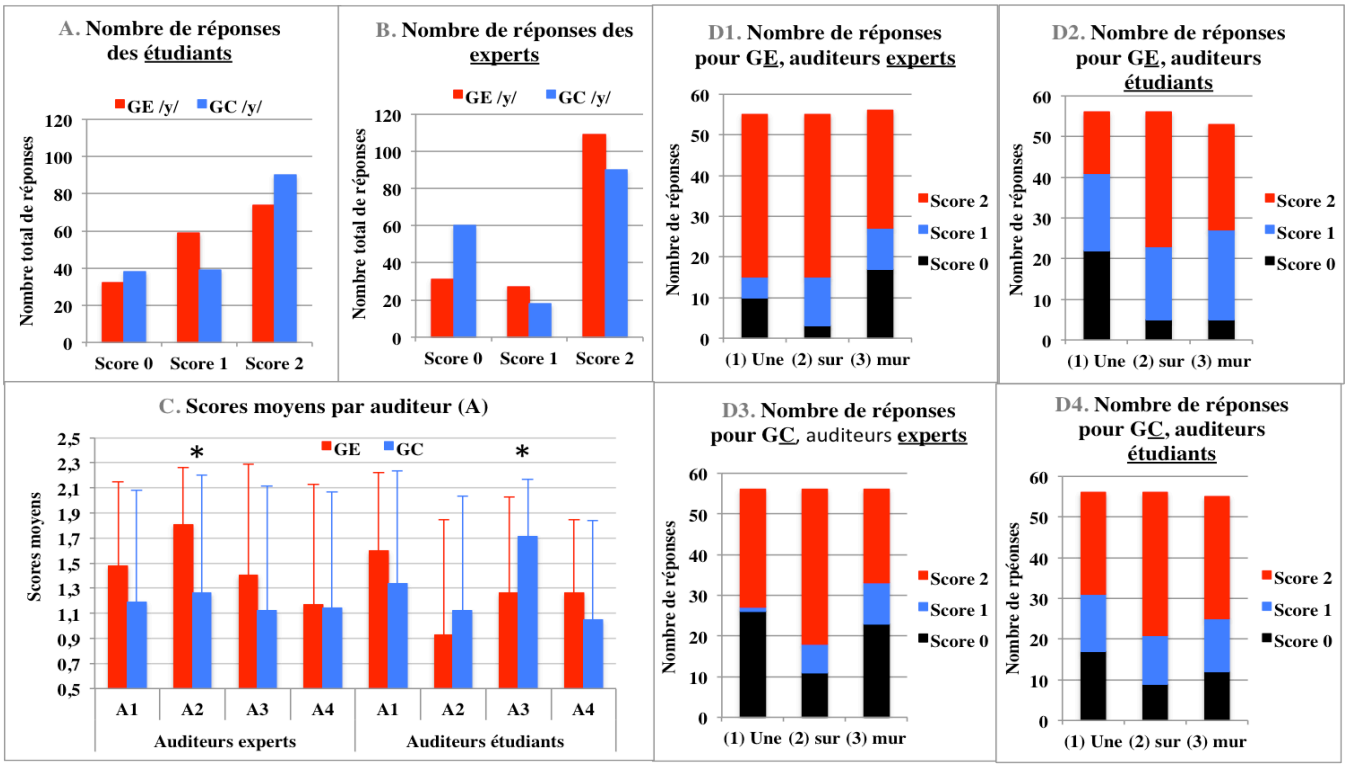


FIGURE 2 : Phonème /y/ : nombre de réponses en fonction du groupe (A, B) et de l'occurrence (D), moyennes et écart-types des scores auditeur par auditeur (C) pour les auditeurs étudiants (A, C, D2, D4) et experts (B, C, D1, D3).

La figure 2D détaille le nombre de réponses des auditeurs experts et étudiants en fonction du groupe et de l'occurrence : *sur* (contexte favorisant avant /y/) comporte plus de perception de la voyelle comme conforme au phonème attendu /y/ (plus de réponses avec le score 2 et moins avec le score 0), surtout pour le GE et chez les auditeurs experts. Par ailleurs, les scores 1 (absence de confusion phonologique mais faute de réalisation phonétique) sont moins utilisés dans les réponses des auditeurs experts pour ces trois types d'occurrences, au profit des scores 0 (phonème confondu avec un autre), surtout pour le GC.

3.3 Accords intra et interlocuteurs

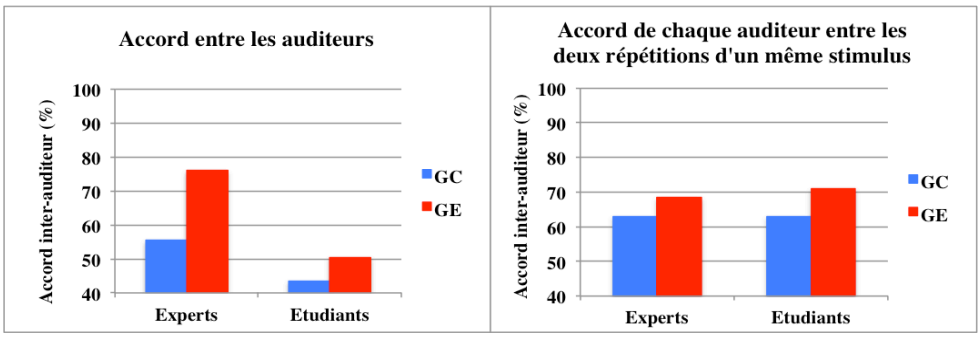


FIGURE 3 : pourcentage d'accord absolu inter-auditeurs (gauche) et intra-auditeurs (droite)

La figure 3 (gauche) représente le pourcentage d'accord (ou concordance) absolu (calculé d'après Kreiman *et al.*, 1993) entre les quatre auditeurs de chaque groupe, pour les GE et GC: ce pourcentage est plus élevé pour les experts d'une part, et pour le GE d'autre part. Concernant l'accord intra-auditeur (droite), les résultats sont similaires entre les experts et les étudiants, avec un accord intra-auditeur supérieur pour le GE.

3.4 Nature des erreurs perçues

La figure 4 montre le nombre de réponses de tous les auditeurs (haut) et des experts (bas) sur la nature du remplacement de /y/ en fonction du score (0 ou 1) et du groupe (GE et GC). Quand la réalisation de /y/ est perçue comme incorrecte (score 0), son phonème de remplacement perçu est majoritairement /i/ chez les enfants du GE, et /u/ pour le GC, surtout pour l'occurrence *une* et pour les auditeurs experts. Quand /y/ est réalisé [y] mais avec une mauvaise articulation de celui-ci (score 1), la même tendance est observée pour *une* que pour les réalisations phonologiquement incorrectes, contrairement aux deux autres occurrences, ce qui n'est pas le cas des auditeurs experts. Les productions du GE du phonème attendu /y/ sont donc davantage perçues par tous les auditeurs francophones natifs comme une voyelle (/i/) dont l'antériorité est similaire à celle de ce phonème cible, contrairement à la voyelle majoritairement perçue comme erronée du GC (/u/, postérieure). L'arrondissement propre à /y/ n'est cependant pas perçu pour le GE. Cette tendance est plus importante pour les auditeurs experts, surtout pour l'occurrence *mur* (figure 4 en bas). Les productions de /y/ perçues comme erronées, avec une perception d'autres voyelles que [i] ou [u], peuvent concerner [e, ɛ, ø, jø, œ, o], en particulier pour l'occurrence *mur* pour les scores 0 et 1 attribués par tous les auditeurs aux enfants du GE.

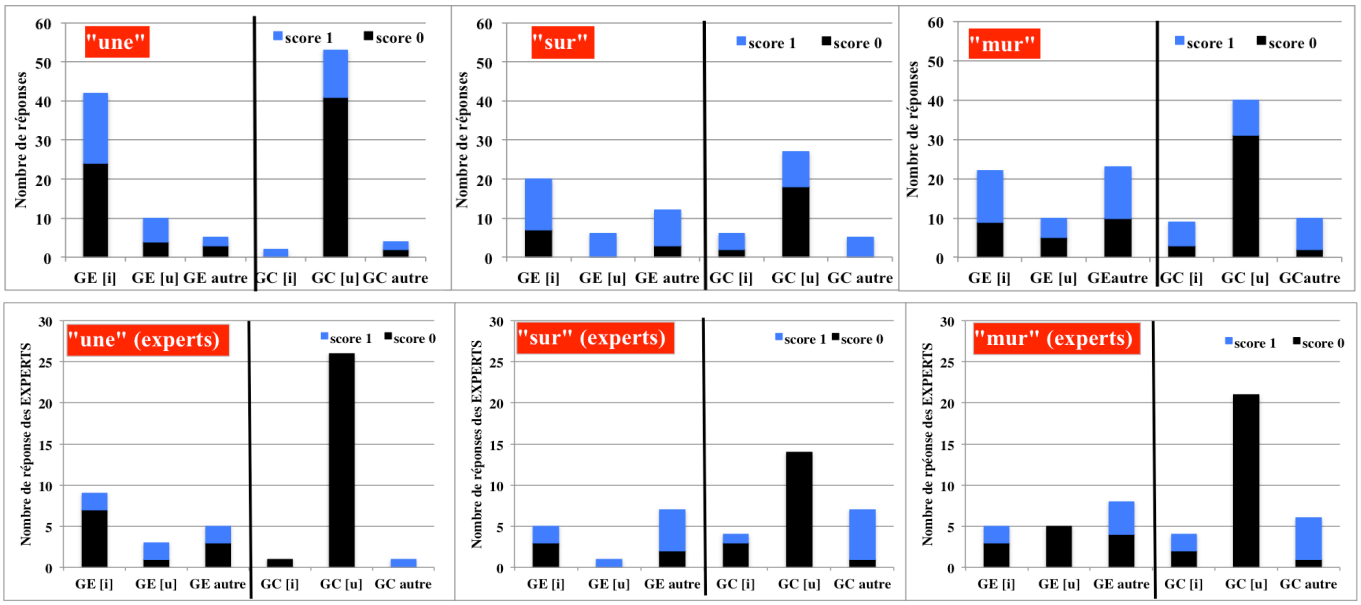


FIGURE 4 : Nature du remplacement de /y/ à chaque occurrence de ***une poule sur un mur***. GC : groupe contrôle ; GE : groupe expérimental. Score 0 : incorrect phonologiquement ; score 1 : faute de réalisation phonétique. En haut : réponses tous auditeurs confondus. En bas : auditeurs experts.

4 Discussion et conclusion

La différence dans la perception des scores corrects de réalisation de /y/ (score 2) entre les GE et GC ne paraît pas notable tous auditeurs confondus : plusieurs auditeurs ont exprimé leur difficulté à effectuer le test de perception. En outre, tous les enfants pratiquent la musique et ont reçu un apprentissage des comptines avec des gestes, efficaces pour de nouvelles acquisitions (Tellier, 2010). Ces résultats seraient probablement plus notables si plus de quatre séances avaient pu être consacrées au travail des /y/ et /u/ français. Toutefois, la perception de la voyelle comme conforme au phonème attendu /y/, est plus importante pour le GE chez les auditeurs experts, surtout dans l'occurrence *sur*, bien que ce mot grammatical ne soit pas situé en fin de groupe rythmique. /y/ de *sur* est produit en contexte favorisant avant /y/, et chanté (pour le GE) comme *Une*, sur une note plus aigue que *mur*[1] et surtout beaucoup plus aigue que *–le* de *poule* qui le précède[1]. *Sur* est aussi

associé avec la montée des bras et mains vers le haut[1] (GE et GC). Les accords inter-auditeurs sont les plus importants chez les experts et pour le GE. Les étudiants détectent globalement plus d'altérations phonétiques et moins d'erreurs phonologiques, peut-être en raison d'une différenciation entre phonétique et phonologie acquise théoriquement mais pas encore exercée perceptivement. Nous n'expliquons pas pourquoi ils jugent plus sévèrement la production de /y/ dans *Une* pour les enfants du GE. Dans les scores perçus comme incorrects (score 1 et surtout score 0), le GE est perçu, surtout par les experts, comme remplaçant le /y/ attendu par une voyelle de même acuité mais étirée, alors que les productions du GC sont perçues comme remplaçant ce phonème par une voyelle grave et arrondie, probablement à cause de l'effet du travail chanté auprès du GE. Notons que les productions des deux enfants bilingues français-italien sont supérieures à celles des autres enfants.

En effet, l'aspect affectif et ludique des comptines chantées, leur structure syntaxique simple, répétitive donc rassurante, facilement compréhensible et mémorisable, et leur contenu positivement émotionnel (Schön et al., 2008), expliquent leur efficacité pour l'amélioration de la prononciation d'une langue étrangère, si elles respectent l'accentuation de la langue et ne la contrecarrent pas musicalement, et si l'accompagnement musical éventuel ne masque pas la correcte perception de la voix du chanteur. Dans le domaine phonétique, « la langue chantée fonctionne comme une loupe des phénomènes articulatoires de tel ou tel autre système phonologique. » (Zedda, 2006 : 258). Chez les enfants, la rythmisation du langage par la comptine est un jeu, une motivation au travail et elle peut être associée aux exercices « musico-moteurs » visant « à perfectionner chez les élèves la coordination audio-motrice », celle entre le geste et le son produits (Pamula, 2008 : 137-138). Si les comptines sont courtes tout en apportant un peu de nouveauté aux enfants (Zedda, 2006), le mouvement au rythme aide les jeunes apprenants à se détendre, à participer plus activement et à « développer une expression créatrice et un travail en groupe. » (Pamula, 2008 : 138).

La catégorisation phonémique est perçue par les experts comme meilleure pour le GE ayant bénéficié d'une expérience musicale ciblée sur /y/. Patel (2011, modèle OPERA), affirme que la formation musicale améliore le codage neuronal de la parole si cinq fonctions sont remplies **O**: Overlap (chevauchement anatomique des réseaux cérébraux traitant une caractéristique acoustique utilisée à la fois en musique et parole) ; **P**: Précision (la musique impose des exigences plus élevées sur ces réseaux partagés que la parole, en termes de précision de traitement) ; **E**: Émotion (les activités musicales suscitent une forte émotion positive) ; **R**: Répétition (ces activités musicales sont souvent répétées) ; **A**: Attention (la musique est associée à une attention ciblée). Il semble que la pratique des comptines dont a bénéficié le GE réunisse les conditions O, E (chanter suscite plus de plaisir que parler, même en rythme), R et A (*Une* et *sur*, attirent l'attention par leur note plus aigue[1]).

Cette étude n'est qu'une amorce des recherches sur ce sujet, qui pourront être poursuivies par : 1) une analyse plus approfondie des résultats perceptifs en fonction des profils de chaque enfant ; 2) une étude sur de plus grands échantillons avec ajout d'analyses statistiques, 3) une analyse acoustique de ces productions et de celles de francophones natifs du même âge (F2 et F3 distinguant /i/, /y/ et /u/) ; 4) l'analyse approfondie de nos données concernant les effets de l'utilisation d'autres comptines sur les autres phonèmes travaillés, /ʒ/ et /ʁ/ ; 5) l'analyse à plus long terme de ce travail spécifique sur ces comptines, plusieurs mois après la fin de cet entraînement spécifique.

Remerciements

Ce travail a bénéficié / bénéficié partiellement d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme "Investissements d'Avenir" portant la référence ANR-10-LABX-0083. Il contribue à l'IdEx Université de Paris - ANR-18-IDEX-0001.

Références

BERRUTO, G. (1987). *Sociolinguistica dell'italiano contemporaneo*, Roma, Carocci editore.

BOERSMA, P. & WEENINK, D. (2018). Praat: doing phonetics by computer, Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

CALLAMAND, M. (1981). *Méthodologie de l'enseignement de la prononciation : organisation de la matière phonique du français et correction phonétique*, Paris, CLE International.

CALVET, L.-J. (1980). *La chanson dans la classe de français langue étrangère*, Paris, CLE International.

CORNAZ, S. & CAUSSADE, D. (2014). Musique, voix chantée et apprentissage: une revue de littérature et quelques propositions d'exploitation en didactique de la phonétique des langues, *Revue Electronique du Centre de Recherche sur les Identités Nationales et l'Interculturalité (e-CRINI)*, 6, 1-34. HAL : hal-01242980

DELASALLE, D. (2005). Repères à l'oral et passage phonie/graphie : comment aider l'élève?, In Delasalle, D. (dir.), *L'apprentissage des langues à l'école : diversité des pratiques*, tome I, Paris, L'Harmattan, p. 104-113.

DI CRISTO, A. (1975). Présentation d'un test de niveau destiné a évaluer la prononciation des anglophones. *Revue de Phonétique Appliquée*, 33-34, pp. 9-35. ERIC NUMBER : EJ124364

FERRERO, F. (1972). Caratteristiche acustiche dei fonemi vocalici italiani, *Parole e metodi*, 3, 9-31.

GARDNER, H. (2001). *Les intelligences multiples*, Paris, Retz.

KREIMAN, J., GERRATT, B.R., KEMPSTER, G.B., ERMAN, A., & BERKE, G.S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, *36*(1), 21-40. DOI : 10.1044/jshr.3601.21

MARTIN, C., TRÉSALLET, E. (1998). *30 phonèmes en 30 chansons*, Paris, Retz.

PAMULA, M. (2008). Sensibiliser les enfants à une langue étrangère par le biais d'une activité musicale, *Synergies Espagne*, 1, pp. 133-140.

PATEL, A.D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*, *2*, 142, 1-14. DOI : 10.3389/fpsyg.2011.00142

ROBERGE, C. (2003). Chapitre 14. Les enfants : « Le Français dans l'espace » (Réflexions à propos d'une expérience d'enseignement précoce), In RENARD, R. (éd.) *Apprentissage d'une langue étrangère / seconde. Vol. 3. La méthodologie*, Bruxelles, De Boeck Université, p. 311-321.

ROGERS, D. & D'ARCANGELI, L. (2004). Italian. *Journal of the International Phonetic Association*, *34*(1), 117-121. DOI: 10.1017/S0025100304001628

SAMSON, C. (2010). *Alex et Zoé et compagnie 1* : chansons et comptines, Paris, CLE international.

SCHÖN, D., BOYER, M., MORENO, S., BESSON, M., PERETZ, I., KOLINSKY, R. (2008). Songs as an aid for language acquisition, *Cognition*, pp. 975–983. DOI: 10.1016/j.cognition.2007.03.005

TELLIER, M. (2010). Faire un geste pour l'apprentissage: le geste pédagogique dans l'enseignement précoce. Impact sur le développement de la langue maternelle. In C. CORBLIN, *L'enseignement des langues vivantes étrangères à l'école*, L'Harmattan, pp.31-54. HAL : hal-00541985

VANTHIER, H. (2009). *L'enseignement aux enfants en classe de langue*, Paris, CLE International.

WATOREK, M. & WAUQUIER-GRAVELINES, S. (2016). Diversité d'approches et de méthodes en acquisition des langues secondes, *Revue française de linguistique appliquée,* 2016/2, XXI, 5-17. DOI : 10.3917/rfla.212.0005

ZEDDA, P. (2006). La langue chantée : un outil efficace pour l'apprentissage et la correction phonétique, *Cahiers de l'Acedle,* 2, p. 257-282.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Évaluation de systèmes apprenant tout au long de la vie

Yevhenii Prokopalo[1] Sylvain Meignier[1]
Olivier Galibert [2] Loïc Barrault [3] Anthony Larcher[1]
(1) LIUM, 72 Le Mans, France, (2) LNE, 78 Trappes, France, (3) University of Sheffield, UK
yevheniiprokopalo@univ-lemans.fr

RÉSUMÉ

Aujourd'hui les systèmes intelligents obtiennent d'excellentes performances dans de nombreux domaines lorsqu'ils sont entraînés par des experts en apprentissage automatique. Lorsque ces systèmes sont mis en production, leurs performances se dégradent au cours du temps du fait de l'évolution de leur environnement réel. Une adaptation de leur modèle par des experts en apprentissage automatique est possible mais très coûteuse alors que les sociétés utilisant ces systèmes disposent d'experts du domaine qui pourraient accompagner ces systèmes dans un apprentissage *tout au long de la vie*. Dans cet article nous proposons un cadre d'évaluation générique pour des systèmes apprenant tout au long de la vie (SATLV). Nous proposons d'évaluer l'apprentissage assisté par l'humain (actif ou interactif) et l'apprentissage au cours du temps.

ABSTRACT

Evaluation of lifelong learning systems

Current intelligent systems need the expensive support of machine learning experts to sustain their performance level when used on a daily basis. To reduce this cost, i.e. remaining free from any machine learning expert, it is reasonable to implement lifelong (or continuous) learning intelligent systems that will continuously adapt their model when facing changing execution conditions. In this work, the systems are allowed to refer to human domain experts who can provide the system with relevant knowledge about the task. Nowadays, the fast growth of lifelong learning systems development rises the question of their evaluation. In this article we propose a generic evaluation methodology for the specific case of lifelong learning systems. Two steps will be considered. First, the evaluation of human-assisted learning (including active and/or interactive learning) outside the context of lifelong learning. Second, the system evaluation across time, with propositions of how a lifelong learning intelligent system should be evaluated when including human assisted learning or not.

MOTS-CLÉS : Apprentissage automatique, évaluation, apprentissage tout au long de la vie.

KEYWORDS: Machine learning, Lifelong learning, Evaluation.

1 Introduction

Les systèmes intelligents utilisent une représentation du monde, un modèle, dont l'apprentissage est réalisé en laboratoire par des experts en apprentissage automatique (EAA) (Bishop, 2006). Le rôle de ces experts est triple : (1) ils sélectionnent et annotent les données d'apprentissage ; (2) ils déterminent les méta-paramètres inhérents à la structure du système en utilisant des données de développement ;

Volume 1 : Journées d'Études sur la Parole, pages 516–524.
HAL : hal-02798580.

(3) ils évaluent et comparent les systèmes afin de déterminer lequel mettre en production sur un jeu de données de test. Dans cet article, nous appellerons "données initiales" l'ensemble de ces trois jeux de données. Une fois en production, ces systèmes sont confrontés aux données réelles. Le cycle de vie d'un tel système est illustré sur la figure 1-A (Chen & Liu, 2016). Si il est raisonnable de considérer que les données d'entrée du système lors de sa mise en production sont proches des données initiales, il arrive la plupart du temps qu'elles s'en éloignent avec le temps. Cet effet est bien connu et entraîne de sévère dégradations de performances (Quionero-Candela *et al.*, 2009) qui ne peuvent être résolues que grâce à l'intervention d'un expert qui entraînera un nouveau modèle. Afin d'éviter le recours couteux à cet expert, la communauté scientifique se mobilise pour développer des systèmes qui apprennent seuls au cours du temps, sans l'aide d'un expert en apprentissage automatique mais en adaptant continuellement leur modèle à l'aide des données disponibles et la possible intervention d'un expert humain du domaine. Ces systèmes sont appelés des systèmes apprenant tout au long de la vie (SATLV). Les SATLV diffèrent des systèmes à apprentissage statique car ils disposent de modules d'adaptation qui doivent permettre de maintenir leurs performances au cours du temps en apprenant des nouvelles données d'entrée. Le cycle de vie d'un système SATLV est décrit par la figure 1-B.

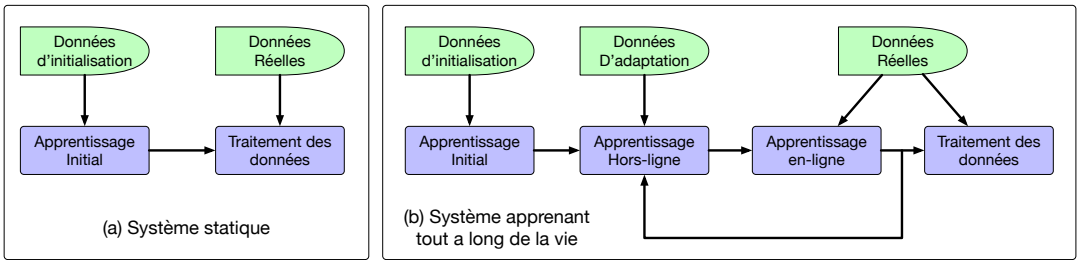


FIGURE 1: Comparaison des cycles de vie d'un système à apprentissage statique et d'un système apprenant tout au long de la vie.

Les SATLV peuvent adapter leur modèle selon deux modes : en-ligne et hors-ligne. Dans ce travail, nous définirons l'adaptation en-ligne comme l'action d'un système qui met à jour ses connaissances (le modèle) en apprenant des données d'entrée qu'il doit traiter. En d'autres termes, il est possible pour le système d'utiliser toutes les données d'entrée qui lui sont fournies lors de son exploitation. Dans ce cas, le SATLV reçoit les mêmes données que n'importe quel système à apprentissage statique.

L'adaptation hors-ligne permet à un SATLV d'utiliser l'ensemble des données qui lui sont accessibles pour améliorer son modèle. Ces données incluent entre autre les données initiales, que le système peut décider de réutiliser, des données additionnelles qui peuvent être fournies par l'expert humain du domaine, des données collectées automatiquement par le SATLV et même des données réelles que le SATLV a dû traiter par le passé.

Lors d'une adaptation hors-ligne, le SATLV n'a pas à générer d'hypothèse (de sortie) et peut bénéficier de ressources de calcul importantes. Dans la suite de cet article, nous supposerons que les données d'adaptation en-ligne sont fournies sans aucune annotation tandis que les données d'adaptation hors-ligne peuvent inclurent des données annotées et des données non-annotées.

Les processus d'adaptation en-ligne et hors-ligne peuvent utiliser des méthodes d'adaptation non-supervisées ou assistées par l'humain. L'apprentissage assisté par l'humain inclue l'apprentissage actif (AA ; pour lequel le système prend l'initiative d'engager un échange avec l'humain) et de l'apprentissage interactif (AI ; dans lequel l'humain est à l'initiative de cet échange).

L'apprentissage non-supervisé permet au système d'adapter son modèle en exploitant les données

d'entrée sans annotation. L'AA autorise le système à demander à l'humain de corriger une partie de l'hypothèse qu'il a générée et l'AI permet à l'humain de renvoyer au système des corrections de son hypothèse courante afin que le système intègre ces corrections et génère une nouvelle hypothèse.

Les SATLV, comme les systèmes à apprentissage statique doivent être évalués avant leur mise en production. Cette évaluation prend souvent la forme d'une comparaison entre systèmes respectant un protocole strict et contrôlé. Notons que le présent article ne traite que de l'évaluation des systèmes automatiques et non de leur développement.

L'évaluation des systèmes à apprentissage statique suit une méthodologie bien établie et requiert un jeu de données d'évaluation, une métrique objective et un protocole d'évaluation. Pour de très nombreuses tâches, ces éléments ne sont pas disponibles pour l'évaluation de systèmes SATLV et la dérivation des modalités d'évaluation des systèmes statiques pour permettre l'évaluation de SATLV effectuant les mêmes tâches n'est pas simple. L'évaluation des SATLV nécessite de pouvoir évaluer les performances du système au cours du temps ainsi que les performances des modules d'AA et d'AI individuellement. Cette évaluation doit également prendre en compte la chronologie des données. En effet, si les systèmes automatiques traitent la plupart du temps les données actuelles, il est possible que pour certaines tâches spécifiques, ce système ait à traiter des données issues du passé. Les performances de ce système sur des données actuelles ou des données du passé peuvent avoir une incidence différente pour l'utilisateur. Nous introduisons dans la section 3.1 une politique-utilisateur que la métrique d'évaluation doit prendre en compte pour évaluer la capacité du SATLV à s'adapter selon la politique dictée par l'utilisateur.

Le travaux existant pour l'évaluation de tels systèmes font souvent état d'un manque de protocoles et de données nécessaires et proposent de simuler la chronologie ou l'évolution des données en permutant les données de corpus existant (Kemker *et al.*, 2018; Lomonaco & Maltoni, 2017). Il apparaît également que les performances de tels systèmes sont évaluées à plusieurs instant dans le temps ou sur des données présentant de nouveaux événement mais qu'aucune évaluation de leur comportement temporel n'est rapportée (Parisi *et al.*, 2019). Dans cet article [1], nous proposons des métriques et protocoles pour l'évaluation de l'apprentissage assisté par l'humain pour les cas d'apprentissage hors-ligne et en-ligne (sections 2.1 et 2.2) (Prokopalo *et al.*, 2020). Nous proposons également une métrique pour évaluer la capacité des SATLV à adapter leur modèle pour respecter la politique fixée par l'utilisateur au cours du temps. Une métrique d'évaluation de l'apprentissage assisté par l'humain est introduite dans la section 2. Conscients que les métriques d'évaluation et les protocoles sont toujours limités et ne permettent pas d'évaluer tous les aspects de systèmes complexes, nous discutons des limitations de nos propositions dans la section 4

L'ensemble des propositions de cet article considèrent que les systèmes à apprentissage statiques sont évalués par des métriques similaires à des taux d'erreurs (le plus faible le meilleur) mais il est important de noter que les solutions proposées peuvent être appliquées à toute métrique scalaire sans perdre de leurs généralités.

2 Évaluation de l'apprentissage assisté par l'humain

Pour supprimer la nécessité de faire intervenir un expert en apprentissage automatique, il est raisonnable de faire appel à un expert humain du domaine (EHD). Dans cette section, nous ne considérons

1. ce travail a fait l'objet d'une publication à LREC 2020

pas l'évaluation au cours du temps mais seulement l'apprentissage assisté par l'humain. L'interaction entre l'EHD et le SATLV peut avoir lieu hors-ligne ou en-ligne.

Hors-ligne, l'EHD interagit avec le SATLV sur les données disponibles sans considérer de contrainte de temps ou de calcul. Dans ce contexte, les SATLV sont libres d'apprendre de n'importe quelles données non-annotées et ont l'opportunité de poser à l'EHD un nombre limité de questions à propos de ces données. Durant cet apprentissage actif, le système doit poser les questions qui maximisent la généralisation des réponses apportées par l'utilisateur. La section 2.1 décrit une façon d'évaluer l'apprentissage actif.

En-ligne, l'EHD est partie prenante de la chaîne de production, ses interactions avec le SATLV ont lieu entre l'arrivée des données à traiter et l'envoie de l'hypothèse finale à l'utilisateur. Dans ce travail, nous restreignons le cadre d'interaction de l'EHD aux seules données à traiter (sans considérer que l'EHD peut apporter une information sur des données additionnelles). Étant donné un ensemble X de données à traiter, le SATLV peut poser des questions à l'EHD. Il est également possible pour l'EHD de surveiller le traitement automatique et de fournir des informations au système au cours du processus de traitement.

Afin d'évaluer les processus d'adaptation hors-ligne et en-ligne, l'EHD doit être simulé par un système qui assure la reproductibilité des tests.

2.1 Apprentissage hors ligne assisté par l'humain

L'efficacité de l'apprentissage assisté par l'humain réside dans la capacité du système à exploiter au mieux les interactions avec l'humain pour en minimiser le coût. De nombreux articles dans la littérature conseillent de mesurer la qualité de l'apprentissage assisté par l'humain en fonction du coût des interactions et des performances du système (Krogh & Vedelsby, 1995; Siddhant & Lipton, 2018; Drugman *et al.*, 2019; Beluch *et al.*, 2018; Pérez-Dattari *et al.*, 2018; Celemin & Ruiz-del Solar, 2019). Cette mesure nous semble pertinente, ainsi nous ne proposons pas de nouveauté dans ce domaine et nous contentons ici de décrire cette métrique. Étant donné un modèle initial, un processus itératif est initié par le système avec pour but de réduire le taux d'erreurs obtenu sur un jeu de données. Le système peut poser des questions relatives à n'importe quel document auquel il a accès et ses performances sont évalués sur un jeu de données de test qui ne peut pas être utilisé pour adapter son modèle. Dans ce contexte, il n'est pas certain que le taux d'erreurs diminue et atteigne zéro au cours de l'apprentissage. Pour cette raison, un coût maximum d'interaction est fixé. Une fois atteint ce coût, l'apprentissage actif est interrompu. Il est important de noter que lors d'un apprentissage hors-ligne, le système n'a pas à traiter de données en particulier et l'apprentissage assisté par l'humain se trouve donc réduit à l'apprentissage actif. Le cas où le EHD initie l'interaction peut être vu comme de l'apprentissage faiblement supervisé ou supervisé.

2.2 Apprentissage en ligne assisté par l'humain

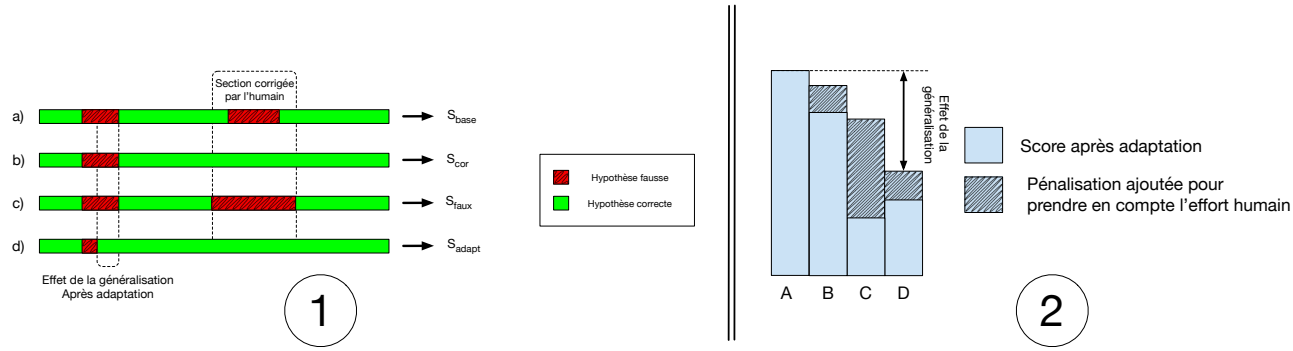
Lors d'un apprentissage en-ligne, le système doit traiter une séquence de données réelles et renvoyer les hypothèses correspondantes. En fonction du scenario, deux types d'interactions sont possibles : soit le SATVL initie une session d'apprentissage actif avant de renvoyer l'hypothèse finale, soit l'EHD initie une session d'apprentissage interactif en corrigeant itérativement les hypothèses générées par le système.

Dans le cas d'un apprentissage interactif, il est raisonnable de considérer que l'EHD est en mesure de

fournir des corrections jusqu'à obtenir une hypothèse entièrement correcte. L'efficacité de l'apprentissage assisté par l'humain peut alors être mesurée par le coût de supervision minimal présenté dans (Geoffrois, 2016).

Différents systèmes peuvent autoriser différentes modalités d'interaction. Il est donc difficile d'estimer le coût de l'interaction humaine de façon homogène et comparable entre systèmes. (Broux *et al.*, 2018). Dans l'idéal, une mesure de performance produirait une comparaison équitable de systèmes offrant des modalité d'interaction diverses. Nous proposons pour ceci de mesurer le coût d'interaction sous forme d'une métrique scalaire donnée dans la même unité que le score de performance du système (taux d'erreur par exemple). Cette pénalité peut s'appliquer à l'apprentissage actif ou interactif. Calculer une pénalité dans la même unité que la performance du système permet de calculer un unique score pénalisé qui reflète à la fois la performance finale du système et l'effort humain nécessaire pour l'obtenir.

Une première option consiste à calculer le coût d'interaction comme la quantité d'information donnée au système par l'utilisateur. Cependant, cette stratégie ne fonctionne pas pour des métriques non linéaires comme par exemple BLEU pour la traduction automatique (Papineni *et al.*, 2002). Nous proposons de calculer la pénalité comme la part du score correspondant aux données corrigées pas l'humain. Afin de calculer cette quantité, nous calculons deux scores : un score corrigé, S_{cor}, et un score dégradé, S_{faux}, dont le calcul est décrit par la figure 2-1.



Une hypothèse (a), produite par un système automatique contenant une partie correcte (vert) et des erreurs (rouge) obtient un score S_{base}. Lors de l'apprentissage assisté par l'humain, une section de l'hypothèse (partiellement fausse) est corrigée. Pour calculer l'effort fourni par l'humain, un nouveau scores S_{cor} est calculé immédiatement après application de la correction (b). Nous introduisons un score S_{faux} qui est calculé en remplaçant la section corrigée par une hypothèse entièrement fausse (c). Après avoir reçu une information provenant de l'humain, le système adapte son modèle et produit une nouvelle hypothèse qui obtient un score S_{adapt}. Un système capable de généraliser l'information reçue devrait améliorer l'hypothèse finale sur des portions où l'utilisateur n'est pas intervenu manuellement (d).

Effet de la pénalisation dans différents cas. Le score A, S_{base}, est obtenu sur l'hypothèse initiale, avant toute intervention de l'humain. Les trois autres colonnes représentent des scores pénalisés, S_{pen}, obtenus en utilisant différentes méthodes d'apprentissage assisté par l'humain. La partie uniforme des colonnes B, C et D représente le score finale après adaptation, S_{adapt}. La partie hachurée représente la pénalité. Le score (B) illustre le cas où le système exploite peu l'information apportée par l'humain. Dans le cas (C), le système obtient un gain significatif au prix d'un effort humain important. Dans le cas (D), le système obtient un taux d'erreur final plus élevé que pour (C) mais l'effort humain nécessaire est beaucoup plus faible. Notons que la différence entre S_{base} et S_{pen} (indiqué pour le score D) correspond au gain obtenu par le système qui a généralisé les corrections appliquées par l'humain.

FIGURE 2: Pénalisation de l'apprentissage assisté par l'humain.

Un système produit une première hypothèse (Figure 2-1-A) et obtient un score S_{base} avant toute interaction avec l'EHD. L'EHD fournit alors des informations en corrigeant une partie de l'hypothèse courante. La partie des données corrigée est indiquée sur la figure 2-1-B et cette correction permet d'obtenir un score S_{cor}. En fonction de la tâche, il est possible qu'une partie de l'hypothèse qui est corrigée ait été en partie correcte. (Dans le cas de la traduction il peut s'agir de la correction d'une phrase qui contenait déjà certains mots corrects). La différence entre S_{base} et S_{cor} correspond à l'amélioration apportée par la seule correction mais ne reflète pas le coût de la correction. Nous calculons alors un nouveau score S_{faux} qui est obtenu en remplaçant l'intégralité de la partie corrigée de l'hypothèse par une hypothèse erronée. (Figure 2-1-C) La différence $S_{faux} - S_{cor}$ mesure la part

du score qui correspond au coût réel de la correction. Finalement, l'hypothèse corrigée est renvoyée au système qui adapte son modèle et re-génère une nouvelle hypothèse qui obtiendra le score S_{adapt}. Le scores pénalisé, S_{pen}, est obtenu de la façon suivante :

$$S_{pen} = S_{adapt} + (S_{faux} - S_{cor}) \quad (1)$$

Notons qu'un système qui ne prendrait pas la correction en compte ou demanderait une information déjà correcte serait pénalisé deux fois : une fois par la valeur élevée de S_{adapt} et une fois par le second terme de l'équation 1.

L'effet de cette pénalisation est illustré par la figure 2-2. Le score S_{base}, illustré par le score A est obtenu avant toute interaction humaine. À partir de cette hypothèse, différentes stratégies d'adaptation pourraient mener à différents scores pénalisés, illustrés par les barres B, C et D de la figure 1. Les parties uniformes de ces barres illustrent le score S_{adapt} tandis que les parties hachurées illustrent la pénalité appliquée à chaque version du système. Le système idéal obtiendra un score pénalisé très faible tout en nécessitant une petite quantité d'interactions.

3 Évaluation au cours du temps

Une fois le modèle initial, M_0, d'un système SATLV entrainé et testé en laboratoire, le système est mis en production et traite une séquence de données, $[X_{l_i}]_{i=1...T}$, lui parvenant au cours du temps. Cette séquence est ici nommée *données au cours du temps* et l_i est le temps auquel X_{l_i} a été produit. Le SATLV peut adapter son modèle après avoir traité des données X_{l_i} pour obtenir une nouvelle version M_i. Les performances du SATLV sont évaluées sur les données X_{l_i} et la séquence de scores obtenue est $[S(M_i, l_i)]_{i=1...T}$.

3.1 Prise en compte d'une politique fixée par utilisateur

Les données que les SATLV doivent traiter évoluent au cours du temps mais il arrive qu'un système rencontre des données semblables à des données vues dans le passé. L'adaptation du modèle aux données actuelles peut entraîner « l'oubli catastrophique »des données du passé au sens de (French, 1999). Le coût de cet oubli pour l'utilisateur dépend de l'application visée et il est essentiel que l'utilisateur puisse l'indiquer au SATLV afin d'adapter la politique d'adaptation du SATLV. Différentes politiques sont illustrées sur la figure 3

Pour l'évaluateur du système, il est essentiel d'évaluer la capacité du SATLV à adapter son modèle de façon à répondre au mieux aux exigences de l'utilisateur. Afin d'évaluer la capacité du SATLV à suivre une politique d'adaptation donnée, une nouvelle séquence de données, $[Y_{t_j}]_{j=1...N}$, est utilisée. La date de création de ces données doit couvrir la même période que les données sur lesquelles le système a été adapté. Il faut que : $l_1 < t_1 < t_N < l_T$. Ainsi, chaque version M_i du modèle peut être évaluée sur l'ensemble de la séquence de test Y_{t_j} pour obtenir une séquence de scores $[s(M_i, t_j)]_{j=1...N}$. Idéalement, ces séries temporelles de données devraient être strictement alternées mais de nombreux critères sur la nature des données et les événements rencontrées au cours du temps doivent être pris en compte afin de produire des études objectives. Notons que cette séquence de scores est calculée pour une unique version, M_i, du modèle. Les performances de ce système sur les données antérieures à l_i montrerons à quel point le système a oublié le passé. La figure 3-1 illustre la

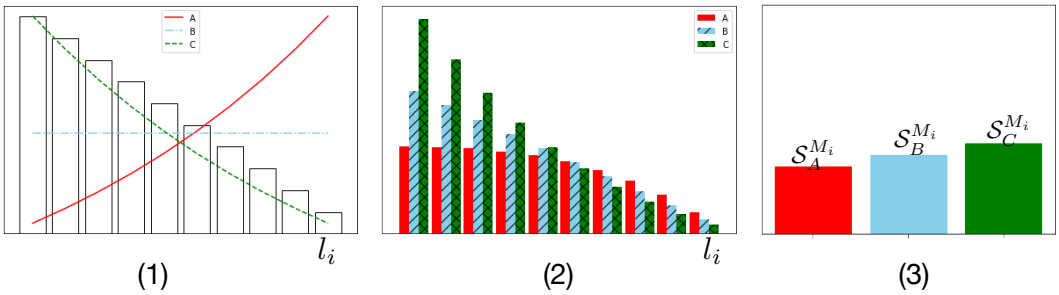


FIGURE 3: Scores obtenus par un système apprenant au cours du temps et 3 politiques d'adaptation différentes (1). Scores pondérés au cours du temps (2). Score final résultat de la somme pondérée des scores au cours du temps pour les 3 politiques (3).

séquence de scores obtenue par la version M_i du système dont le modèle a été adapté jusqu'à l_i. Cet exemple factice illustre le cas d'un système optimisé pour les données les plus récentes mais dont les performances sur les données du passé sont moins bonnes. La même figure montre trois politiques d'adaptation : (A) l'utilisateur désire que le système privilégie les données actuelles ; (B) les données du passé sont aussi importantes que les données actuelles : (C) l'utilisateur privilégie les données du passé. Évaluer la politique d'adaptation pour ce système consiste à pondérer les scores obtenus au cours du temps par la politique définie par l'utilisateur (Figure 3-2) La somme de ces scores pondérés est un indicateur des performances du sytème mais également de sa capacité à satisfaire la politique d'adaptation, $\mathcal{P}$, fixée.

$$\mathcal{S}_{\mathcal{P}}^{M_i} = \sum_{t_j \in [l_1; l_i]} s(M_i, t_j) \cdot \mathcal{P}_{t_j} \tag{2}$$

On observe sur la figure 3-3, que le système factice obtient un meilleur score pour la politique qui privilégie les données actuelles, comme attendu.

3.2 Évaluation au cours du temps de l'apprentissage assisté par l'humain

Dans ce travail nous considérons que l'adaptation hors-ligne au cours du temps représente un coût fixe pour le système car elle nécessite la présence permanente d'un expert humain, mais ne modifie pas la disponibilité du système. En revanche, l'adaptation en-ligne influe sur le temps de réponse du système qui effectue des cycles d'adaptation entre la réception des données et la communication de l'hypothèse au client. Ainsi nous proposons d'intégrer la coût de l'adaptation en-ligne en remplaçant, dans l'équation 2, le score $s(M_i, t_j)$ par sa version pénalisée proposée dans la section 2.2. Le score pénalisé et pondéré par la politique d'adaptation devient alors :

$$\mathcal{S}_{\mathcal{P}}^{M_i} = \sum_{t_j \in [l_1; l_i]} S_{pen}(M_{ij}, t_j) \cdot \mathcal{P}_{t_j} \tag{3}$$

4 Discussion

Dans cet article, nous proposons trois contributions pour évaluer les systèmes automatiques apprenant tout au long de la vie (SATLV). Nous proposons une nouvelle mesure du coût de l'interaction humaine dans l'apprentissage assisté par l'humain, donnée dans la même unité que la mesure de performance du système. Nous introduisons le concept de politique d'adaptation afin d'évaluer la capacité des

SATLV à respecter une politique définie par l'utilisateur. Enfin, nous avons combiné ces éléments pour proposer un cadre complet d'évaluation des systèmes SATLV. Notre travail considère des métriques de type « taux d'erreur »mais peut également être appliqué à des métriques scalaires définies sur un intervalle semi-ouvert.

La pénalité proposée pour sanctionner le coût de l'interaction humaine dans l'adaptation du modèle est exprimée dans la même unité que la mesure de performance du système afin d'être directement combinée et d'offrir un indicateur de performance unique. Cette pénalité permet d'évaluer la capacité du système à généraliser l'information reçue de l'expert du domaine. Dans certains cas, le calcul du score intermédiaire, S_{faux}, pose des questions complexes. Par exemple dans le cas de la traduction automatique, la définition de la traduction la plus fausse est très complexe car il est toujours possible de dégrader le scores en insérant des mots. Une borne raisonnable consiste à limiter la taille de l'hypothèse fausse à la taille de la référence. Ce calcul sera appliqué dans une tâche de l'évaluation internationale WLT 2020 dédié aux SATLV. Le calcul de la pénalité considère que l'information fournie par l'humain peut être directement appliquée à l'hypothèse pour calculer un nouveau score, S_{cor}. Cette hypothèse n'est pas toujours réaliste car elle dépend des interactions autorisées.

Afin de garantir la reproductibilité et l'équité de l'évaluation de l'adaptation assistée par l'humain, il est nécessaire d'implémenter une simulation d'expert humain. Différentes implémentations sont possibles et tous les systèmes comparés devront interagir avec le même simulateur. Idéalement, plusieurs simulateurs pourraient être utilisés afin de confronter les systèmes à différents cas de figure. Le développement de ces simulateur est un sujet de recherche en soi qui n'est pas l'objet de cet article.

Notre deuxième contribution est l'introduction d'une politique d'adaptation. Cette politique permet à l'utilisateur d'introduire une connaissance du domaine relative à l'évolution des données au cours du temps (par exemple le cas ou un système rencontrera des données qui varient da façon cyclique). La définition d'une telle fonction par un utilisateur non-expert en apprentissage automatique nécessite une interface utilisateur intuitive et bien définie qui peut être difficile à évaluer en elle même.

Dans ce travail, nous avons tenté de définir un cadre général d'évaluation qui pourrait être transposé à un grand nombre de tâches et de mesures de performance. Comme c'est le cas pour de nombreuses métriques existante et utilisées, nos propositions ne répondent pas à tous les besoins et doivent être combinées à d'autres afin d'évaluer tous les aspects de système complexes par définition.

Enfin les métriques et protocoles définis dans ce travail ont été appliqués au cas de la segmentation en locuteur et de la traduction automatique et seront utilisés lors d'évaluations internationales organisées en 2020 : ALLIES/ALbayzin 2020 et WMT2020. À la fin de ces évaluations, les métriques, protocoles et données collectées pour ces évaluations seront distribuées gratuitement pour les fins de recherche.

Remerciements

Ce travail a été financé par le projet Chist-ERA ALLIES (ARN-17-CHR2-0004-01) [2]

2. https://lium.univ-lemans.fr/allies/

Références

BELUCH W. H., GENEWEIN T., NÜRNBERGER A. & KÖHLER J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 9368–9377.

BISHOP C. M. (2006). *Pattern recognition and machine learning*. springer.

BROUX P.-A., DOUKHAN D., PETITRENAUD S., MEIGNIER S. & CARRIVE J. (2018). Computer-assisted speaker diarization : How to evaluate human corrections. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

CELEMIN C. & RUIZ-DEL SOLAR J. (2019). An interactive framework for learning continuous actions policies based on corrective feedback. *Journal of Intelligent & Robotic Systems*, **95**(1), 77–97.

CHEN Z. & LIU B. (2016). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **10**(3), 1–145.

DRUGMAN T., PYLKKONEN J. & KNESER R. (2019). Active and semi-supervised learning in asr : Benefits on the acoustic and language models. *arXiv preprint arXiv :1903.02852*.

FRENCH R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, **3**, 128–135.

GEOFFROIS E. (2016). Evaluating interactive system adaptation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, p. 256–260.

KEMKER R., MCCLURE M., ABITINO A., HAYES T. L. & KANAN C. (2018). Measuring catastrophic forgetting in neural networks. In *Thirty-second AAAI conference on artificial intelligence*.

KROGH A. & VEDELSBY J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems*, p. 231–238.

LOMONACO V. & MALTONI D. (2017). Core50 : a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, p. 17–26.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)* : ACL.

PARISI G. I., KEMKER R., PART J. L., KANAN C. & WERMTER S. (2019). Continual lifelong learning with neural networks : A review. *Neural Networks*.

PÉREZ-DATTARI R., CELEMIN C., RUIZ-DEL SOLAR J. & KOBER J. (2018). Interactive learning with corrective feedback for policies based on deep neural networks. *arXiv preprint arXiv :1810.00466*.

PROKOPALO Y., MEIGNIER S., GALIBERT O., BARRAULT L. & LARCHER A. (2020). Evaluation of lifelong learning systems. In *International Conference on Language Resources and Evaluation (LREC)*.

QUIONERO-CANDELA J., SUGIYAMA M., SCHWAIGHOFER A. & LAWRENCE N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press.

SIDDHANT A. & LIPTON Z. C. (2018). Deep bayesian active learning for natural language processing : Results of a large-scale empirical study. *arXiv preprint arXiv :1808.05697*.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La voix actée : pratiques, enjeux, applications

Mathias Quillot[3] Lauriane Guillou[2] Adrien Gresse[3] Rafaël Ferro[1]
Raphaël Roth[2] Damien Malinas[2] Richard Dufour[3], Axel Roebel[1]
Nicolas Obin[1], Jean-François Bonastre[3], Emmanuel Ethis[2]
(1) STMS Lab - Ircam, CNRS, Sorbonne Université, Paris, France
(2) Laboratoire Culture et Communication, Avignon Université, Avignon, France
(3) Laboratoire d'Informatique d'Avignon, Avignon Université, Avignon, France

Résumé

La voix actée représente un défi majeur pour les futures interfaces vocales avec un potentiel d'application extrêmement important pour la transformation numérique des secteurs de la culture et de la communication, comme la production ou la post-production de voix pour les séries ou le cinéma. Un aspect central de la voix actée repose sur la notion d'interprétation, un aspect peu étudié dans la communauté scientifique de la parole. Cet article propose un état des lieux et une réflexion sur les défis scientifiques et les applications technologiques de la voix actée : à la croisée de l'acoustique, de la linguistique, de la culture, et de l'apprentissage machine. Une analyse préliminaire des pratiques permet de rendre compte de la diversité de l'écosystème des "*métiers de la voix*" et de pointer les fonctions et les conventions qui s'y rattachent. Nous nous intéresserons ensuite à la pratique particulière du doublage de voix, en faisant ressortir ses enjeux et problématiques spécifiques puis en présentant des solutions proposées pour modéliser les codes expressifs de la voix d'un acteur ou les choix d'un opérateur pour le doublage.

Abstract

Acted voice : practices, challenges, applications
The acted voice represents a major challenge for the next generation of voice interfaces with an extremely important application potential for the digital transformation of cultural and communication sectors, such as production or post-production of voice for series or films. A central aspect of the acted voice relies on the notion of interpretation, an aspect that has been little studied in the speech community. This article offers an overview and a reflection on the scientific challenges and technological applications of acted voice: at the crossroads of acoustics, linguistics, culture, and machine learning. A preliminary analysis of practices is presented to account for the diversity of the "voice professions" ecosystem and to point out the functions and conventions associated with it. The remaining of the paper focuses on the specific practice of voice dubbing, highlighting its specific issues and presenting some solutions proposed to model the expressive codes of an actor or the choices of an operator for dubbing.

Mots-clés : voix actée, interprétation, analyse des pratiques, doublage
Keywords: acted speech, interpretation, practice analyse, dubbing

Volume 1 : Journées d'Études sur la Parole, pages 525–533.
HAL : hal-02798582.

1 Introduction

Les avancées spectaculaires réalisées cette décennie en traitement automatique de la parole, pour des tâches comme la reconnaissance ou la synthèse de parole, ont permis la réalisation de nombreuses applications devenues d'usage quotidien, comme l'interaction vocale avec un smartphone. Tour à tour et successivement voix de "laboratoire", "pathologique", "ordinaire", "spontanée", "expressive" ou "conversationnelle", la recherche sur la parole s'attaque à des domaines variés et s'affranchit peu à peu des défis posés par des facteurs de variabilité de plus en plus nombreux et complexes. Par contraste, l'étude de la parole "actée" demeure aujourd'hui marginale dans la communauté de la parole au point où un article sur le doublage était encore catalogué il y a quelques années dans la catégorie "parole anormale" ("abnormal speech"). Contrairement à la parole "ordinaire" ou "spontanée", la voix "actée" est un artifice, une construction fruit d'une interprétation maîtrisée et planifiée de sa voix par un acteur afin de produire un effet désiré chez un spectateur, par exemple rendre manifeste le comportement d'un personnage fictif, et faciliter la crédibilité et l'immersion du public dans une situation ou une trame. La voix est souvent dénaturée, grossie, pour rendre audible et sensible les effets expressifs produits par le comédien. Ainsi, la voix actée élargit le champs de recherche des fonctions expressives de la voix humaine et ouvre sur le domaine encore peu étudié de l'*interprétation*. Par ailleurs, une brève revue des "*métiers de la voix*" (voix-off pour le commentaire, publicité, voix doublée, voix au théâtre, voix au cinéma, etc…) montre la diversité des pratiques et les usages spécifiques de la voix principalement dans le secteur de la culture et de la communication, et des industries créatives. Pour comprendre les fonctions de la voix dans ces usages, la voix ne peut être considérée seule, mais dans une relation à un être humain dont elle ne constitue que l'une des modalités d'expression : l'incarnation de la voix par un corps (les voix-off sont ainsi totalement désincarnées) ou l'adéquation d'une voix et de son corps (le doublage offre un exemple de composition du corps d'un acteur réel ou fictif et de la voix d'un autre acteur) (Le Breton, 2011).

L'appréhension de la voix actée soulève à la fois un ensemble de questions inédites pour la recherche et de nouvelles possibilités technologiques et applicatives, en ouvrant sur la dimension interprétative de la voix humaine. Premièrement, l'étude de la voix actée — qui est par essence expressive — nécessite d'être en mesure de se confronter à des modalités de productions inusuelles et des registres extrêmes de variations acoustiques. Par ailleurs, elle pose la vaste question de la notion d'interprétation appliquée au domaine de la parole. La compréhension de cette notion nécessite d'étudier les pratiques des métiers de la voix : depuis la diversité de ses pratiques et de ses fonctions jusqu'aux conventions et aux choix de l'interprétation. On peut citer, à titre d'exemple, la production des voix dans les industries créatives qui suit un protocole extrêmement codifié, faisant intervenir un ensemble de conventions implicites depuis le choix d'un acteur capable d'incarner un physique ou un personnage, la supervision de son interprétation par un directeur artistique lors de séances d'enregistrements, jusqu'à la diffusion finale à un public cible. L'appréhension de la voix actée ne se limite manifestement pas à des facteurs acoustique ou linguistique et fait intervenir des facteurs sociologiques et culturels nécessaires pour sa compréhension et à sa modélisation. Cet article tente de présenter un état des lieux et une réflexion d'ensemble sur les défis et les applications de la voix actée, à la croisée de l'acoustique, de la linguistique, de la culture, et de l'apprentissage automatique. Il propose en outre un cadre méthodologique pour mieux définir le champs de la voix actée et en particulier les critères utilisés par un opérateur humain expert pour qualifier la voix et l'interprétation vocale d'un acteur. La compréhension des processus cognitifs mis en œuvre par un opérateur humain dans ses choix d'un acteur ou d'une interprétation devrait permettre, à terme, d'apprendre à la machine à reproduire ces choix.

L'article est organisé de la manière suivante : une première partie présente une enquête sociologique menée sur les pratiques des métiers de la voix, une seconde partie esquisse les défis et problématiques spécifiques au traitement de la voix actée, et une dernière partie présente une application dans le cadre du doublage au cinéma.

2 Enquête sociologique sur le rapport à la voix dans le travail d'interprétation

Dans l'optique de produire des connaissances sur les qualités de la voix actée, nous avons mis en place une enquête auprès de professionnels du doublage vocal. Artistes interprètes, directeurs artistiques et adaptateurs ont été interrogés dans le cadre de 7 entretiens semi-directifs conduits auprès de 9 enquêtés dans trois contextes en 2019 (Festival d'Avignon, festival international des voix du cinéma d'animation Voix d'Étoiles, échanges téléphoniques). L'analyse de ce matériau a permis de mettre en exergue un ensemble de caractéristiques propres à la voix dans le contexte du jeu d'acteur et du doublage, celui des fictions audiovisuelles (prise de vue réelle et animation) et des documentaires ; une diversité assurant une représentativité au sein de ce secteur professionnel.

Dérivé de l'anglais *dubbing*, le doublage est aujourd'hui un terme intégré dans le langage courant. Ce sont les Cahiers du cinéma qui, pour la première fois en mai 1930, ont présenté cette technique au public francophone : « Le *dubbing* ou doublage, est un système qui consiste à tourner des paroles en n'importe quelle langue et à l'adapter sur une bande tournée primitivement en parlant américain » (Cornu, 2014). Le doublage s'est institutionnalisé en un secteur à part entière l'industrie audiovisuelle. Il est particulièrement sollicité au regard de l'expansion du nombre de productions en circulation nécessitant une réponse technique et artistique de qualité pour leur diffusion internationale. Le doublage ne concerne pas tous les pays de la même manière, certains comme les pays scandinaves privilégient les versions originales sous-titrées, à l'inverse de la France qui est l'un des principaux *doubleurs* (en découle aussi une grande qualité des doublages français). D'autres pays en restent éloignés pour des raisons socio-culturelles : la Chine réserve ainsi le doublage aux films d'animation pour le jeune public. Un doublage dont le synchronisme labial n'est pas suffisant, voire non adapté au genre cinématographique ou audiovisuel, peut générer une suspension volontaire d'incrédulité pour les spectateurs (une notion entendue au sens de Samuel Taylor Coleridge dans Biographia (1817), relevant d'abord du champ littéraire, qui renvoie à l'acceptation, par le récepteur d'une œuvre, d'un univers fictionnel).

Le doublage tel que nous le connaissons aujourd'hui est un aboutissement d'une histoire culturelle et technologique. Avant « d'être parlant, le cinéma était muet, mais pas silencieux. Certes, les films muets n'offraient pas une parfaite *synchronisation*, mais ils n'étaient pas dénués de sons et d'illustrations musicales » (Roth, 2017). Le doublage se caractérise par un impératif de synchronisation, à commencer par un synchronisme labial (Bosseaux, 2015). La généralisation du doublage va entraîner la création de métiers qui lui sont spécifiques à l'instar des traducteurs-adaptateurs, des repéreurs et détecteurs, des coordinateurs linguistiques ou encore des monteurs en synchronisation. Elle va également générer une spécialisation de certains comédiens dont la carrière peut se concentrer presque exclusivement sur le doublage (post-synchronisation, narration, *voice over*, habillage d'antenne, publicité, bandes annonces, etc.). D'ailleurs, les équipes artistiques tentent de garantir une certaine continuité dans la relation au spectateur en gardant le même doubleur pour un comédien très reconnu (i.e. Patrick Poivey pour Bruce Willis ou Céline Monsarrat pour Julia Roberts). Si « L'arrivée du parlant va influencer la façon de produire le cinéma, de le voir et de l'entendre » (Roth, 2017 : 26), l'expansion du doublage va générer une spécialisation au sein de ce secteur industriel. Nous en distinguons quatre grands

champs : la postsynchronisation (domaine de la fiction), la narration (une voix *off* qui décrit le développement d'une trame, le plus souvent un documentaire), le *voice over* (traduire la personne qui parle à l'écran sans synchronisme labial), la publicité et les autres activités comme l'habillage d'antenne, les voix institutionnelles ou les bandes annonces.

L'enquête conduite auprès des comédiens a permis de mieux qualifier le travail sur la voix dans un contexte d'interprétation, notamment du doublage qui s'y consacre pleinement, soit le passage de la voix naturelle à la voix actée, et quelquefois chantée. Les artistes apprennent ainsi à *fixer leur voix*, à stabiliser leur *identité*, voire leur *signature de voix*. Certains acteurs admettent le fait de *changer leur voix* pour s'adapter au personnage ou à l'acteur en version originale. Pour d'autres, il s'agit plutôt de *prendre un accent*, de *moduler la voix*, de la *modifier* ou de l'*adapter*, mais non pas de la changer ou de la *transformer*. D'autres ont un langage plus artistique et parlent de *modeler la voix*, d'imaginer une sculpture, voire d'être *artisan de la voix*, de travailler à partir de leurs *tessitures* et de l'*amplitude de la palette vocale* qu'ils maîtrisent. Le travail sur la voix s'accompagne de deux autres démarches : une attention portée à la *respiration* et au *rythme du personnage* ou de l'acteur à doubler. Celui-ci peut relever de la *prosodie* de l'acteur ou de la langue (les acteurs doublant des productions audiovisuelles sont le plus souvent tributaires des traducteurs et des adaptateurs du scénario ou du texte de la version originale vers la version française). Plus encore, le travail sur la voix doit être conforme « au genre que l'on perçoit » qui se construit dans « l'immédiateté d'un objet concret avant de rejoindre les catégories de discours et vient fermer à sa façon une des fonctions inachevées telle qu'elle est directement reçue depuis le film projeté » (Ethis, 2004 : 166).

Les enquêtés ont également été interrogés sur une part de mimétisme vis-à-vis de la version originale. Il est intéressant de constater que la plupart des comédiens s'appuient sur la version à doubler, tout en s'attachant à conserver leur identité vocale dans la mesure où c'est à partir de celle-ci qu'ils ont été choisis. La voix actée se caractérise par des qualificatifs qui renvoient en particulier à l'idée d'*intention* : elle est le reflet d'une *énergie* qui appartient à l'artiste en version originale ou à la situation présentée à l'écran. La voix actée est ainsi une adresse qui ne doit ni trahir une œuvre, ni le jeu en version originale. Elle se situe dans un système de contraintes communicationnelles et symboliques (genre audiovisuel, registre de jeu, registre de langue). Les artistes devant être en capacité d'incarner une multitude de personnages durant leur carrière, et donc s'inscrire dans divers registres de jeu, au-delà de la voix elle-même, c'est leur palette, autrement dit leur amplitude vocale, qui importe. Enfin, la question de la ressemblance physique entre un artiste en version originale et en version française a été posée : si elle est parfois un repère, c'est souvent la personnalité de l'interprète qui va importer dans un casting.

Pour conclure, il convient d'insister sur le positionnement du doublage dans l'industrie audiovisuelle contemporaine : il est à la fois un espace technique, voire technologique, économique, artistique et culturel. Le travail sur la voix actée doit s'inscrire dans des registres filmiques ou audiovisuels pour être au service d'un projet artistique et des publics auxquels il s'adresse.

3 Problématiques et verrous scientifiques pour le traitement de la voix actée

La suite de cet article s'intéresse à la voix au cinéma et plus particulièrement les voix pour le doublage, dont le *voice over*. Plusieurs protagonistes interviennent dans la chaîne de traitement du *voice over*. Tout d'abord, les voix des doubleurs doivent à la fois rester fidèles à l'oeuvre

d'origine, mais aussi correspondre aux attentes du public cible. Le public est l'auditeur final du travail et doit être immergé dans l'oeuvre à son visionnage. Ensuite, la qualité d'un doublage ressort de la responsabilité des Directeurs Artistiques (DA). Ce métier nécessite une expérience en comédie et une culture cinématographique développée afin de sélectionner des voix, via le casting vocal, et de guider les acteurs lors d'enregistrements. Pour finir, le client a des objectifs commerciaux ou artistiques et contrôle les prises de décision. Il intervient lors de la sélection des acteurs et valide la qualité du doublage. Cette chaîne implique de nombreux choix humains, ainsi que des conventions culturelles et esthétiques qui sont implicites ou explicites. Identifier les codes explicitement ou modéliser ces choix implicitement imposent des verrous scientifiques que nos travaux cherchent à briser. Pour cela, nous avons concentré nos efforts sur trois points principaux : la taxonomie de la voix, la modélisation d'un processus artistique et la dépendance de réception de la voix à la langue et la culture.

Dans le but de comprendre et modéliser les mécanismes du doublage, il est nécessaire de savoir représenter une voix actée. Cette représentation peut être construite à partir de codes explicites donnant définition au terme palette vocale. Il n'existe cependant aucune taxonomie partagée par les DA et autres protagonistes de la chaîne de production du doublage. Ces faits rendent difficile la création d'un protocole d'annotation. Quand bien même une taxonomie existerait, les DA ont hélas très peu de temps à accorder à l'annotation. Sans ces annotations, il est difficile à la fois de construire des modèles d'apprentissage supervisé pour détecter automatiquement des facteurs de variabilité dans la voix, d'isoler et d'analyser les impacts de certaines caractéristiques dans des prises de décisions. Evaluer des systèmes automatiques de classification de la voix actée, et expliquer de manière intelligible à l'utilisateur les critères qui ont guidé les décisions de nos systèmes sont aussi des limites auxquelles nous nous confrontons. Autant dans une problématique sociologique qu'informatique, se pose alors la question suivante : comment définir une taxonomie de la voix actée sans l'accès à une connaissance de ses codes explicites ?

Le jugement de l'adéquation entre une voix et son personnage conduit par les DA implique de nombreux facteurs humains et perceptifs. Pour modéliser un tel processus, il nous faut comprendre quels critères ou filtres y interviennent. Nous supposons que cette adéquation porte autant sur la nature de la voix d'un comédien que sur l'interprétation qu'il y apporte en jouant le personnage cible. Nous étudions dans nos travaux comment inférer ces critères et filtres depuis des enregistrements de voix et des décisions de DA. Les critères peuvent être différents en fonction du DA. (chacun présente un profil différent, avec une expérience et des préférences artistiques qui lui sont propres) A l'instar de la reconnaissance de la parole où certains modèles sont adaptés au locuteur, il nous est nécessaire d'étudier l'adaptation de nos modèles au DA cible. Différentes pistes sont à explorer. La construction d'un modèle par DA. La construction d'un modèle universel. L'adaptation d'un modèle universel à chaque DA. Ces modèles doivent aussi s'adapter dans le temps. Le DA, à force d'expérience, évolue et ses décisions aussi.

La voix actée est intrinsèquement liée à la perception humaine. Partant de ce postulat, nous recherchons quels codes sont universels et lesquels varient en fonction des individus. Ces codes sont sujets à un processus cognitif qui rend leur modélisation difficile. Les facteurs de variation de ces derniers peuvent être d'ordre culturel ou liés à la structure de la langue parlée. Cette variabilité peut intervenir dans différentes dimensions. Comme, par exemple, la dimension de l'émotion. Certaines émotions sont universelles et ne varient pas d'une langue à l'autre, là où d'autres sont dépendantes de l'individu. Ohala (Ohala, 1996) unifie des théories guidant les recherches sur l'expression d'émotion par la voix et tend à montrer que certaines émotions sont partagées par différentes cultures. De la même manière, nous supposons que le jeu des personnages est construit pour mettre en avant des stéréotypes qui diffèrent d'une culture à l'autre.

Pour finir, tout comme notre société évolue, ces codes évoluent, et nous devons chercher à prendre en compte dans nos modèles l'évolution temporelle de la réception de la voix.

4 Application : l'aide au doublage de voix d'acteurs

Un cadre applicatif possible pour la voix actée pourrait être un système de recommandation de voix pour les DA. Un tel système consisterait à proposer automatiquement un ensemble de voix selon une requête formulée. Dans notre contexte d'étude, nous proposons de nous intéresser au casting vocal et plus particulièrement au doublage : considérant une voix dans une langue source, quelles voix d'acteurs dans une langue cible peuvent le mieux correspondre aux attentes d'un DA ? En réponse à ces problématiques, des travaux ont proposé d'étudier la similarité de voix, qui est une notion centrale dans l'élaboration de systèmes de recommandation. Elle consiste à définir une mesure de similarité entre deux voix. Comment considère-t-on que deux voix sont similaires ? A partir de quelles données, vérité, peut-on construire une telle mesure ? L'apparition de la similarité de la voix au sens du personnage a donné des réponses à ces questions. Celle-ci part de l'hypothèse suivante : il existe, dans le signal acoustique de parole, des signes acoustiques caractéristiques du personnage joué. La similarité de la voix au sens du personnage consiste alors à définir une mesure qui, si les deux acteurs jouent le même personnage, doit informer que les voix sont similaires. Dans le cas contraire, la mesure doit indiquer que les deux voix paraissent dissimilaires. Deux approches différentes ont été proposées pour cette mesure de similarité.

Dans la première approche (Gresse, 2017), deux modèles *i*-vecteurs (Dehak, 2010) sont appris, l'espace source (anglais), l'espace cible (français). Le modèle *i*-vecteur est un modèle de représentation de voix plus communément utilisé pour des tâches de reconnaissance du locuteur. Le système proposé consiste à appliquer une matrice de passage *W* à l'espace cible pour le projeter dans un nouvel espace, que nous nommerons l'espace modifié, comparable avec l'espace source. La matrice *W* est apprise sur un ensemble de données d'entraînement. Un score de similarité est ensuite calculé entre les vecteurs de l'espace source et de l'espace modifié dans le cas du système B et C. Pour le système *baseline* A, la mesure est calculée entre l'espace source et l'espace cible.

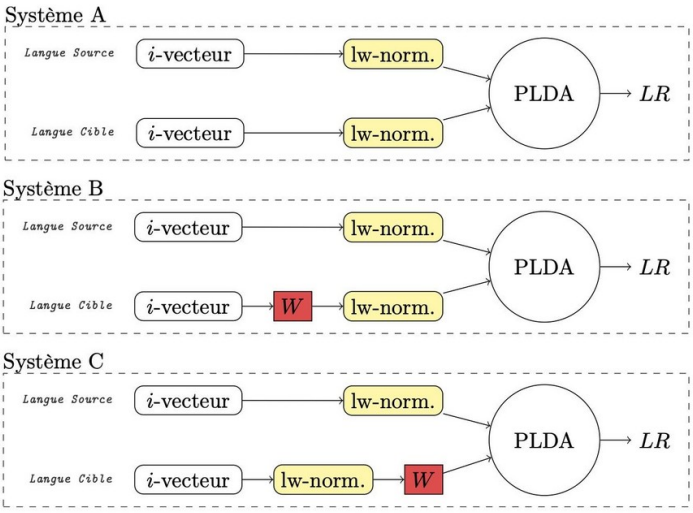


Figure 1: Systèmes de similarité proposés dans (Gresse, 2017)

Dans la seconde approche (Gresse, 2018), un réseau de neurones siamois est entraîné à projeter les deux voix fournies en entrée, une en anglais et une en français, dans un espace commun où celles-ci sont comparables au moyen d'un score de similarité. L'objectif de cet espace est de représenter la dimension personnage des voix qui y sont projetées indépendamment du locuteur, de la langue et du contenu linguistique. Dans la continuité de ces travaux, nous nous sommes intéressés à la création d'un espace personnage en s'inspirant des modèles x-vecteurs (Snyder, 2018). Nous entraînons un réseau de neurones profond à reconnaître le personnage joué quelque soit sa langue. Le réseau fournit en sortie des probabilités indiquant l'appartenance de l'enregistrement à un personnage spécifique. Une fois le réseau entraîné, sa dernière couche sert d'espace de représentation, le p-vecteur. La figure 2 donne un exemple de cet espace de représentation orienté "personnage". Des enregistrements de 4 paires de voix, chacune composée d'une voix anglaise et d'une voix française jouant le même personnage, sont projetée dans l'espace p-vecteur dans lequel nous avons appliqué un algorithme de regroupement pour tenter de retrouver les classes personnage. Chaque paire est représentée avec une couleur donnée. Notons que le système n'a jamais vu durant la phase d'apprentissage, ni les personnages, ni les locuteurs concernés.

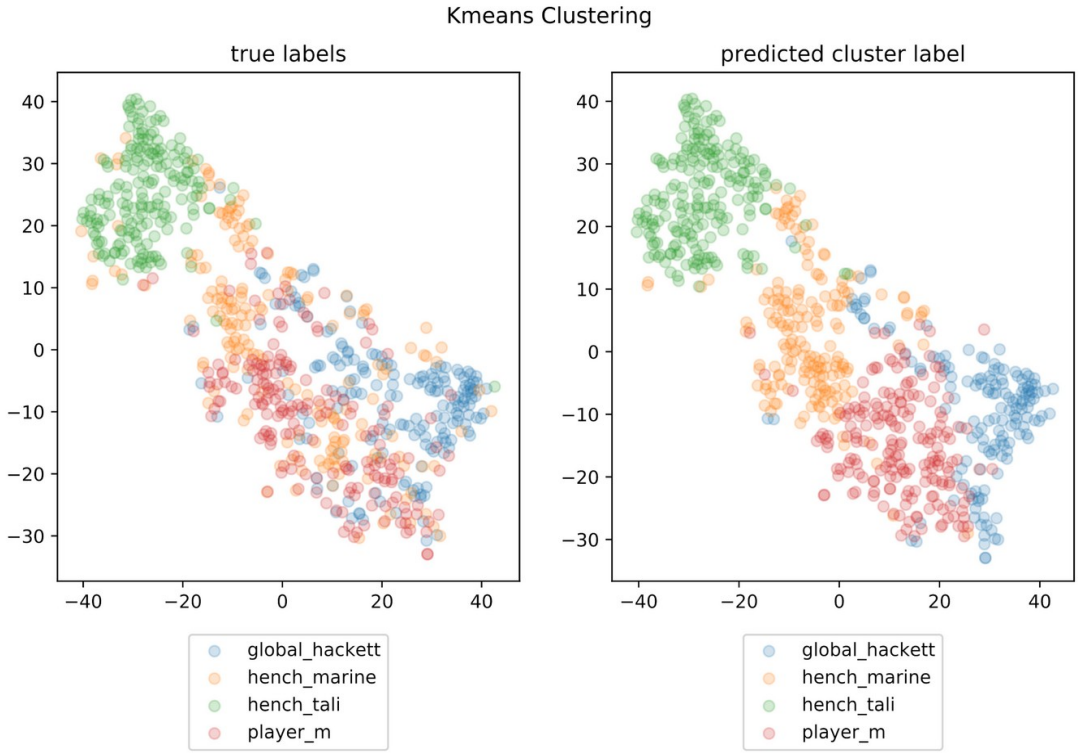


Figure 2: Comparaison des vrais labels de personnage avec les labels associés à chaque enregistrement par un algorithme de clustering sur un espace de représentation de la dimension personnage de la voix actée.

D'autres travaux ont proposé une mesure de similarité perceptuelle (Obin, 2014). Cette contribution se démarque des précédentes en proposant d'abord de représenter un enregistrement par des classes au lieu de calculer un score directement sur un espace acoustique. Ces classes sont liées à la physiologie, la phonétique, au timbre, à l'articulation, à la prosodie ou au jeu d'acteur. Un système de classification multi-labels est alors entraîné sur des enregistrements étiquetés par un expert. Pour un enregistrement donné, le système calcule un vecteur de probabilités où chaque valeur représente l'affinité de l'enregistrement pour le label. Chaque enregistrement est donc représenté par un vecteur d'affinité, référé comme la signature vocale de l'enregistrement. Appliquer une distance entre deux signatures revient à calculer un score de similarité entre deux

enregistrements. L'article confronte deux signatures vocales, l'une issue de vecteurs d'affinités appris depuis les *i*-vecteurs, l'autre des *i*-vecteurs + *PLDA*, et montre au travers d'une expérience perceptive que l'apprentissage du classifieur a permis de mettre significativement en lumière, dans l'espace i-vecteur, des informations caractéristiques de la perception de la voix actée.

D'un point de vue applicatif, ces travaux sur la similarité et la représentation de voix ouvrent des perspectives pour construire des systèmes de recommandation pour le casting vocal. Ces espaces de représentations vont permettre d'initialiser les systèmes de recommandation pour proposer des voix qui seront pertinentes ou inattendues pour les DA. Un tel système pourra prendre la forme d'une application informatique où le DA pourra demander des recommandations et réagir aux résultats. Dans une dynamique itérative, cette application fournira de nouvelles données à la recherche qui alimenteront nos travaux pour améliorer encore la finesse des recommandations proposées aux DA.

5 Conclusion

Nous avons proposé dans cet article un état des lieux et une réflexion sur les défis scientifiques et les applications technologiques de la voix actée. L'appréhension de la voix actée soulève un ensemble de questions inédites de recherche pour la communauté de la parole. Tout d'abord, une analyse des pratiques montre qu'il existe une grande diversité de métiers de la voix, chacun avec ses fonctionnalités et ses objectifs propres. Les pratiques de ces métiers ne sont cependant pas homogènes et relèvent d'une dimension artistique. En particulier, l'étude des voix de doublage montre l'existence d'un cadre partagé pour la sélection et la direction de voix d'acteurs dans le but de produire des effets définis, mais les codes sous-jacents restent implicites et difficiles à formaliser. Ainsi, l'appréhension de la voix actée ne se limite manifestement pas à des facteurs purement acoustiques ou linguistiques, mais fait intervenir des facteurs sociologiques et culturels reflétés par les choix d'opérateurs humains. La modélisation des choix des opérateurs humains nécessite donc de modéliser explicitement ou implicitement les filtres cognitifs ayant amené à ces choix. La mise en évidence de codes explicites par une taxonomie partagée et standardisée et/ou la modélisation implicite des choix représente clairement un défi pour la recherche dans ce domaine. Une étude est en cours pour essayer de faire apparaître une taxonomie partagée par les professionnels pour qualifier une voix dans les métiers en production ou post-production de voix. Par ailleurs, nous avons présenté dans cet article une solution pour modéliser implicitement les choix d'un opérateur autour de la notion de "personnage", avec les résultats préliminaires obtenus en doublage de voix. Les études à venir visent à préciser les méthodologies présentées pour permettre de mieux comprendre et modéliser les nombreux facteurs de la voix actée et, à terme, les exploiter pour créer des applications technologiques pour les industries créatives.

Remerciements

Cette recherche a été menée avec le projet TheVoice (ANR-17-CE23-0025) financé par l'Agence Nationale de la Recherche.

Références

Bosseaux C. (2015). *Dubbing, Film and Performance. Uncanny Encounters.* Bern : Peter Lage.

Cornu, J.-F. (2014). *Le doublage et le sous-titrage: Histoire et esthétique*. Presses universitaires de Rennes.

Ethis E., Malinas D. (2012). *Films de Campus. L'Université au cinéma*. Paris : Armand Colin.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

Ethis E. (2004). *Pour une po(ï)étique du questionnaire en sociologie de la culture. Le spectateur imaginé*. Paris : L'Harmattan.

Gresse A., Rouvier M., Dufour R., Labatut V., Bonastre J.-F. (2017). *Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization*. INTERSPEECH.

Gresse A., Quillot M., Dufour R., Labatut V., Bonastre J.-F. (2018). Similarity Metrics Based on Siamese Neural Networks for Voice Casting. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouelle (2010). "Front-End Factor Analysis For Speaker Verification," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788 – 798

Launier J.-J. (2016). *L'Art des studios d'animation Walt Disney. Le mouvement par nature*. Art Ludique - Le Musée.

Le Breton, D. (2011). *Eclats de voix.. Une anthropologie des voix*. Métailié.

Obin N., Roebel A., (2016). Similarity search of acted voices for automatic voice casting, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24 , Issue. 9, p. 1642 – 1651.

Obin N., Roebel A., Bachman G. (2014). On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification

Ohala J., (1996). *Ethological Theory and the Expression of Emotion in the Voice*. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96

Roth R. (2017). À l'écoute de Disney. *Une sociologie de la réception de la musique au cinéma*. Paris : L'Harmattan.

Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. (2018). *X-Vectors: Robust DNN Embeddings for Speaker Recognition*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude des facteurs affectant la compréhensibilité de documents multimodaux : une étude expérimentale

Estelle Randria[1,2], Lionel Fontan[2], Maxime Le Coz[2], Isabelle Ferrané[1], Julien Pinquier[1]

(1) IRIT-UPS, 118 route de Narbonne 31062 Toulouse, France
(2) Archean LABS, 20 place Prax-Paris, 82000 Montauban, France
{estelle.randria, isabelle.ferrane, julien.pinquier}@irit.fr, {lfontan, mlecoz}@archean.tech

RÉSUMÉ

La compréhensibilité de documents audiovisuels peut dépendre de facteurs propres à l'auditeur/spectateur (ex. langue maternelle, performances cognitives) et de facteurs propres aux contenus des documents (ex. complexité linguistique, intelligibilité de la parole). Dans ces travaux, nous étudions les effets de facteurs propres aux contenus sur la compréhensibilité de 55 dialogues extraits de films, présentés à 15 experts (enseignants de français langue étrangère) selon cinq modalités différentes (transcription, transcription + audio, audio, audio + vidéo, transcription + audio + vidéo). Les experts ont évalué les dialogues en termes de compréhensibilité générale, de complexité du vocabulaire, de complexité grammaticale, et d'intelligibilité de la parole. L'analyse de leurs évaluations montre que (1) la complexité du vocabulaire, la complexité grammaticale, et l'intelligibilité de la parole sont significativement corrélées à la compréhensibilité générale, et (2) que les évaluations de compréhensibilité générale ont tendance à être plus élevées lors de présentations multimodales.

ABSTRACT

Factors affecting the comprehensibility of multimodal documents : an experimental study

It has been shown that the comprehensibility of audiovisual documents can be influenced both by factors related to the recipient (e.g., native language, cognitive performance) and factors related to the document itself (e.g., linguistic complexity, speech intelligibility). The aim of this study was to investigate the effects of the document-related factors on the comprehensibility of 55 movie clips, presented to 15 experts (language teachers) under five different (combinations of) presentation modalities (text transcript, audio, text transcript + audio, audio + video, text transcript + audio + video). The experts rated the documents in terms of overall comprehensibility, lexical complexity, grammatical complexity, and speech intelligibility. The analysis of the experts ratings indicated that (1) lexical complexity, grammatical complexity, and speech intelligibility are all three significantly correlated to the ratings of comprehensibility and (2) that the combination of different presentation modalities lead to higher ratings of comprehensibility.

MOTS-CLÉS : corpus, film, compréhensibilité, compréhension orale, multimodalité, complexité linguistique, complexité grammaticale, intelligibilité de la parole.

KEYWORDS: corpus, movie, comprehensibility, listening comprehension, multimodality, linguistic complexity, grammatical complexity, speech intelligibility.

Volume 1 : Journées d'Études sur la Parole, pages 534–542.
HAL : hal-02798583.

1 Introduction

Le développement et la popularisation des plateformes de vidéo à la demande a eu un impact majeur sur la diffusion de contenus audiovisuels. La compréhension de contenu est un élément clef pour favoriser l'accessibilité à ces collections massives de documents audiovisuels. Il est possible de classer du contenu audiovisuel par genre, et il existe également des labels permettant de définir l'âge minimum conseillé pour regarder une vidéo, mais aujourd'hui il n'existe pas de classification permettant d'indiquer à quel point un document audiovisuel est compréhensible pour un public donné. En effet, ce type de document peut être perçu comme plus ou moins compréhensible selon la personne qui y a accès (personnes âgées, enfants, étrangers...). Dans ce contexte, il est intéressant d'étudier si ce type de classification est envisageable. Cet article traite de la compréhensibilité de contenu audiovisuel, plus spécifiquement de la difficulté potentielle de compréhension qui peut leur être associée, qui sera appelée par la suite « difficulté globale ». Dans un premier temps, la difficulté globale sera abordée principalement dans le cadre de l'apprentissage des langues. Les objectifs de notre étude sont : (1) d'étudier de quelle façon les humains perçoivent la difficulté globale de contenus audiovisuels (2) de savoir ce qui influence cette difficulté entre la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité de la parole (c'est-à-dire à quel point la parole est facile à percevoir) et puis (3) de savoir si l'accès aux différentes modalités (vidéo, audio et texte) prises séparément ou combinées, joue un rôle sur la perception humaine de la difficulté globale.

2 Etudes autour de la compréhensibilité de documents

La compréhensibilité de documents, que ce soit de documents écrits, audio ou audiovisuels, a fait l'objet de nombreuses études. La grande majorité a cependant porté sur la compréhensibilité de textes et de documents audio plutôt que sur les contenus audiovisuels. Dans cette partie, nous présentons les éléments les plus souvent dégagés par les études comme étant des facteurs de complexification ou de facilitation de la compréhension orale.

Tout d'abord, il est important d'aborder des aspects linguistiques : il a en effet été démontré très tôt que le vocabulaire et la grammaire (notamment la syntaxe), jouaient un rôle prépondérant dans la compréhension écrite et orale. Les premières preuves ont été apportées dans des études portant sur la lisibilité (*readability*), qui prennent en compte les aspects linguistiques et conceptuels d'un texte pour évaluer sa difficulté globale. De nombreuses formules de lisibilité, prédisant la difficulté des textes, font appel à des descripteurs liés à la difficulté du vocabulaire et à la complexité grammaticale. Les formules les plus connues sont celles de Lively et Pressey (Lively & Pressey, 1923), Flesch (Flesch, 1948) et la formule (revisitée) de Dale et Chall (Chall & Dale, 1995) pour l'anglais, et la formule de Henry (Henry, 1975) pour le français. Ces formules utilisent notamment des variables liées à la fréquence d'occurrence des mots, à la longueur des phrases et des mots ou à la diversité lexicale du texte. Les nouvelles techniques d'apprentissage informatique (*machine learning*) ont permis d'aller beaucoup plus loin dans cette voie en analysant un grand nombre de descripteurs dérivés du texte afin d'obtenir de nouvelles formules. Cependant les nouvelles formules proposées par les linguistes construisant continuent toujours à faire appel à des variables liées aux complexités grammaticales et de vocabulaire (François, 2009). Tout comme pour la compréhension écrite, les connaissances linguistiques jouent un rôle prépondérant dans la compréhension orale, que ce soit pour la compréhension orale de la langue native ou étrangère (Buck, 2001) : des connaissances en vocabulaire et grammaire étant nécessaires pour décoder le contenu. Le lien de ces complexités

avec la difficulté de compréhension orale a été montré à plusieurs reprises (Nissan *et al.*, 1995; Carrow-Woolfolk, 1999).

Mais si la complexité intrinsèque du message est importante, sa vocalisation comporte également des aspects important pour la compréhension orale. Pour l'anglais l'emphase et l'intonation peuvent être déterminantes pour la compréhension du message (Wong & Waring, 2010) en aidant à insister sur les mots importants. Le débit de parole ou l'accent d'un locuteur peuvent également affecter la compréhension (Boyle, 1984). Un débit rapide peut également être une complication pour la compréhension des non natifs (Goh, 1994). Les pauses et les hésitations (aussi appelées disfluences) ont des effets qui varient selon les auditeurs, elles peuvent être informatives pour les auditeurs natifs (Corley & Hartsuiker, 2003) ou porter préjudice à la compréhension des non natifs (Voss, 1979). Chang et Read ont étudié l'impact des accents sur la compréhension orale (Chang & Read, 2008) et ont conclu que la compréhension des auditeurs était affectée par la présence d'accents qui ne leur sont pas familiers.

L'environnement sonore a lui aussi un impact sur la difficulté perçue.Dans l'étude de Boyle (Boyle, 1984), plusieurs professeurs le citent comme un élément important dans la compréhension orale en L1 et L2. Les auditeurs natifs peuvent rencontrer plus de difficultés à comprendre un document audio si les conditions sonores sont mauvaises (Adank *et al.*, 2009). Les effets de l'environnement sonore sont encore plus importants pour des auditeurs non natifs qui sont confrontés à la fois aux conditions d'écoute dégradées et à leurs potentielles lacunes dans la langue cible.

Les études concernant la compréhension de contenus audiovisuels sont moins nombreuses, mais, si l'influence des aspects linguistiques sur la compréhension de contenus audiovisuels n'a visiblement pas été étudiée, on trouve des études s'intéressant à l'influence des sous-titres sur la compréhension de documents audiovisuels. Ce sujet a tout son intérêt dans le cadre de l'apprentissage des langues. Plusieurs tests menés pour déterminer si les sous-titres sont bénéfiques ou nuisibles pour la compréhension, ont principalement conclu que les sous-titres avaient un effet positif en accroissant la compréhension. (Perez *et al.*, 2013; Markham *et al.*, 2001).

Dahl a mené une étude qui a souligné l'importance des gestes dans la facilitation de la compréhension orale L1 et L2 (Dahl & Ludvigsen, 2014) : même si les locuteurs natifs et non natifs ne les exploiteront pas de la même façon, les gestes les aideront à atteindre une meilleure compréhension. En partant de ces observations, l'hypothèse peut être faite que les gestes peuvent faciliter la compréhension des contenus audiovisuels. L'accès à des vidéos qui contiennent des gestes et des indices visuels permet effectivement aux apprenants en langue seconde à mieux réussir la tâche de compréhension orale (Sueyoshi & Hardison, 2005).

Les recherches sur la compréhension de contenu ont mis en avant l'influence des aspects linguistiques sur la compréhension, que ce soit pour la compréhension écrite et orale. Elles ont aussi montré que la qualité de production de la parole (la prosodie, le débit....) et l'environnement sonore ont une influence sur la compréhension de contenu audio. Pour les contenus audiovisuels, la présence d'indices visuels et de sous-titres apparaissent comme des facilitateurs pour la compréhension. Notre étude s'intéresse aux aspects inhérents aux documents audiovisuels qui affectent la compréhension orale : la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité de la parole.

3 Matériel et méthodes

3.1 Corpus

L'étude porte sur la perception humaine de la difficulté globale d'extraits issus de films. Pour mener cette étude, la création d'un corpus approprié était nécessaire, les sujets de l'expérience devant évaluer la difficulté globale des extraits en fonction des modalités disponibles. Trois modalités ont été exploitées : la modalité audio, la modalité vidéo et la modalité texte. Un corpus composé de 55 extrait de 15 films populaires français a été constitué. Les films ont été choisis pour que le corpus contienne des films de genres et d'époques différents, avec de la variété dans la langue et dans le registre. Les extraits ont été sélectionnés de telle sorte à ce qu'ils répondent à la définition d'une interaction (Traverso, 2013). La compréhension de situations de communication et de leurs éléments contextuels (qui ? quand ? où ? comment ?) étant la première étape menant à la compréhension de contenu audiovisuel. Les scènes d'action, ne comprenant pas de parole ou comportant des interactions trop courtes ont été exclues car elles ont été jugées comme n'étant pas suffisamment pertinentes pour évaluer la difficulté globale. Trois à cinq extraits par films ont été choisis, puis les flux audio et vidéo, ainsi que les transcriptions de ces extraits ont été récupérés. Le corpus compte 55 extraits (2541 secondes), chacun disponible sous trois formats différents :

— texte : il s'agit de la transcription des extraits (7225 mots),
— audio : les pistes sonores seules,
— vidéo : qui inclut les images et le son.

3.2 Participants

Les experts de l'étude ont été choisis parmi des professeurs de langue car ils sont souvent confrontés à l'évaluation de documents (de type texte, audio ou vidéo) pour déterminer s'ils sont appropriés pour leurs étudiants en terme de difficulté globale. Quinze professeurs de français langue étrangère (FLE) ont été recrutés, il était requis qu'ils soient francophones natifs avec au moins trois ans d'expérience dans l'enseignement (avec des apprenants de niveaux variés). Ils devaient être familiers avec l'utilisation de documents audiovisuels en classe et être normo-entendants. Les 15 professeurs –13 femmes, 2 hommes, tranche d'âge : 27-63 ans, âge moyen : 37 ans –ont une expérience dans l'enseignement qui varie entre 3 et 40 ans (moyenne : 11 ans, écart-type : 9 ans). Tous les experts ont reçu une compensation financière pour leur participation.

3.3 Évaluation

Une interface graphique a été développée puis déployée en ligne, afin de présenter les 55 extraits de films aux différents participants. Les extraits pouvaient être présentés sous cinq conditions différentes, chaque condition correspondant à une combinaison de modalités : texte (T) (seule la transcription de l'extrait était accessible), audio (A), audio+texte (AT), audio+vidéo (AV) et audio+vidéo+texte (AVT). Pour chaque participant, chacun des extraits a été présenté une seule fois sous une des conditions citées ci-dessus. Les participants ont réalisé l'évaluation en ligne, en utilisant leur propre matériel. Il leur a été spécifié de réaliser l'expérience dans un lieu calme, en utilisant un ordinateur et des écouteurs. À la fin de l'évaluation, chacun des 55 extraits de films a été présenté sous chacune des conditions à exactement trois participants.

Les extraits vidéos ont été présentés dans un ordre aléatoire. Pour chacun des extraits, l'interface présentait une zone dédiée à la transcription, un lecteur audio/vidéo ainsi que trois à quatre curseurs, en fonction de la condition présentée. Les participants avaient comme instruction d'utiliser chacun des curseurs mis à leur disposition pour évaluer chacun des extraits en terme de difficulté globale (de 0 –très facile à 100 –très difficile), de complexité du vocabulaire (de 0 –très facile à 100 –très difficile), de complexité grammaticale (de 0 –très facile à 100 –très difficile) et (pour chaque extrait présenté avec la modalité audio) d'intelligibilité de la parole (de 0 –totalement intelligible à 100 –totalement inintelligible). La position initiale de chaque curseur était à 50. L'interface mettait également à disposition des zones de texte où les participants pouvaient laisser un commentaire pour justifier la position donnée à chaque curseur. Les commentaires étaient obligatoires pour justifier l'évaluation de la difficulté globale. Les participants ont été familiarisés avec la tâche d'évaluation lors d'une phase d'entraînement sur cinq extraits de films, qui précédait la véritable évaluation, ils étaient libres de réaliser l'évaluation des extraits en plusieurs fois.

4 Résultats

Cette partie traite, dans un premier temps, de l'influence de la complexité du vocabulaire, de la complexité grammaticale et de l'intelligibilité sur la difficulté globale. La dernière partie étudie l'influence des modalités sur la difficulté globale perçue et l'intelligibilité de la parole. La diversité dans la perception des experts pouvant mettre en évidence des facteurs intéressants pour la suite de l'étude, nous n'avons pas procédé à des accords inter-annotateurs.

4.1 Relation entre les scores de difficulté globale et les scores de complexité du vocabulaire, de la grammaire et de l'intelligibilité

Dans cette partie les relations entre les scores de difficulté globale et les « sous-scores » (*i.e.* les scores de complexité du vocabulaire, de complexité grammaticale et d'intelligibilité) sont étudiées, à l'aide de corrélations bivariées et de régressions linéaires multiples.

4.1.1 Corrélations bivariées

La table 1 présente les résultats des corrélations par rang de Spearman, calculées entre les scores de difficulté globale et les sous-scores. Dans le cas de l'intelligibilité, les extraits présentés aux participants qui ne contenaient pas la modalité audio ont été ignorés.

TABLE 1 – Corrélations par rang de Spearman entre la difficulté globale et la complexité du vocabulaire, la complexité grammaticale, l'intelligibilité de la parole (*** $p \leq 0,001$)

Complexité	Vocabulaire	Grammaire	Intelligibilité parole
Difficulté globale	0,74***	0,56***	0,63***

Des corrélations significatives positives, modérées à fortes, ont été trouvées entre la difficulté globale et les trois sous-dimensions évaluées. Ces résultats montrent que (1) comme cela pouvait être attendu,

plus le vocabulaire et la grammaire sont perçus comme étant complexes plus la difficulté globale perçue augmente et (2) plus la parole est intelligible, plus la difficulté globale perçue diminue. Pour étudier plus en profondeur le lien entre la difficulté globale et les trois sous-dimensions, des régressions linéaires multiples ont été réalisées.

4.1.2 Régressions linéaires multiples

Les scores de difficulté globale et les scores de complexité du vocabulaire, de complexité grammaticale et d'intelligibilité ont été moyennées pour chaque extrait, en prenant en compte les scores donnés sous toutes les conditions. Deux régressions linéaires multiples ont été réalisées en utilisant ces moyennes. La première régression a été calculée en prenant la difficulté globale comme variable dépendante et la complexité du vocabulaire et la complexité grammaticale comme variables indépendantes.

La régression linéaire multiple donne un coefficient de détermination élevé, avec un R^2 ajusté de 0,76. Ceci signifie que le modèle permet d'expliquer 76% de la variance de la difficulté globale. Les coefficients non-standardisés (NsCoef) montrent que la complexité du vocabulaire (NsCoef = 0,69) a plus de poids dans la régression que la complexité grammaticale (NsCoef = 0,34). La figure 1 représente un diagramme le dispersion reliant les scores de difficulté globale moyens prédits aux scores humains de difficulté globales moyens.

Une nouvelle régression linéaire multiple est réalisée pour voir si prendre en considération les scores évaluant l'intelligibilité de la parole permet d'améliorer la prédiction de la difficulté globale. Cette régression prend la difficulté globale comme variable dépendante, et la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité comme variables indépendantes. Pour cette régression, les extraits ayant été notés avec la modalité texte seule ont été ignorés. Rajouter l'intelligibilité comme variable dépendante dans cette seconde régression linéaire permet d'obtenir un meilleur coefficient de détermination, avec un R^2 ajusté de 0,82. La complexité du vocabulaire a toujours plus de poids dans la prédiction de la difficulté globale (NsCoef = 0,55), suivie de la complexité grammaticale (NsCoef = 0,31) et de l'intelligibilité de la parole (NsCoef = 0,28). La relation entre les valeurs prédites et les valeurs réelles sont visibles dans la figure 2.

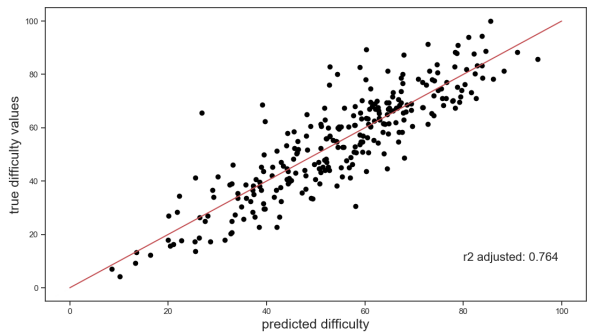


FIGURE 1 – Diagramme de dispersion reliant les évaluations humaines de la difficulté globale aux prédictions de la difficulté globale, en utilisant une régression linéaire multiple avec la complexité du vocabulaire et la complexité grammaticale comme variables indépendantes

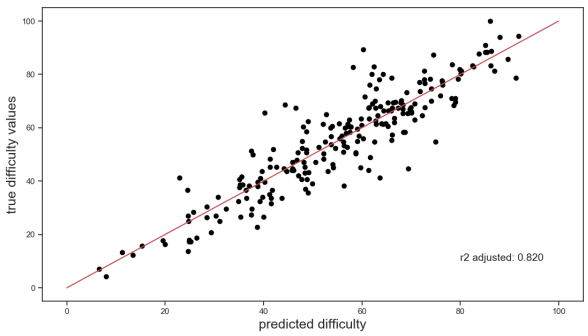


FIGURE 2 – Diagramme de dispersion reliant les évaluations humaines de difficulté globale aux scores prédits de difficulté globale, en utilisant une régression linéaire multiple avec la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité comme variables indépendantes

4.2 Influence des modalités

Un autre aspect intéressant à étudier est la variation des scores en fonction des modalités disponibles. Les hypothèses suivantes ont été faites : (1) la difficulté globale perçue sera plus élevée pour la condition A que pour la condition AT et que pour la condition AV, (2) la difficulté globale perçue sera plus élevée pour les conditions AV et AT que pour la condition AVT, (3) l'intelligibilité de la parole augmente si la modalité vidéo et/ou la modalité texte sont combinées avec la modalité audio. Aucune relation n'est attendue entre les modalités et la complexité du vocabulaire et la complexité grammaticale.

La moyenne et l'écart-type des scores de difficulté globale et d'intelligibilité ont été calculés pour chaque condition, pour visualiser l'évolution des scores en fonction des modalités disponibles. Les résultats sont représentés dans les figures 3 et 4.

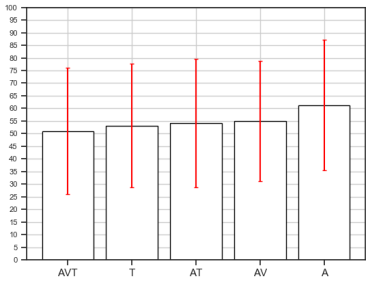


FIGURE 3 – Scores moyens de difficulté globale en fonction des combinaisons de modalités. Les barres d'erreur représentent ± 1 écart-type

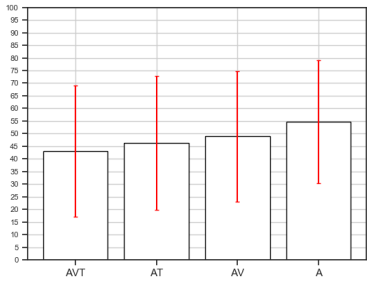


FIGURE 4 – Scores moyens d'intelligibilité de la parole (scores bas = intelligibilité haute) en fonction des combinaisons de modalité. Les barres d'erreur représentent ± 1 écart-type.

4.2.1 Influence des modalités sur la difficulté globale

En premier lieu, il est possible d'observer sur la figure 3 que la moyenne des scores de difficulté globale est la plus élevée dans la condition A, c'est-à-dire si la modalité audio est seule ; dans ce corpus, les extraits présentés avec la modalité audio seule étaient donc les plus compliqués. Comme supposé, les moyennes sont moins élevées quand les modalités vidéo et texte s'ajoutent à la modalité audio seule, la moyenne la plus basse est obtenue pour la condition AVT, c'est-à-dire quand les trois modalités audio, vidéo et texte sont disponibles. Pour ce corpus, avoir accès aux trois modalités en même temps permet de minimiser la difficulté globale perçue.

4.2.2 Influence des modalités sur l'intelligibilité de la parole

Avoir accès à la vidéo et au texte devrait représenter un avantage pour la compréhension : si l'accès à la vidéo ne permet pas totalement de désambiguïser ce qui est entendu, la présence du texte permet de définitivement éliminer les problèmes d'inintelligibilité (même s'il est possible qu'un surplus d'information amène une surcharge cognitive). Cette théorie peut être supportée par les observations faites sur l'évolution de la moyenne des scores d'intelligibilité en fonction des modalités disponibles. La figure 4 montre que les extraits vidéos présentés avec la modalité audio seule, sont

perçus comme moins intelligibles par les participants et que l'ajout des modalités vidéo et/ou texte améliore l'intelligibilité. Les extraits vidéos les plus intelligibles sont ceux qui ont été présentés avec les trois modalités audio, vidéo et texte. Ces résultats montrent que, pour ce corpus, combiner les modalités audio, vidéo et texte permet d'optimiser l'intelligibilité.

5 Discussion et conclusions

Dans le cadre d'une étude de la compréhension de contenu audiovisuel, un corpus composé d'évaluations subjectives de 55 extraits extraits de 15 films, réalisées par 15 professeurs de français langue étrangère, a été constitué. Ces évaluations ont permis d'étudier les facteurs qui influent sur la difficulté perçue mais aussi d'étudier l'influence des modalités sur la compréhension.

Concernant l'influence de la complexité du vocabulaire, de la complexité grammaticale et de l'intelligibilité de la parole sur la difficulté globale, au niveau du corpus, il a été confirmé qu'une corrélation positive existait entre ces facteurs et la difficulté globale. Des régressions linéaires multiples, prenant la difficulté globale comme variable dépendante et la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité comme variables indépendantes, confirment que la combinaison de ces trois variables permet de construire un modèle prédisant efficacement la difficulté globale, avec un R^2 ajusté atteignant une valeur de 0,82. Le fait que le R^2 ajusté n'atteigne pas la valeur de 1 peut s'expliquer par le fait que certains éléments jouant sur la difficulté globale n'ont pas été pris en considération. Des facteurs secondaires affectant la difficulté globale (par exemple des facteurs liés à l'aspect cognitif) n'ont pas été inclus dans cette étude. De plus, comme les évaluations ont été réalisées par des humains, d'autres facteurs, comme la fatigue des participants, ont pu influencer les évaluations.

En ce qui concerne l'influence des modalités sur la difficulté globale, il a été montré que les extraits présentés avec la modalité audio seule ont été perçus comme étant les plus difficiles par les participants. Rajouter les combinaisons vidéo et texte diminue la difficulté globale perçue et améliore l'intelligibilité de la parole : les extraits les plus faciles et les plus intelligibles du corpus étant ceux qui étaient présentés sous la condition AVT. La présence du texte doit permettre de contourner les problèmes liés à l'intelligibilité. La présence de la vidéo peut aider à améliorer l'intelligibilité de part les indices que peuvent donner les gestes, les expressions faciales et les mouvements des lèvres. En résumé, l'exploitation de toutes les modalités semble permettre de maximiser l'intelligibilité et de minimiser la difficulté globale perçue.

Ces conclusions résultent d'une étude subjective qui pourra conduire au développement d'un outil de prédiction automatique de la difficulté globale et à l'obtention d'une mesure objective de cette difficulté. La prochaine étape consistera à comparer les scores recueillis avec les scores obtenus automatiquement à partir de paramètres extraits des modalités audio, texte et vidéo, et à différentes modélisations pour voir si la complexité du vocabulaire, de la grammaire et l'intelligibilité peuvent être prédites. L'intérêt de cette étude était de recueillir des scores reflétant la perception d'experts de la difficulté des documents audiovisuels et de déterminer les facteurs influençant la compréhensibilité tout en incluant les variations de perception de chacun des experts. Les critères mis en évidence pourront être intégrés dans les travaux futurs sur la prédiction de la difficulté globale, à leur tour évalués par des enseignants de langue ou des apprenants.

Références

ADANK P., EVANS B., STUART-SMITH J. & SCOTT S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology : Human perception and performance*, **35(2)**, 520.

BOYLE J. (1984). Factors affecting listening comprehension. *ELT Journal*, **38(1)**, 34–38.

BUCK G., Éd. (2001). *Assessing Listening*. Cambridge University Press.

CARROW-WOOLFOLK E. (1999). Comprehensive assessment of spoken language. *Bloomington, MN : Pearson Assessment*.

CHALL J. S. & DALE E., Éds. (1995). *Readability revisited : The new Dale-Chall readability formula*. Brookline Books.

CHANG A.-S. & READ J. (2008). Reducing listening text anxiety through various forms of listening support. *TESL-EJ*, **12(1)**, 1–25.

CORLEY M. & HARTSUIKER R. (2003). Hesitation in speech can... um... help a listener understand. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **25(25)**, 276–281.

DAHL T. & LUDVIGSEN S. (2014). How I see what you're saying : The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, **98(3)**, 813–833.

FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32(3)**, 221–233.

FRANÇOIS T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. *Actes de RECITAL 2009*.

GOH C. (1994). How much do learners know about the factors that influence their listening comprehension ? *Hong Kong Journal of Applied Linguistics*, **4(1)**, 17–42.

HENRY G., Éd. (1975). *Comment mesurer la lisibilité*. Labor.

LIVELY B. & PRESSEY S. (1923). A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, **9**, 938–398.

MARKHAM P., PETER L. & MCCARTHY T. (2001). The effects of native language vs target language captions on foreign language students' dvd video comprehension. *Foreign Language Annals*, **34(5)**, 439–445.

NISSAN S., DEVINCENZI F. & TANG K. (1995). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *ETS Research Report Series*, **2**, i–42.

PEREZ M., NOORTGATE W. & DESMET P. (2013). Captioned video for L2 listening and vocabulary learning : A meta-analysis. *System*, **41(3)**, 720–739.

SUEYOSHI A. & HARDISON D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, **55**, 661–699.

TRAVERSO V., Éd. (2013). *L'analyse des conversations*. Armand Colin.

VOSS B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, **22(2)**, 129–144.

WONG J. & WARING H., Éds. (2010). *Conversation Analysis and Second Language Pedagogy*. Taylor & Francis.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Évaluer l'intelligibilité, mots ou pseudo-mots ? Comparaison entre deux groupes d'auditeurs

Marie Rebourg[1], Muriel Lalain[1], Alain Ghio[1], Corinne Fredouille[2], Nicolas Fakhry[1, 3], Virginie Woisard[4]

(1) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Laboratoire d'Informatique d'Avignon, Avignon, France
(3) Service ORL, APHM, La Conception, Marseille, France
(4) Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France
marie.rebourg@univ-amu.fr, muriel.lalain@univ-amu.fr, alain.ghio@univ-amu.fr, corinne.fredouille@univ-avignon.fr, nicolas.fakhry@ap-hm.fr, woisard.v@chu-toulouse.fr

RÉSUMÉ

La perte d'intelligibilité représente une plainte importante des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. L'évaluation de l'intelligibilité est essentielle dans le parcours de soin, mais les tests existants ne sont pas satisfaisants. Basés sur la perception de listes de mots par des auditeurs entraînés à restaurer des séquences sonores dégradées, ils conduisent souvent à une sous-évaluation des déficits. Nous avons proposé une nouvelle tâche d'évaluation de l'intelligibilité, la tâche de décodage acoustico phonétique (DAP), basée sur l'utilisation de pseudo-mots (Astésano *et al.*, 2018; Ghio *et al.*, 2018; Ghio et al., soumis, Lalain et al., sous presse). Dans cette étude, nous évaluons la capacité de la tâche DAP à neutraliser les effets de restauration lexicale et d'expertise auditive clinique. Les résultats montrent que contrairement à une évaluation de l'intelligibilité basée sur des mots, une évaluation basée sur des pseudo-mots permet d'obtenir des scores de Déviation Phonologique Perçue (DPP) stables au cours du temps quel que soit le degré d'expertise des auditeurs, naïfs ou cliniciens.

ABSTRACT

Assess intelligibility, words or pseudo-words? Comparison between two groups of listeners

Loss of intelligibility is a significant complaint from patients treated for Head and Neck cancer. Intelligibility assessment is essential in the treatment process, but the existing tests are not satisfactory. Based on the perception of word lists by listeners trained to restore degraded sound sequences, they often lead to an underestimation of deficits. We have proposed a new task for evaluating intelligibility, the acoustical phonetic decoding (DAP) task, based on the use of pseudo-words (Astésano *et al.*, 2018; Ghio *et al.*, 2018; Ghio et al., soumis). In this study, we assess the ability of the DAP task to neutralize the effects of lexical restoration and clinical auditory expertise. The results show that, unlike a word-based intelligibility assessment, a pseudo-word assessment makes it possible to obtain stable Perceived Phonological Deviation (PPD) scores over time regardless of the level of expertise. listeners, naive or clinicians.

MOTS-CLÉS : Phonétique Clinique, Intelligibilité, Trouble de la Production de la Parole, Cancer VADS

KEYWORDS: Clinical phonetic, Intelligibility, Speech disorders, Head and Neck cancer

Volume 1 : Journées d'Études sur la Parole, pages 543–551.
HAL : hal-02798584.

1 L'intelligibilité

La perte d'intelligibilité constitue une plainte récurrente des patients atteints de Troubles de la Production de la Parole (TPP). Dans la cadre de la prise en charge des cancers ORL, les traitements dont bénéficient les patients (chirurgie, chimiothérapie, radiothérapie) entraînent un déficit communicationnel notable. Dans le parcours de soin, l'évaluation de l'intelligibilité est essentielle puisqu'elle permet de mesurer le handicap à la communication en évaluant les composantes linguistiques préservées et/ou dégradées pour établir la prise en charge orthophonique. Elle permet également de mesurer l'effet du traitement préalablement établi dans le parcours de soin.

La définition de l'intelligibilité ne fait pas consensus dans la littérature, elle est généralement définie comme « le degré de précision avec lequel un message est compris par un auditeur » (Yorkston, Dowden et Beukelman, 1992) ; dans le cadre de notre étude nous préférons adopter le point de vue de Hustad, Jones et Dailey (2003) selon lesquels les processus de bas niveaux supportent des informations « dépendantes du signal » alors que les informations « indépendantes du signal » sont liées aux processus de haut niveau. Il apparaît essentiel de préciser le concept de compréhension, défini par Lindblom (1990), comme « L'intégration à la fois des informations acoustico-phonétiques et de toutes les informations pertinentes indépendantes du signal qui permettent de comprendre un message parlé dans une situation de communication particulière ». Cette dernière met en lien le niveau de la compréhension avec les informations de haut niveau (top down) portées par le contexte et les connaissances générales d'un auditeur. En revanche, dans le cadre de ces travaux nous définissons l'intelligibilité, au sens de Keintz, Bunton et Hoit (2007), comme « la quantité de parole comprise à partir du seul signal acoustique ». Ainsi, l'évaluation de l'intelligibilité repose sur la quantité de signal acoustique perçu et correctement identifié par un auditeur. Cette définition établit un lien étroit entre l'intelligibilité et les informations dépendantes du signal liées aux processus de décodage des sons de parole, soit le bas niveau (bottom up).

1.1 L'évaluation clinique de l'intelligibilité

La méthode d'évaluation utilisée par les cliniciens repose sur la perception d'un auditeur. Cette évaluation est peu satisfaisante, mais elle constitue le *gold standard* de l'évaluation (Balaguer *et al.*, 2019)**.** La perception de la parole repose sur l'intégration d'informations de haut niveau, issues du contexte, et de bas niveau qui concernent l'identification des sons de parole. L'intégration de ces informations permet à un auditeur de restaurer les séquences sonores dégradées (Warren, 1984; Warren *et al.*, 1997) et ainsi optimise le décodage et la compréhension d'un message. La performance de ces mécanismes est essentielle à la communication dans la vie quotidienne, mais perturbe l'évaluation de l'intelligibilité en contexte clinique en entraînant une sous-évaluation des déficits.

En effet, le contexte spécifique à l'évaluation clinique augmente les effets top down. Le matériel linguistique des tests existants, Batterie d'Evaluation Clinique de la Dysarthrie (BECD) (Auzou et Rolland-Monnoury, 2006), Frenchay Dysarthry Assessment 2 (FDA2) adapté au français, (Blanc *et al.*, 2014) se composent de listes de mots courtes et fermées qui, facilement mémorisables, finissent par intégrer le stock de connaissances linguistiques générales des auditeurs. De plus, les connaissances du clinicien lui permettent de faire des prédictions sur les déficits des patients : la connaissance du dossier médical, du parcours de soin, du patient en tant que personne et de la pathologie influencent son jugement / évaluation. Enfin, une étude (Jarzé *et al.*, 2017) a mis en évidence une grande variabilité dans la perception des troubles de la parole pathologique. L'ensemble de ces facteurs, associés à la pratique et à l'utilisation répétée du test par le clinicien le conduisent à

apprendre le matériel linguistique ce qui entraîne la restauration lexicale des items et donc une sous-évaluation des déficits.
En définitive, les processus de restauration perceptive sont permis par le recours aux informations de haut niveau, accentués par le contexte spécifique de l'évaluation clinique et conduisent à une évaluation subjective. Bien que peu satisfaisante, la méthode d'évaluation de l'intelligibilité repose sur une évaluation perceptive, dont on ne peut s'affranchir, puisque par définition l'intelligibilité implique la production d'un locuteur et la perception d'un auditeur. On peut alors se demander comment rendre l'évaluation plus objective. Comment générer une évaluation qui force le recours aux informations de bas niveau et neutralise les processus de restauration lexicale ?

La solution que nous avons proposée (Ghio *et al.*, 2017) consiste à utiliser des listes de pseudo-mots pour limiter le recours aux informations de haut niveau et forcer l'auditeur à s'appuyer essentiellement sur des informations de bas niveau en le plaçant dans une situation de décodage acoustico-phonétique.

1.2 Le recours aux pseudo-mots

Les pseudo-mots sont des séquences qui suivent les règles phonotactiques du français. Dans le cadre du projet de recherche C2SI, plus de 80 000 formes ont été générées et permettent de constituer aléatoirement un très grand nombre de listes respectant des contraintes prédéfinies (Ghio *et al.*, 2018). Le principe consiste à faire produire une liste de pseudo-mots qui sera ensuite perçue et transcrite par un auditeur. Ex de pseudo-mots : *chouvi, granchu, soga, vosso, zinvo, rouzant, rumo, preglo, coubin*. Nous calculons ensuite un score de Déviation Phonologique Perçue (DPP) en termes de nombre de traits moyen altérés par phonème. Ce score est issu de la comparaison entre une cible qui devait être produite et ce qui a été perçu par un auditeur. Le calcul du score DPP repose sur une matrice de confusion intégrant la distance de Levenshtein à l'algorithme de Wagner Fischer.

Les études conduites jusqu'à présent pour éprouver la robustesse de la tâche de DAP ont montré des résultats positifs (i) pour sa capacité à discriminer deux groupes de locuteurs : patients VS témoins (Ghio *et al.*, 2018), (ii) car quelle que soit la liste utilisée l'évaluation reste stable, les listes de pseudo-mots utilisées sont donc équivalentes (Ghio *et al.*, 2019), (iii) et les scores de Déviation Phonologique Perçues sont corrélés avec les évaluations globales de l'intelligibilité (Lalain, M. *et al*., sous presse).

Il reste toutefois des questions pour valider les critères de pertinence et d'objectivité de la tâche de DAP. Ainsi, on peut se demander si une évaluation basée sur l'utilisation de pseudo-mots est plus pertinente qu'une évaluation basée sur des mots, autrement dit, quand les auditeurs sont confrontés plusieurs fois aux mêmes stimuli, sont-ils capables de les mémoriser et de les restaurer ? Enfin, l'utilisation de pseudo-mots, si elle permet une évaluation de l'intelligibilité plus objective chez des auditeurs naïfs, permet-elle de neutraliser l'expertise auditive des cliniciens ? Répondre à ces deux questions est l'objectif de l'étude que nous présentons ici.

2 Méthodologie

Afin de répondre à ces questions de recherche nous avons constitué un corpus composé des enregistrements de 20 locuteurs. Ce corpus rassemble le matériel linguistique que nous avons utilisé dans le protocole expérimental des tests de jugement perceptif de l'intelligibilité au cours desquels nous collectons les transcriptions de différents groupes d'auditeurs. Ces transcriptions font l'objet d'un traitement spécifique permettant d'obtenir un score de Déviation Phonologique Perçue (DPP). Ces scores (VD) seront mis en lien avec la nature du matériel linguistique (VI : Mots VS Pseudo-mots) et les groupes d'auditeurs (VI : Naïf VS Expert) dans nos analyses statistiques.

2.1 Corpus

Pour les besoins de cette étude 10 patients et 10 témoins ont chacun été enregistrés sur la production d'une liste de 50 mots (items de la BECD) et sur la production de 2 listes de 52 pseudo-mots issues du matériel linguistique constitué pour la tâche de DAP. Soit 154 productions par locuteur. Nous avons donc enregistré 154 stimuli x 20 locuteurs, soit 3080 stimuli en tout et respectivement 1000 stimuli pour le corpus BECD et 2080 stimuli pour le corpus DAP. Les locuteurs ont été divisés en deux groupes (A et B) de 10 locuteurs dans lesquels on retrouve 5 patients et 5 témoins, impliquant que les corpus BECD et DAP ont eux aussi été divisés en deux sous corpus composés chacun de 500 stimuli BECD et 1040 stimuli DAP. Ceci afin de limiter le nombre de stimuli transcrit par les auditeurs.

A partir de chaque sous corpus, des listes de stimuli ont été constituées. Nous obtenons alors 6 listes de productions pour les corpus BECD avec 167 stimuli par liste et 12 listes pour les corpus DAP avec 174 stimuli par liste. Les listes 1 à 3 du corpus BECD et les listes 1 à 3 et 7 à 9 du corpus DAP sont produites par le groupe de locuteurs A. Les listes 4 à 6 du corps BECD et listes 4 à 6 et 10 à 12 du corpus DAP ont été produites par le groupe de locuteurs B. Ces listes ont été utilisées pour effectuer des tests de jugement perceptif de l'intelligibilité auprès de deux populations d'auditeurs, naïfs et cliniciens.

2.2 Design expérimental

Dans ces tests, l'auditeur écoute dans un casque les productions de patients (traités pour un cancer des VADS) et de témoins, et transcrit, sur un clavier, ce qu'il entend.

Chaque auditeur écoute et transcrit 3 listes BECD et 6 listes DAP, produites par un groupe de locuteurs (A ou B). Chaque liste est transcrite par trois auditeurs différents et l'ordre de présentation des listes est contrôlé de manière que chaque liste apparaisse dans chaque position possible (T1, T2 et T3) et soit transcrite par trois auditeurs différents. Nous obtenons donc 3 transcriptions, proposées par 3 auditeurs différents, et ce dans chacune des positions (T1, T2 et T3), soit au total 9 transcriptions par liste. L'ordre de passation des listes est résumé dans le tableau de design expérimental suivant :

			T1	T2	T3
Groupe de locuteurs A	3 auditeurs ≠	BECD	1	2	3
		DAP 1	1	2	3
		DAP 2	7	8	9
	3 auditeurs ≠	BECD	2	3	1
		DAP 1	3	1	2
		DAP 2	9	7	8
	3 auditeurs ≠	BECD	3	1	2
		DAP 1	2	3	1
		DAP 2	8	9	7
Groupe de locuteurs B	3 auditeurs ≠	BECD	4	5	6
		DAP 1	4	5	6
		DAP 2	10	11	12

Chaque auditeur transcrit 3 blocs de stimuli (BECD, DAP1 puis DAP2)
Chaque bloc comprend 3 listes de stimuli

L'ordre de passation des listes au cours de la tâche est contrôlé. Ex : La liste 1 est transcrite en T1, puis, T2 et T3

3 auditeurs ≠	BECD	6	4	5
	DAP 1	5	6	4
	DAP 2	11	12	10
3 auditeurs ≠	BECD	5	6	4
	DAP 1	6	4	5
	DAP 2	12	10	11

TABLE 1 : Tableau récapitulatif du design expérimental du test de jugement perceptif

Ce protocole expérimental a permis de réaliser des tests de jugement perceptif de l'intelligibilité auprès de deux groupes de population.

2.3 Test de jugement perceptif

18 auditeurs naïfs, natifs de langue française, sans problème de vue ou d'audition non corrigés, ayant un bon niveau en orthographe et 18 auditeurs experts en écoute de la parole pathologique, tous orthophonistes, ont participé à cette étude. Les auditeurs naïfs ont réalisé la tâche en débutant par la transcription orthographique des mots issus de la BECD puis en transcrivant les listes de pseudo-mots du DAP. Pour des questions d'organisation les auditeurs experts ont réalisé la tâche en débutant par la transcription des listes de pseudo-mots du DAP, puis ont poursuivi l'exercice en transcrivant les mots issus de la BECD. Les auditeurs ont reçu pour consigne de toujours donner une transcription, au plus près de ce qu'ils perçoivent et identifient, en suivant les règles orthographiques du français. Ils ont été dédommagés en ticket Kadeos. Ce design expérimental supporte un double objectif, il permet de comparer les scores d'intelligibilité obtenus par la production et la transcription de deux types de matériel linguistique : des mots et des pseudo-mots ; et de comparer les scores obtenus par les transcriptions de deux catégories d'auditeurs : des auditeurs naïfs et experts. Ce test de jugement perceptif de l'intelligibilité a été mené au CEP (http://cep.lpl-aix.fr/), à l'aide de la station de perception PercEval (André *et al.*, 2003).

2.4 Traitement des données

Pour réaliser l'analyse statistique des données collectées lors des tests de jugement perceptif de l'intelligibilité, plusieurs étapes sont nécessaires.

Tout d'abord, nous récupérons les données brutes au sortir du logiciel d'expérimentation PercEval (André *et al.*, 2003). Nous réalisons différentes opérations de prétraitement des données visant à leur donner un format compatible avec les outils de calcul du score DPP. En premier lieu nous recodons les transcriptions orthographiques des auditeurs : si un auditeur propose une transcription avec une lettre accentuée (e.g : « é ») elle est encodée par un symbole dans le fichier de réponse (e.g : « %E9 »), nous remplaçons les symboles du fichier réponse par la lettre correspondante transcrite initialement par l'auditeur. Ensuite nous transformons ces transcriptions orthographiques en phonétique, c'est l'étape de phonétisation ((LIA_Phon, Béchet, 2001) et Lexique.org). Ces transcriptions phonétisées, compatibles avec la matrice de confusion utilisée pour calculer les scores de Déviation Phonologique Perçue (DPP), sont utilisées pour l'analyse des données.

Nous comparons alors les transcriptions phonétiques des réponses des auditeurs aux tests de jugement perceptif de l'intelligibilité avec les transcriptions phonétiques des cibles qui devaient être prononcées par les locuteurs. L'algorithme de Wagner Fischer permet le calcul de distance d'édition entre deux

chaînes de caractères phonétiques. Nous utilisons une matrice de confusion qui attribue un coût en termes de distance basé sur la théorie des traits, développé par A. Ghio dans le cadre de l'analyse des données DAP (Ghio *et al.*, 2018). L'algorithme intègre la distance de Levenshtein, qui considère trois opérations d'éditions élémentaires : la suppression, l'insertion ou la substitution d'un caractère, ce qui permet de traiter les altérations sur les deux axes syntagmatique et paradigmatique. Soit, une distance entre les transcriptions des cibles qui devaient être prononcées par les locuteurs et les transcriptions effectives des réponses des auditeurs.

Les scores de distance cumulée sont obtenus en divisant le score donné par la matrice de confusion pour chaque stimulus par le nombre de caractères de la cible phonétique. Nous obtenons alors des scores de Déviation Phonologique Perçue (DPP) en termes de distance cumulée à la cible, qui représentent le nombre moyen de traits altérés par phonème. Ces scores sont les variables dépendantes (VD) des analyses statistiques. De plus, pour chaque auditeur, un numéro a été assigné à chaque item en fonction de l'ordre de la passation, ainsi, le premier stimulus transcrit est numéroté 1, le second stimulus transcrit est numéroté 2, et ainsi de suite jusqu'à 500 pour les items lexicaux BECD et jusqu'à 1040 pour les items non-lexicaux DAP.

Pour l'analyse statistique, nous comparerons les scores DPP à l'aide d'une analyse de variance appliquée afin de tester les différences significatives des scores d'intelligibilité (VD) en fonction du matériel d'évaluation (Variables Indépendantes (VI) : Mots vs Pseudo-mots) et en fonction du groupe d'auditeurs (VI : Naïfs vs Experts).

3 Résultats et discussion

Une analyse de variance (ANOVA) a été conduite, afin de tester les effets simples et les effets d'interaction entre les différents facteurs et ainsi de déterminer les liens entre les variables testées. Nous avons utilisé le score DPP en VD et en VI le groupe d'auditeurs (facteur 2 niveaux : Experts vs Naïfs), le type de matériel linguistique (facteur 2 niveaux : Mots BECD vs Pseudo-mots DAP), et le groupe de locuteurs (Patients et Témoins) en facteur. Les résultats montrent un effet simple des groupes de locuteurs, d'auditeurs et du matériel linguistique ($p < 0.001$). Cela signifie que, dans nos données, les groupes de locuteurs sont distingués, tout comme les groupes d'auditeurs et type de matériel linguistique. Les résultats révèlent également un effet d'interaction significatif entre Locuteur et Auditeur ($p = 0.027$) et entre Matériel linguistique et Auditeur ($p = 0.047$). Cela montre que les locuteurs sont évalués différemment par les deux groupes d'auditeurs et que les groupes d'auditeurs ne proposent pas des scores équivalents avec les deux matériaux linguistiques. Cependant, ils ne montrent pas d'effet d'interaction Locuteur et Matériel linguistique ($p = 0.44$), donc les deux groupes de locuteurs ne sont pas évalués différemment par les matériaux linguistiques. En revanche, ils révèlent également un effet d'interaction Locuteur, Auditeur et Matériel linguistique ($p < 0.001$) ; les groupes d'auditeurs évaluent différemment les groupes de locuteurs en fonction du matériel linguistique utilisé.

Pour établir le sens du lien entre nos variables nous avons établi une matrice de corrélation qui intègre l'ordre de passation des items (VI : Order) à la place du groupe de locuteurs. Les résultats révèlent un effet de l'ordre des items dans la passation pour le matériel linguistique lexical (Mots BECD) et ce pour les deux groupes d'auditeurs, Expert ($r = -0.197$, $p < 0.001$) et Naïfs ($r = -0.302$, $p < 0.001$) (FIGURE 1). En revanche, aucun effet n'est mis en exergue dans une évaluation basée sur l'utilisation de pseudo-mots (DAP) et ce pour les deux groupes d'auditeurs, Experts (r -0.058, $p = 0.06$) et Naïfs ($r = -0.028$, $p = 0.37$) (FIGURE 3).

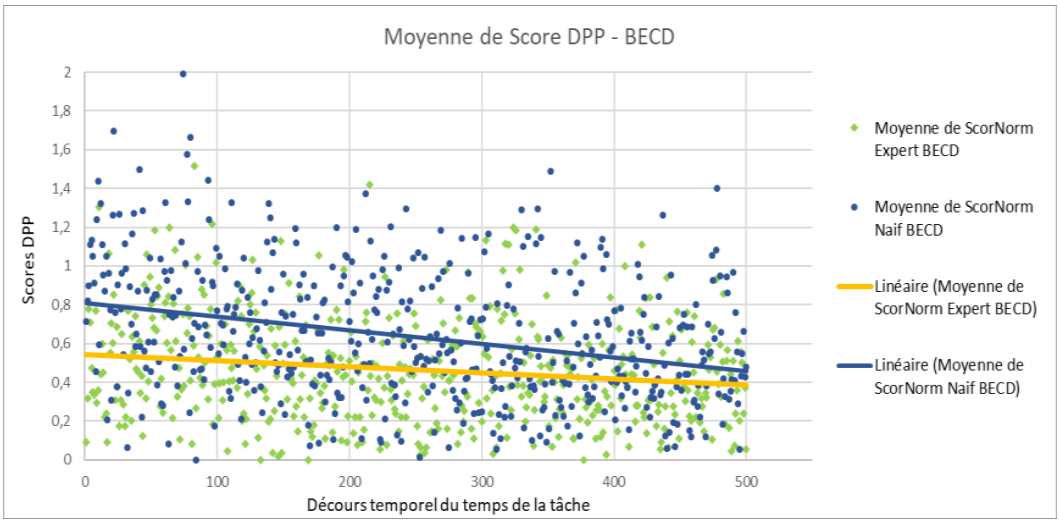


FIGURE 1 : Scores DPP attribués par une évaluation BECD par deux types d'auditeurs

Ceci indique que, dans une évaluation utilisant du matériel linguistique lexical (Mots BECD) il existe un effet d'apprentissage du matériel linguistique au cours de la tâche. L'accès au lexique mental permis par l'emploi de mots favorise la mémorisation des items et leur restauration perceptive lorsque leur production est dégradée. Les listes de mots courtes et fermées classiquement utilisées (Auzou et Rolland-Monnoury, 2006; Blanc *et al.*, 2014) facilitent ces effets. Défavorables dans le cadre de l'évaluation clinique de l'intelligibilité, car conduisant à la sous-évaluation des déficits, ces effets se trouvent neutralisés dans une évaluation s'appuyant sur du matériel linguistique non lexical (Pseudo-mots DAP). En effet, les pseudo-mots DAP permettent d'obtenir une évaluation stable au cours du temps, critère qui valide la fiabilité de la tâche et du matériel linguistique, et de neutraliser l'effet d'apprentissage et de mémorisation, et donc les effets de restauration perceptive liés au matériel linguistique.

Ces résultats confirment l'hypothèse selon laquelle l'emploi de pseudo-mots est pertinent pour évaluer l'intelligibilité (au sens de la définition de Keintz, Bunton et Hoit (2007)) et qu'il permet de neutraliser l'effet d'apprentissage du matériel linguistique, et ce, pour les deux groupes d'auditeurs, naïfs et cliniciens.

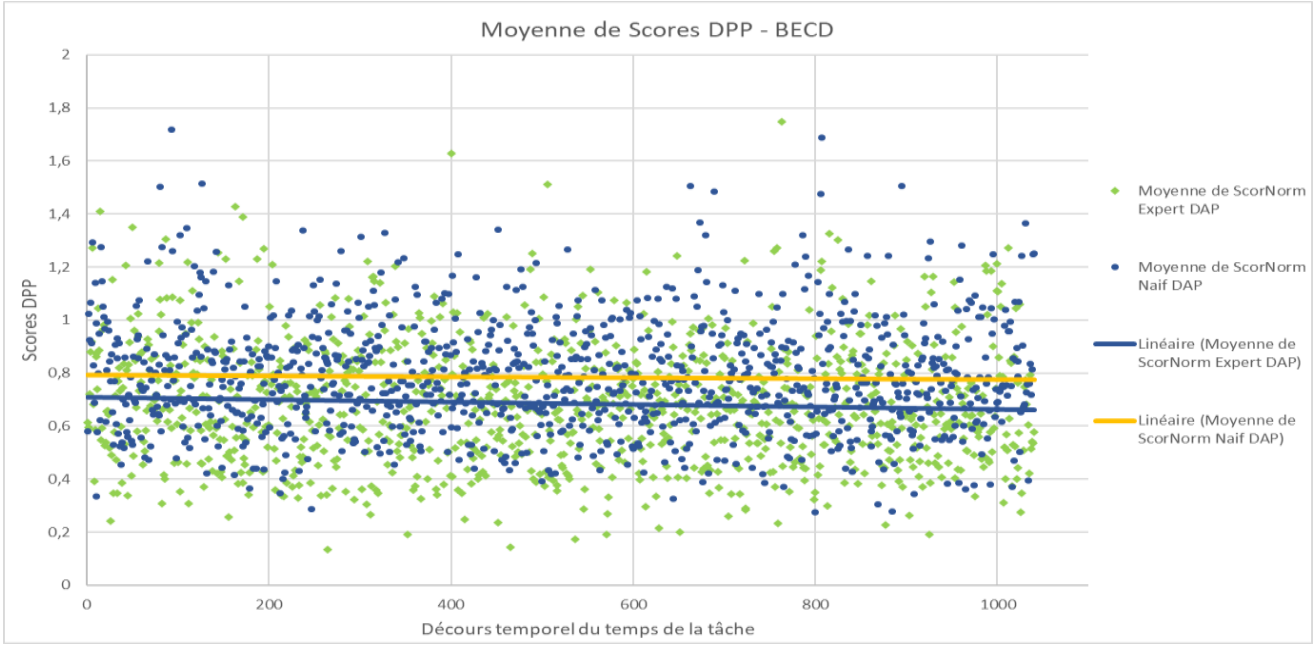


FIGURE 2 : Scores DPP attribués par une évaluation DAP par deux types d'auditeurs

Dans les évaluations BECD la différence de scores moyens entre les deux groupes d'auditeurs, Experts ($\mu = 0{,}464$) et Naïfs ($\mu = 0{,}634$) était attendue car les auditeurs Experts ont déjà été confrontés aux items de la BECD au cours de leur pratique clinique. Cependant, cette différence nous interpelle dans les évaluations DAP, Experts ($\mu = 0.685$) et Naïfs ($\mu = 0.787$). Ces scores indiquent que les auditeurs cliniciens sont de meilleurs décodeurs que les auditeurs naïfs, du fait de leur expertise auditive en écoute de la parole pathologique. Ainsi, pour mesurer l'effet d'expertise sur la tâche de DAP, nous avons comparé les scores DPP moyens attribués par les deux groupes d'auditeurs. La différence observée s'est révélée significative ($p < 0.001$). Les experts proposent des scores DPP significativement plus bas que les auditeurs naïfs. Cette différence moyenne correspond à une différence moyenne de 0.1 trait d'écart moyen par phonème, ce qui est très faible. Nous avons donc cherché à évaluer la taille de cet effet en calculant un coefficient d de Cohen (« cohen.d » in R). Un 'd' aux alentours de 0.2 est considéré comme « faible », à 0.5 « médium » et à 0.8 « fort ». Le d de Cohen estimé pour le DAP, entre les auditeurs naïfs et experts est (d : 0.102) dit négligeable. La différence de scores DPP moyens attribués par les auditeurs naïfs et les auditeurs experts ne perturbe donc pas l'évaluation car sa faiblesse montre que les deux groupes d'auditeurs n'évaluent finalement pas différemment les locuteurs avec le DAP. Cela indique que l'effet d'expertise auditive des cliniciens est neutralisé lors d'une évaluation de l'intelligibilité par la tâche de DAP.

Ces résultats corroborent l'hypothèse selon laquelle l'utilisation de pseudo-mots dans l'évaluation de l'intelligibilité permet de neutraliser l'effet d'expertise auditive des cliniciens.

4 Conclusion

Cette étude visait à valider la pertinence de l'utilisation de pseudo-mots, par rapport à des mots pour évaluer l'intelligibilité. Plus précisément, nous avons cherché à montrer qu'une évaluation basée sur l'utilisation de pseudo-mots permettait d'obtenir des scores de Déviation Phonologique Perçue (DPP) plus stables dans le temps qu'une évaluation basée sur des mots, et ce lorsque cette évaluation est conduite par des auditeurs naïfs aussi bien que par des experts. Nos résultats ont confirmé nos hypothèses : la tâche DAP permet d'obtenir un score DPP objectif, en nombre moyen de traits altérés par phonème, qui est représentatif de la sévérité du déficit des patients et du degré de dégradation du signal perçu. La tâche de DAP permet ainsi de neutraliser les effets d'apprentissage et mémorisation du matériel linguistique, ainsi que l'effet d'expertise auditive des cliniciens.

Ces résultats s'inscrivent dans la lignée des études précédemment conduites (Voir Section 1.1) pour évaluer les critères d'objectivité et de pertinence de la tâche DAP. Une partie des données de cette étude ont également permis d'évaluer la stabilité des scores DPP obtenus lorsque ceux-ci sont calculés non plus à partir de 52 pseudo-mots, mais à partir de 16 (Marczyck et al, soumis). Cette réduction, basée sur des critères phonologiques a montré de très bons résultats dans le cadre d'une évaluation par des auditeurs naïfs. Dans la suite de notre étude présentée ici, nous évaluerons la robustesse de l'évaluation de l'intelligibilité basée sur des pseudo-mots, auprès d'auditeurs experts, en utilisant des listes réduites à 16 items.

Remerciements

Ce travail de recherche a été réalisé dans le cadre d'un contrat doctoral soutenu par la Ligue Contre le Cancer. Il s'inscrit également dans le cadre de projet de recherche soutenu par la subvention n ° 2014-135 de l'Institut National pour le Cancer (INCA) projet C2SI et par la subvention ANR-18-CE45-0008 de l'Agence Nationale de la Recherche en 2018 Projet RUGBI

Références

ANDRE, C. *ET AL.* (2003): «PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception», in *XVth ICPhS. ICPhS*, Barcelone, Espagne, p. 1421-1424.

ASTESANO, C. *ET AL.* (2018): «Carcinologic Speech Severity Index Project: A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer», in. *Language Resources and Evaluation Conference*, Miyazaki.

AUZOU, P. ET ROLLAND-MONNOURY, V. (2006): *Batterie d'évaluation clinique de la dysarthrie*. Ortho Edition. France: ORTHO.

BALAGUER, M. *ET AL.* (2019): «Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer», *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 136(5), p. 355-359. doi: 10.1016/j.anorl.2019.05.012.

BECHET, F. (2001): «LIA PHON : Un système complet de phonétisation de textes», *Traitement Automatique des Langues, TAL*. (TAL - ATALA), 42(1), p. 47-67.

BLANC, E. *ET AL.* (2014): «Adaptation en français du test d'intelligibilité de la version révisée du « Frenchay Dysarthria Assessment » (FDA-2)», in *Congrès de la Société Française de Phoniatrie*. Paris, France. Disponible sur: https://hal.archives-ouvertes.fr/hal-01615204.

GHIO, A. *ET AL.* (2017): «Du décodage acoustico-phonétique pour mesurer l'intelligibilité de locuteurs atteints de troubles de production de la parole», in. *Journée de Phonétique Clinique (7éme)*, Paris.

GHIO, A. *ET AL.* (2018): «Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique», in *XXXIIe Journées d'Études sur la Parole*. *XXXIIe Journées d'Études sur la Parole*, ISCA, p. 285-293. doi: 10.21437/JEP.2018-33.

GHIO, A. *ET AL.* (2019): «Application d'un test d'intelligibilité à partir de pseudo- mots dans le cas de patients post traitement de cancers des VADS», in *Journées de Phonétique Clinique*. Mons, Belgium. Disponible sur: https://hal.archives-ouvertes.fr/hal-02098836.

GHIO, A. *ET AL.* (2020), «Evaluation de l'intelligibilité de patients avec traitement du cancer des cavités orales et pharyngales», submitted in this volume.

HUSTAD, K. C., JONES, T. ET DAILEY, S. (2003): «Implementing Speech Supplementation Strategies: Effects on Intelligibility and Speech Rate of Individuals With Chronic Severe Dysarthria», *Journal of Speech, Language, and Hearing Research*, 46(2), p. 462-474. doi: 10.1044/1092-4388(2003/038).

JARZE, S. *ET AL.* (2017): «Analyse perceptive des voix dysphoniques. Influences et corrélations entre les dimensions G, R et B de l'échelle d'Hirano», in *Journées de Phonétique Clinique*. Paris, France. Disponible sur: https://hal.archives-ouvertes.fr/hal-01615030.

KEINTZ, C. K., BUNTON, K. ET HOIT, J. D. (2007): «Influence of Visual Information on the Intelligibility of Dysarthric Speech», *American Journal of Speech-Language Pathology*, 16(3), p. 222-234. doi: 10.1044/1058-0360(2007/027).

LINDBLOM, B. (1990): «On the communication process: Speaker listener interaction and the development of speech.», in *Augmentative and Alternative Communication*. (6), p. 220-230.

WARREN, R. M. (1984): «Perceptual restoration of obliterated sounds», *Psychological Bulletin*, 96(2), p. 371-383. doi: 10.1037/0033-2909.96.2.371.

WARREN, R. M. *ET AL.* (1997): «Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps», *Perception & Psychophysics*, 59(2), p. 275-283. doi: 10.3758/BF03211895.

YORKSTON, K. M., DOWDEN, P. A. ET BEUKELMAN, D. R. (1992): «Intelligibility measurement as a tool in the clinical management of dysarthric speakers», in *Intelligibility in speech Disorders : Theory, measurement and management*. Raymond D. Kent. Madison, Wisconsin: John Benjamins Publishing Company, p. 265-286. Disponible sur: https://benjamins.com/catalog/sspcl.1.08yor

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Sur le voisement des consonnes fricatives finales en français du Québec

Josiane Riverin-Coutlée
Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität
Schellingstraβe 3, 80799 Munich, Allemagne
josiane.riverin@phonetik.uni-muenchen.de

RÉSUMÉ

Cette étude s'intéresse aux indices acoustiques qui concourent à distinguer les fricatives non voisées /f s ʃ/ et voisées /v z ʒ/ en position de finale absolue en français du Québec. La durée de la consonne elle-même, celle de la voyelle accentuée précédente et le taux de voisement consonantique sont les indices acoustiques examinés. La durée intrinsèque des voyelles, caractéristique importante de la variété à l'étude, est prise en compte lors de l'analyse des résultats, qui indiquent que les deux groupes de consonnes se distinguent en tous points. Les consonnes voisées ont une durée plus courte, présentent un taux de voisement supérieur quoique moindre que celui rapporté dans la littérature antérieure, et allongent les segments vocaliques précédents. Ce dernier phénomène se produit même lorsque la voyelle est intrinsèquement longue, révélant la robustesse de l'effet allongeant des consonnes voisées et l'extensibilité de la durée vocalique en français québécois.

ABSTRACT

On final fricative consonant voicing in Quebec French
This study looks at the acoustic cues distinguishing between unvoiced fricatives /f s ʃ/ and voiced /v z ʒ/ in absolute final position in Quebec French. The duration of the consonant itself, duration of the preceding stressed vowel and rate of consonantal voicing are the acoustic cues examined. Intrinsic vocalic length, an important feature of the studied variety, is taken into account in the analysis, the results of which indicate that the two groups of consonants are distinct in all manners. The voiced consonants have a shorter duration, greater rate of voicing – although it is smaller than that reported in previous literature – and they lengthen preceding vocalic segments. This latter phenomenon takes place even with intrinsically long vowels, therefore revealing the robustness of the voiced consonants' lengthening effect and expandability of vocalic length in Quebec French.

MOTS-CLÉS : français québécois, consonnes fricatives, voisement, longueur vocalique intrinsèque
KEYWORDS: Quebec French, fricative consonants, voicing, intrinsic vocalic length

1 Voisement : problématique et objectifs

Le voisement est à la base de nombreux contrastes phonologiques dans les langues du monde. Si en théorie, les segments de la parole peuvent être qualifiés de voisés lorsqu'ils sont produits avec vibration des cordes vocales, dans les faits, un ensemble d'indices concourent à ce que les auditeurs perçoivent certains segments comme étant voisés. Par exemple, en anglais, un très court intervalle de silence suivant l'explosion et précédant le début du segment suivant mène les auditeurs à percevoir

Volume 1 : Journées d'Études sur la Parole, pages 552–560.
HAL : hal-02798586.

qu'une occlusive est voisée (Lisker & Abramson, 1964). Dans la même langue, une voyelle longue est un indicateur aussi efficace, sinon meilleur, du voisement de la fricative suivante que le fait que celle-ci ait été produite avec vibration des cordes vocales (Denes, 1955; Raphael, 1972). Ce dernier phénomène se rattache à l'effet de voisement, ou *voicing effect*, attesté dans plusieurs langues (voir par exemple Pape & Jesus, 2015). L'effet de voisement consiste en l'allongement d'une voyelle en syllabe accentuée fermée par une consonne voisée, comparativement à la même voyelle en syllabe fermée par une consonne non voisée. Si l'effet peut se produire aussi bien avec les consonnes occlusives que fricatives, il serait encore plus marqué avec ces dernières (House & Fairbanks, 1953; Hogan & Rozsypal, 1980).

De fait, l'effet de voisement est reconnu se produire avec les consonnes fricatives du français. Rappelons que cette langue, y compris la variété parlée au Québec, comporte entre autres phonèmes trois paires de consonnes fricatives qui s'opposent par le voisement : /f v/, /s z/ et /ʃ ʒ/ (Fougeron & Smith, 1993; Martin, 1996). Lorsqu'elles se retrouvent en fin d'unité lexicale, on qualifie les consonnes fricatives voisées /v z ʒ/ (ainsi que /ʁ/) d'« allongeantes » (Fouché, 1959), puisque la voyelle accentuée les précédant se trouve sensiblement allongée à leur contact. Par exemple, dans le mot *sage*, la voyelle /a/ est plus longue que dans le mot *sache*. Parallèlement à la présence de cet effet de voisement, les consonnes /v z ʒ/ tendent à être dévoisées en finale absolue (Jatteau *et al.*, 2019). L'allongement vocalique joue donc un rôle prépondérant dans l'expression du voisement de ces consonnes en français. Ces tendances bien connues présentent néanmoins un intérêt de recherche particulier dès lors qu'il est question du français parlé au Québec (désormais FQ). En effet, les locuteurs de cette variété conservent des caractéristiques de durée vocalique historiques que plusieurs francophones d'Europe (quoique pas tous; voir entre autres Walter, 1982) ont perdues. Toutes choses étant égales par ailleurs, en syllabe fermée, les voyelles /o ø aɛ ɑ ɑ̃ ɔ̃ ɛ̃ œ̃/[1] possèdent une durée intrinsèquement plus longue que les autres voyelles. Par exemple, un locuteur du FQ prononcera *saute* [soːt], *sainte* [sɛ̃ːt], *sotte* [sɔt] et *sept* [sɛt]. Dans ces circonstances, on peut s'interroger sur ce qu'il advient de l'effet allongeant de /v z ʒ/ et de l'expression de leur voisement de manière plus générale.

Dans l'une des rares études cherchant à quantifier le voisement des fricatives en FQ, Jacques (1990) analyse un corpus de parole produite par quatre locuteurs montréalais en mesurant des durées vocaliques et consonantiques et en évaluant le taux de voisement consonantique en fonction de l'identification visuelle de barres de voisement sur des spectrogrammes. L'auteur parvient notamment aux trois constatations suivantes. Premièrement, bien que les fricatives soient plus sujettes au dévoisement lorsqu'elles occupent la position de finale absolue qu'initiale ou intervocalique, elles demeurent voisées sur environ un tiers de leur durée pour /z ʒ/ et deux tiers pour /v/ lorsqu'en finale. Deuxièmement, les fricatives finales voisées sont plus courtes que leurs homologues non voisées. Troisièmement, les voyelles suivies de /v z ʒ/ sont en moyenne légèrement plus longues que celles suivies de /f s ʃ/, mais sans doute moins que ne le laissait prévoir le titre de « consonnes allongeantes », et les écarts-types sont considérables. Jacques (1990) attribue cette importante variabilité et l'effet apparemment peu allongeant de /v z ʒ/ au fait que le corpus exploité dans le cadre de son étude ne lui permettait pas de tenir compte de la nature des voyelles analysées. Selon l'auteur, la durée intrinsèque des voyelles québécoises pourrait l'emporter sur l'allongement contextuel.

L'objectif de la présente contribution est de reprendre les choses là où Jacques (1990) les a laissées et de proposer une description des consonnes fricatives finales voisées et non voisées du FQ en tenant compte de la durée intrinsèque des voyelles accentuées les précédant, tout en cherchant à confirmer

[1] La voyelle /aɛ/ est celle que l'on retrouve dans des mots comme *fête*, *bêle* ou *maître*. Elle est phonologiquement distincte de /ɛ/ en FQ et habituellement diphtonguée.

les résultats obtenus par le chercheur en ayant recours à un échantillon de plus grande taille et à des techniques d'analyse modernes, en particulier une méthode de détection automatisée du voisement. Nous procéderons ainsi à l'examen de trois caractéristiques contribuant potentiellement à exprimer le voisement des fricatives finales en FQ : la durée vocalique, la durée consonantique et le taux de voisement consonantique.

2 Méthodologie

2.1 Locuteurs et corpus

Cinquante-cinq (55) locuteurs natifs du FQ, 42 femmes et 13 hommes âgés de 18 à 23 ans, ont été recrutés pour cette étude. Ils étaient tous étudiants universitaires au moment des enregistrements, qui se sont déroulés en septembre et octobre 2016 dans les villes de Saguenay et de Québec. Vingt-cinq (25) d'entre eux étaient originaires de la ville où ils étudiaient, les 30 autres venaient tout juste d'y emménager pour entreprendre un cursus universitaire et provenaient de diverses municipalités à travers la province. Aucun participant n'a déclaré avoir vécu de manière prolongée hors du Québec. Les enregistrements ont eu lieu en chambre anéchoïque et ont été effectués au format numérique à l'aide d'un appareil Zoom H4n (44 100 Hz, 16 bits).

Les participants ont été amenés à effectuer différentes tâches de lecture, parmi lesquelles deux sont exploitées dans le cadre de la présente étude, puisque les locuteurs y ont produit des consonnes fricatives en finale absolue. Notre objectif n'étant pas d'effectuer une comparaison inter-tâches, ce potentiel facteur de variation sera plutôt traité comme effet aléatoire lors de l'analyse statistique des données. La première tâche consistait en la lecture de courtes phrases hors contexte mais sémantiquement signifiantes, par exemple *Au soleil on bronze*, où le mot cible est *bronze*. Lors de la seconde tâche, les mêmes mots cibles étaient présentés à nouveau, mais de manière isolée. Les 26 mots cibles mono- ou bisyllabiques se terminant par une consonne fricative ainsi produits par les participants sont reproduits dans la Table 1. Les voyelles /e ø œ̃/ en sont absentes, car en raison des contraintes du lexique français, elles ne peuvent se retrouver en syllabe finale accentuée suivies de fricatives voisées et/ou non voisées. Toutes les voyelles de la Table 1 sont suivies d'une consonne voisée et d'une consonne non voisée, mais les combinaisons voyelle-consonne résultantes ne forment pas nécessairement de paires minimales fondées seulement sur le voisement. Puisque notre analyse mène à rassembler les voyelles en deux grandes catégories (courtes vs longues), l'absence de paires minimales pour chacune d'entre elles ne constitue pas un obstacle majeur. Enfin, si les segments composant des mots polysyllabiques sont parfois de durée plus courte que ceux des monosyllabes, les résultats présentés ci-après ne laissent pas entrevoir une telle tendance pour *chétive* et *débouche*.

	/i/	/y/	/u/	/o/	/ɛ/	/aɛ/	/œ/
C voisée	*chétive*	*fuse*	*rouge*	*chose*	*brève*	*beige*	*neuve*
C non voisée	*friche*	*puce*	*débouche*	*fosse*	*crèche*	*caisse*	*œuf*
	/ɔ/	/a/	/ɑ/	/ɑ̃/	/ɔ̃/	/ɛ̃/	
C voisée	*toge*	*page*	*âge*	*change*	*bronze*	*quinze*	
C non voisée	*bosse*	*face*	*lâche*	*France*	*ponce*	*rince*	

TABLE 1 : Liste des mots cibles produits à deux reprises par les participants. Les correspondances phonème-lexème reflètent la manière dont les locuteurs du FQ prononcent ces mots cibles

2.2 Analyses

Le signal sonore a été analysé à l'aide du logiciel *Praat* (Boersma & Weenink, 2020). Les voyelles et les consonnes fricatives ont été segmentées manuellement. Les segments vocaliques ont été identifiés d'après la présence de formants sur le spectrogramme et d'une intensité relativement importante, tandis que le début et la fin des fricatives ont été déterminés par l'identification visuelle de bruits de friction en hautes fréquences. En cas d'incertitude quant à la frontière entre voyelle et consonne, l'apparition des bruits de friction a été utilisée comme indice marquant le début de la consonne. Sur les 2 860 items du corpus analysé (55 locuteurs * 26 mots * 2 tâches), 148 ont été éliminés en raison de la production d'un schwa final, qui d'une part, faisait en sorte que la fricative ne se retrouvait plus en finale absolue, et d'autre part, modifiait la structure syllabique du mot de manière à ce que la fricative ne fasse plus partie de la même syllabe que la voyelle précédente (un mot C**VC** comme *rouge* devient C**V**.**C**V). Au final, ce sont donc 2 712 occurrences qui ont été analysées.

Rappelons que trois indices acoustiques sont examinés dans cette contribution : la durée vocalique, la durée consonantique et le taux de voisement consonantique. Si mesurer la durée des segments ne pose pas de défi particulier une fois réglés les problèmes de segmentation, il en va autrement du voisement. En effet, la méthode la plus directe pour évaluer le taux de voisement d'un segment dans *Praat* est d'utiliser la fonction *Voice Report* sous l'onglet *Pulses* de la fenêtre *Praat Editor.* Le logiciel renvoie alors, entre autres données, le nombre de portions non voisées dans l'intervalle sélectionné par rapport au nombre total de portions détectées, qui lui, dépend de la fréquence d'échantillonnage de la fréquence fondamentale (f0). Par contre, comme le souligne Eager (2015), la fraction de portions non voisées renvoyée par *Praat* est appelée à changer grandement en fonction du degré d'agrandissement ou du positionnement de la fenêtre d'analyse dans le signal. Il est théoriquement possible de contrer ce problème de fiabilité en augmentant le nombre de portions détectées, c'est-à-dire en réduisant l'intervalle temporel entre deux points d'échantillonnage de la f0, mais cette stratégie cause un sérieux ralentissement de toute opération dans *Praat Editor*. Eager (2015) suggère ainsi d'éviter cette fenêtre et recommande plutôt d'extraire, à partir de *Praat Objects*, des « objets » *Pitch* (bouton *Analyse periodicity*). Ces objets renseignent sur l'évolution temporelle de la périodicité au sein du fichier sonore à partir duquel ils sont créés. Divers paramètres, au moment de leur création, sont à la discrétion de l'analyste. Il est possible, par exemple, de définir les fréquences plancher et plafond de détection de la f0, mais surtout, l'intervalle temporel entre deux points d'échantillonnage. Si l'analyste choisit de le réduire sensiblement, la précision de la détection de la f0 s'en trouve grandement améliorée sans que les inconvénients susmentionnés ne surviennent. Suivant les recommandations de Eager (2015), qui a notamment mis en évidence la fiabilité de cette méthode par rapport à une analyse manuelle et l'absence de biais inhérent à surestimer ou à sous-estimer la quantité de voisement d'un segment, nous avons analysé le taux de voisement des fricatives au moyen d'objets *Pitch* où l'intervalle temporel entre deux points d'échantillonnage était de 0,001 s (soit une prise de mesure toutes les millisecondes). Pour les locutrices, les fréquences plancher et plafond de détection de la f0 ont été établies à 70 Hz et 300 Hz, et pour les locuteurs, à 70 Hz et 250 Hz.

3 Résultats

La Figure 1 illustre les résultats relatifs à la durée vocalique. Les six graphiques de la partie supérieure, qui représentent les voyelles intrinsèquement longues, montrent que leur durée est moins influencée par le voisement attendu de la consonne suivante que les voyelles intrinsèquement courtes, représentées dans les sept graphiques de la partie inférieure. Ces dernières subissent effectivement un allongement marqué lorsqu'elles sont suivies des consonnes voisées /v z ʒ/ (en bleu), leur durée

devenant alors similaire à celle des voyelles longues. Que l'allongement soit intrinsèque ou contextuel, il semble mener à davantage d'hétérogénéité dans les productions des participants, comme en témoigne la distribution plus étendue des voyelles longues et allongées.

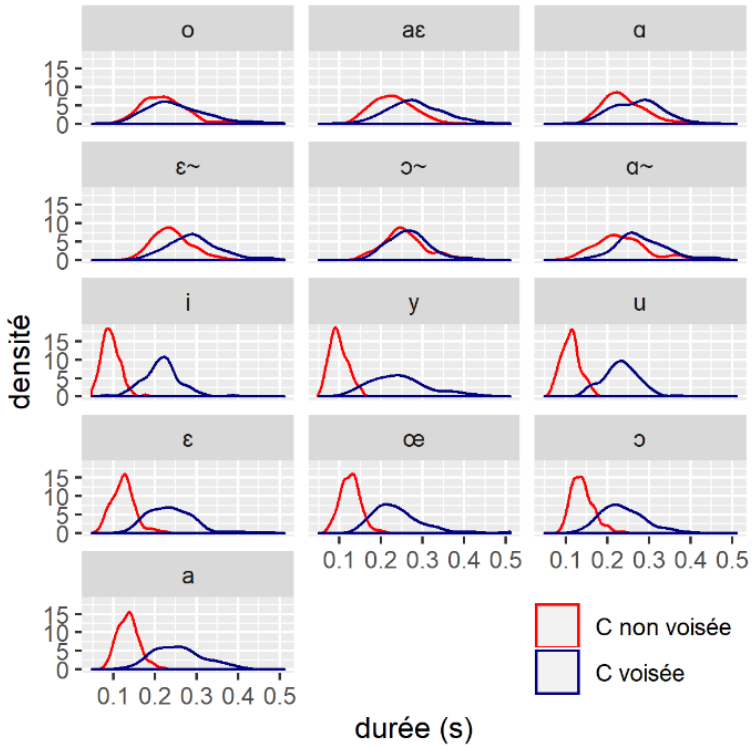


FIGURE 1 : Distribution de la durée vocalique en fonction du voisement attendu de la consonne suivante

L'analyse statistique des données a été effectuée à l'aide d'un modèle de régression linéaire à effets mixtes (environnement R, bibliothèques *lme4* et *emmeans*; R Core Team, 2020; Bates *et al.*, 2015; Lenth *et al.*, 2018). Ce modèle permet d'explorer la relation entre la durée vocalique et deux effets fixes (ou variables indépendantes) : le voisement consonantique attendu (consonnes voisées vs non voisées) et la longueur intrinsèque (voyelles courtes vs longues). Deux effets aléatoires ont également été pris en compte : les locuteurs et les tâches de lecture. Puisqu'un problème d'hétéroscédasticité a été détecté lors de la vérification des postulats sur lesquels sont fondés les modèles de ce type, une transformation logarithmique a été opérée sur la distribution de la variable dépendante. Les résultats font état d'une interaction significative entre les deux effets fixes. Les comparaisons multiples post-hoc montrent, sans surprise, une différence significative entre les voyelles courtes suivies des consonnes non voisées et voisées (β=0,703, e.s.=0,009, t(2655)=73,244, p<0,0001). Cet allongement des voyelles courtes en contexte voisé est si important qu'elles ne se distinguent plus des voyelles intrinsèquement longues en contexte non voisé (β=0,006, e.s.=0,009, t(2655)=0,628, p=0,9232), mais des voyelles longues en contexte voisé, si (β=-0,146, e.s.=0,01, t(2654)=-14.596, p<0,0001). C'est donc dire que les voyelles intrinsèquement longues subissent un allongement significatif au contact des consonnes voisées (β=0,152, e.s.=0,01, t(2653)=15,123, p<0,0001).

La Figure 2 illustre de deux manières la durée moyenne des consonnes fricatives. Le graphique de gauche permet de voir que les consonnes non voisées /f s ʃ/ ont une durée plus longue que les consonnes voisées /v z ʒ/. À droite, on observe que les consonnes non voisées représentent une plus grande part des séquences VC que les voyelles, que celles-ci soient longues ou courtes. Cette relation est toutefois inversée lorsque les consonnes sont voisées : les voyelles représentent alors une plus

grande portion de la séquence VC. L'analyse statistique de la durée consonantique, pour laquelle un modèle mixte identique à celui décrit précédemment a été ajusté, révèle à nouveau une interaction significative entre les deux effets fixes (voisement consonantique attendu et longueur intrinsèque). D'abord, peu importe la longueur intrinsèque de la voyelle précédente, les consonnes non voisées sont toujours significativement plus longues que les consonnes voisées. En ce qui concerne /v z ʒ/, leur durée ne varie pas en fonction de la voyelle précédente (β=0,002, e.s.=0,001 t(2654)=1,323, p=0,5483). Par contre, les consonnes /f s ʃ/ sont significativement plus longues lorsqu'elles sont précédées de voyelles courtes que longues (β=0,031, e.s.=0,001, t(2653)=16,817, p<0,0001), leur poids au sein de la séquence VC s'en trouvant d'autant plus augmenté.

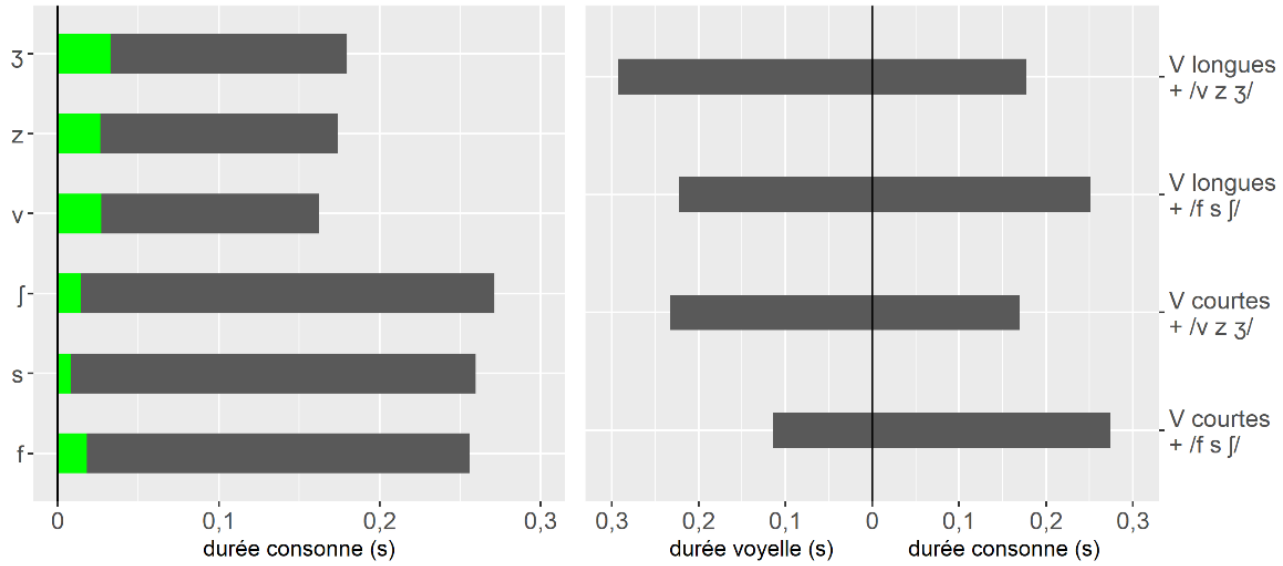


FIGURE 2 : À gauche, durée moyenne de chacune des consonnes, où la portion verte correspond à la durée moyenne du voisement relevé. À droite, durée moyenne des voyelles longues ou courtes et des consonnes voisées ou non voisées dans les séquences VC

Le voisement consonantique réel est quant à lui illustré aux Figures 2, 3 et 4. Dans le graphique de gauche de la Figure 2, on observe qu'une certaine portion des consonnes /f s ʃ/ est voisée, vraisemblablement une conséquence de la coarticulation avec les voyelles précédentes. Visuellement, il semble que la durée moyenne du voisement dans les consonnes /v z ʒ/ soit également plutôt réduite, bien qu'elle représente une proportion plus importante de la durée consonantique totale (17%, 15% et 18% respectivement). La Figure 3 ajoute un complément d'information en montrant la distribution du taux de voisement consonantique pour chacune des consonnes, au-delà de la seule moyenne. La série de graphiques, où le voisement est cette fois présenté sous forme de proportion plutôt que de durée absolue, révèle qu'une grande majorité des consonnes /f s ʃ/ est voisée sur moins de 25% de leur durée, voire moins dans le cas de /s/. Bien que la distribution du taux de voisement de /v z ʒ/ tire aussi vers la gauche, les valeurs sont davantage réparties sur l'axe des abscisses. Enfin, les graphiques de la Figure 4 suggèrent que les consonnes tendent davantage à être non voisées lorsqu'elles sont précédées de voyelles longues (courbes noires) que courtes (courbes discontinues rouges).

L'analyse statistique du taux de voisement vocalique a été effectuée en deux temps. D'abord, une régression logistique à effets mixtes a été effectuée sur deux valeurs de voisement : 0 représentant les consonnes voisées sur 0% de leur durée et 1 représentant toutes les autres. Cette première analyse révèle un effet significatif des deux effets fixes (sans interaction). Ainsi, le groupe de consonnes /f s ʃ/ comporte significativement plus d'occurrences voisées sur 0% de leur durée que le groupe /v z ʒ/ (β=0,290, e.s.=0,016, t(2660)=17,185, p<0,0001). Les résultats confirment également que les

consonnes sont plus fréquemment voisées sur 0% de leur durée lorsqu'elles suivent des voyelles intrinsèquement longues (β=0,122, e.s.=0,016, t(2656)=7,257, p<0,0001). Dans un second temps, une régression linéaire à effets mixtes sur une transformation logarithmique de toutes les valeurs autres que 0% a été effectuée. Une interaction significative entre les effets fixes ressort à nouveau. Quelle que soit la durée intrinsèque de la voyelle précédente, /v z ʒ/ présentent toujours un taux de voisement significativement plus élevé que /f s ʃ/. Le taux de voisement de ces dernières n'est pas influencé par la durée intrinsèque de la voyelle précédente (β=0,108, e.s.=0,063, t(1215)=1,712, p=0,3177). À l'inverse, conformément à ce que l'on observe dans le graphique de droite de la Figure 4, /v z ʒ/ possèdent un taux de voisement significativement plus élevé lorsqu'elles suivent des voyelles courtes que longues (β=0,353, e.s.=0,046, t(1211)=7,539, p<0,0001).

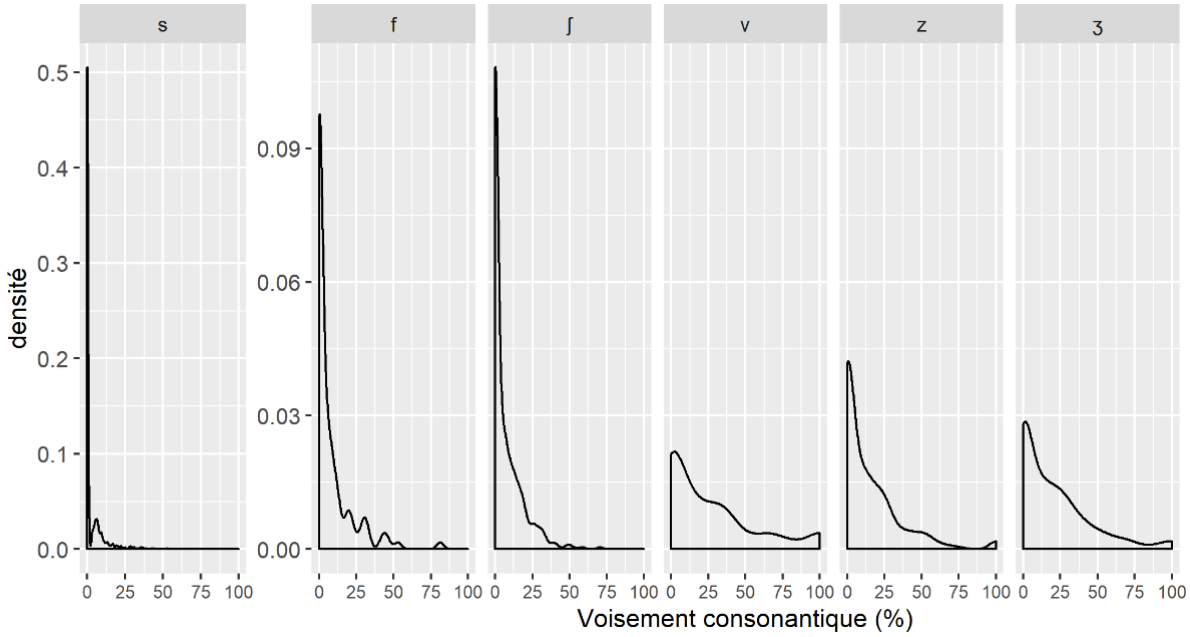


FIGURE 3 : Distribution de la proportion de voisement mesuré dans chaque consonne en fonction de la durée normalisée. L'échelle de l'axe des ordonnées varie entre /s/ et les autres consonnes

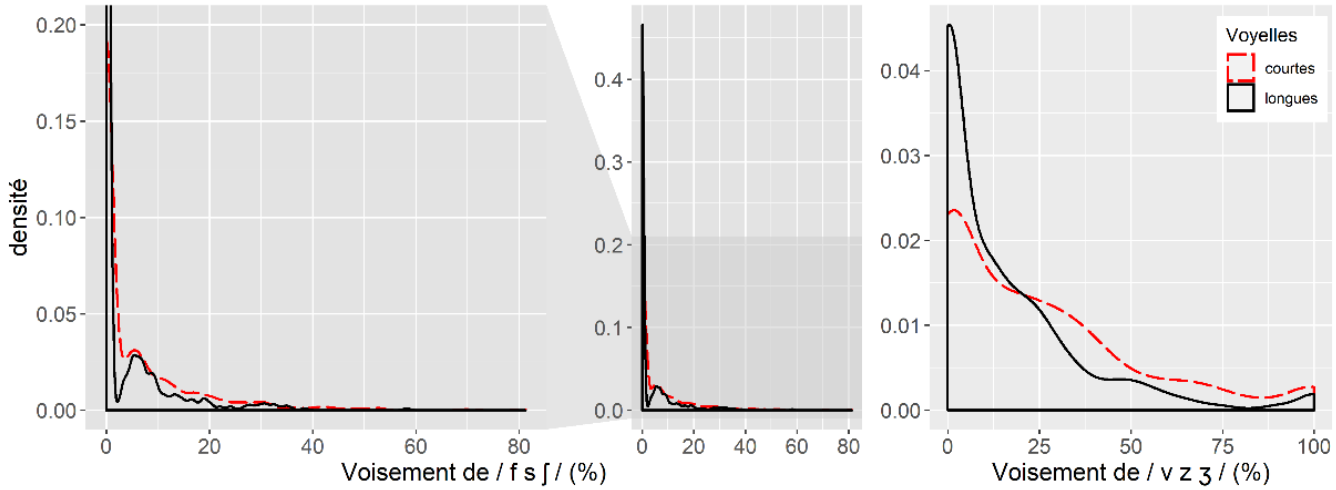


FIGURE 4 : Distribution de la proportion de voisement mesuré dans les consonnes non voisées (gauche) et voisées (droite) selon la longueur intrinsèque de la voyelle précédente. Le graphique de gauche est un agrandissement de la partie inférieure du graphique du centre

4 Discussion et conclusion

Rappelons que l'objectif principal de cette contribution était de proposer une description des consonnes fricatives finales voisées et non voisées du FQ. Notamment, nous cherchions à confirmer certains résultats précédemment obtenus par Jacques (1990) en utilisant un plus grand corpus (55 locuteurs comparativement à 4), en exploitant des techniques d'analyse plus modernes, en particulier une méthode de détection automatisée et précise du voisement (Eager, 2015), et en tenant compte de la durée intrinsèque des voyelles accentuées précédant les fricatives. En résumé, les résultats obtenus indiquent que les fricatives non voisées /f s ʃ/ possèdent une durée plus longue et qui représente une proportion plus importante des séquences VC que /v z ʒ/. Le faible taux de voisement parfois détecté au cours de /f s ʃ/ constitue une part très réduite de leur durée totale et résulte probablement de la coarticulation avec les voyelles précédentes. Dans ce contexte, la durée desdites voyelles semble essentiellement dépendre de leurs propriétés intrinsèques, confirmant de ce fait la vitalité du phénomène en FQ contemporain. Pour ce qui est des fricatives voisées /v z ʒ/, leur durée est moins importante que celle de /f s ʃ/ d'une part, et que celle des voyelles dans les séquences CV d'autre part. Si un voisement total est plus marginal qu'un dévoisement total, il demeure que /v z ʒ/ possèdent un taux de voisement supérieur à celui de /f s ʃ/ et une tendance moindre à être entièrement dévoisés. Leur effet sur la durée des voyelles précédentes est sans équivoque : elles allongent sensiblement les voyelles intrinsèquement courtes, mais également les voyelles déjà longues, qui voient leur durée significativement accrue à leur contact. Lorsque tel est le cas, le taux de voisement de /v z ʒ/ diminue. Il est probable que la production d'une voyelle à la durée ainsi maximisée rende physiologiquement plus difficile le maintien du voisement dans la consonne suivante (Ohala, 1983). Au final, les consonnes /f s ʃ/ se distinguent en tous points de /v z ʒ/.

Si nous comparons à présent nos résultats avec ceux de Jacques (1990), celui-ci rapportait qu'en finale absolue, /z ʒ/ étaient voisés sur un tiers de leur durée et /v/, sur deux tiers. La Figure 2 révèle des taux moyens de voisement nettement inférieurs et peu d'écart entre les trois consonnes (/v/ : 17%; /z/ : 15% ; /ʒ/ : 18%). Il n'est pas impossible que les usages des locuteurs du FQ aient changé au cours du temps, mais il est également probable que les instruments de mesure utilisés soient en cause. Jacques (1990) rapportait ensuite que les fricatives finales voisées étaient plus courtes que leurs homologues non voisées. Une tendance similaire se dégage de nos résultats, la Figure 2 laissant également entrevoir une gradation parallèle, entre les voisées et les non-voisées, de la durée consonantique en fonction du lieu d'articulation. Finalement, Jacques (1990) suggérait de tenir compte de la durée intrinsèque des voyelles dans une future étude sur l'effet allongeant de /v z ʒ/ en raison de résultats plus mitigés qu'attendu. Nos résultats confirment l'intuition de l'auteur, en plus d'avoir révélé la robustesse de l'effet de voisement et mis au jour le fait que les voyelles intrinsèquement longues du FQ sont, dans une certaine mesure, encore extensibles.

Cette étude préliminaire gagnera bien entendu à être enrichie. Trois indices acoustiques ont été examinés à ce jour, mais l'analyse des consonnes fricatives peut également reposer sur des variables telles que les moments spectraux ou les coefficients de transformées en cosinus discrètes (DCT). Tenir compte du potentiel effet abrégeant des consonnes non voisées en analysant en outre des voyelles produites en syllabe ouverte pourrait également permettre de mieux prendre la mesure de l'effet allongeant des consonnes voisées. Intégrer à l'analyse d'autres catégories de consonnes, en particulier les occlusives, permettrait par ailleurs une description plus générale de l'effet du voisement consonantique sur la durée des voyelles intrinsèquement courtes et longues du FQ. Cela dit, les résultats présentés confirment une fois de plus que le voisement n'est binaire qu'en perception. En production, les locuteurs utilisent un ensemble de stratégies concomitantes dont le poids perceptif relatif en FQ demeure à explorer.

Remerciements

Les données présentées ont été récoltées dans le cadre d'un projet financé par le Conseil de recherches en sciences humaines du Canada (CRSH) et supervisé par Johanna-Pascale Roy (Université Laval). Merci aux deux évaluateurs anonymes pour leurs questions et suggestions, ainsi qu'à Michele Gubian et Stefano Coretta (IPS-LMU) pour les discussions enrichissantes.

Références

BATES D., MAECHLER M., BOLKER B. & WALKER S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

BOERSMA P. & WEENINK D. (2020). Praat: Doing Phonetics by Computer (version 6.1.x). *Logiciel*, https://www.fon.hum.uva.nl/praat/.

DENES P. B. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27(4), 761-764.

EAGER C. D. (2015). Automated voicing in Praat: Statistically equivalent to manual segmentation. In THE SCOTTISH CONSORTIUM FOR ICPHS 2015, Éds., *Proceedings of ICPhS XVIII*, Glasgow.

FOUCHÉ P. (1959). *Traité de prononciation française*, 2e éd. Éditions Klincksieck : Paris.

FOUGERON C. & SMITH C. L. (1993). French. *Journal of the International Phonetic Association*, 23(2), 73-76.

HOGAN J. T. & ROZSYPAL A. J. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America*, 67(5), 1764-1771.

HOUSE A. S. & FAIRBANKS G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25(1), 105-113.

JACQUES B. (1990). Étude de trois indices acoustiques du voisement des consonnes fricatives en français de Montréal. *Revue québécoise de linguistique*, 19(2), 59-71.

JATTEAU A., VASILESCU I., LAMEL L., ADDA-DECKER M. & AUDIBERT N. (2019). "Gra[f]e!" Word-final devoicing of obstruents in Standard French: An acoustic study based on large corpora. *Proceedings of Interspeech 2019*, Graz, 1726-1730.

LENTH R., LOVE J. & HERVE M. (2018). emmeans: Estimated marginal means, aka least-squares means (version 1.1.2). *Logiciel*, https://CRAN.R-project.org/package=emmeans.

LISKER L. & ABRAMSON A. S. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, 20(3), 384-422.

MARTIN P. (1996). *Éléments de phonétique avec application au français*. PUL : Ste-Foy.

OHALA J. J. (1983). The origin of sound patterns in vocal tract constraints. In P. F. MACNEILAGE, Éd., *The Production of Speech*, 189-216. Springer : New York.

PAPE D. & JESUS L. M. (2015). Stop and fricative devoicing in European Portuguese, Italian and German. *Language and Speech*, 58(2), 224-246.

R CORE TEAM. (2020). R: A language and environment for statistical computing. *Logiciel*, https://www.R-project.org.

RAPHAEL L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America*, 51(4), 1296-1303.

WALTER H. (1982). *Enquête phonologique et variétés régionales du français*. PUF : Paris.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Imprécision dans la production des voyelles : un potentiel marqueur infraclinique dans la maladie de Parkinson

Virginie Roland[1], Véronique Delvaux[1, 2], Kathy Huet[1], Myriam Piccaluga[1] et Bernard Harmegnies[1]
(1) Institut de Recherches en Sciences et Technologies du Langage ; Service de Métrologie et Sciences du Langage, 18 place du Parc, 7000 Mons, Belgique
(2) FNRS, Belgique
Virginie.roland@umons.ac.be

RÉSUMÉ

La maladie de Parkinson est une maladie neurodégénérative qui affecte le système neuro-moteur. Une grande variété de troubles de la parole, généralement regroupés sous les termes de dysarthrie hypokinétique, peuvent apparaitre. Dans cette contribution, nous présentons les résultats d'une étude acoustique comparative de la production de voyelles par 63 locuteurs MP dysarthriques et non dysarthriques, avec 35 locuteurs sains. Notre objectif est d'étudier la production de voyelles isolées afin de répondre à la question : l'imprécision dans la production des voyelles peut-elle être considérée comme un marqueur infraclinique de la dysarthrie ?

ABSTRACT

Imprecision of vowel production: a potential subclinical marker in Parkinson's disease. Parkinson's disease is a neurodegenerative disease affecting neuromotor system. A wide variety of speech disorders usually regrouped under the label of hypokinetic dysarthria may be appear. In this paper, we present the results of an acoustic study of vowel' production by dysarthric PD, non dysarthric PD and control speakers. Our aim is to study the production of vowel in isolation to address the following question: can imprecision of vowel's production be considered as an infraclinic marker of dysarthria?

MOTS-CLÉS : voyelles, maladie de Parkinson, métriques acoustiques, marqueur infraclinique
KEYWORDS: vowels, Parkinson's disease, acoustic metrics, subclinical marker

1 Introduction

Notre recherche s'intègre dans un questionnement plus large qui s'intéresse au potentiel informatif du signal de parole produit par des locuteurs s'exprimant dans des situations de communication adverses, au rang desquelles figurent les états à caractère pathologique. La présente contribution se

Volume 1 : Journées d'Études sur la Parole, pages 561–569.
HAL : hal-02798587.

centre plus particulièrement sur la parole produite par des locuteurs atteints de la maladie de Parkinson (MP), une affection neurodégénérative présentant une large symptomatologie, au sein de laquelle figurent des troubles dysarthriques. Cette dysarthrie, dite *hypokinétique* dans la MP, a longtemps été considérée comme d'apparition tardive, en particulier en ce qui concerne ses répercussions sur l'intelligibilité de la parole, et plus largement sur la communication. Toutefois, ce caractère tardif n'a pas toujours fait consensus et est, aujourd'hui, en partie réfuté. Ainsi, les réflexions menées sur l'évolution de la maladie (Braak et al., 2003) ainsi que celles menées par l'*International Parkinson and Movement Disorder Society* suggèrent que les troubles de la parole pourraient au contraire faire partie des premières manifestations de la MP (Sapir, 2014). Notre objectif est ainsi d'identifier des mesures acoustiques sensibles aux altérations de parole présentes dans la MP. En effet, puisque la dysarthrie semble pouvoir apparaître tôt dans le décours de la maladie, il nous semble opportun de tenter d'identifier des éléments permettant sa détection aussi précocement que possible, dans une perspective de prise en charge, voire de diagnostic précoces.

2 Méthodologie

2.1 Locuteurs et corpus

La participation de 98 locuteurs francophones, répartis en deux groupes, a été sollicitée. Le premier groupe se compose de 63 participants (36 hommes et 27 femmes) atteints de la MP, âgés entre 38 et 85 ans (m = 70 ans). Le diagnostic de la maladie a été posé au minimum un an avant le recrutement jusqu'à une durée de maladie de 25 ans, la durée moyenne de la MP étant de 7 ans. Les participants présentent un éventail complet de l'ensemble des stades de la maladie tel que proposé par Hoehn et Yahr (1967). À la suite de l'évaluation à partir de l'item parole de l'échelle d'évaluation UPDRS, 20 de nos participants ont été identifiés comme ne présentent pas de troubles dysarthriques (11 hommes et 9 femmes – proportions similaires à celles du groupe complet). Le second groupe (groupe témoin) se compose quant à lui de 35 participants, 19 hommes et 16 femmes (proportions similaires à celles du groupe MP), âgés entre 41 et 84 ans (m = 66 ans), ne présentant pas de pathologies ni d'antécédents médicaux pouvant altérer la production de la parole.

Les participants ont été amenés à réaliser différentes tâches de production de parole, dont la production isolée et brève des phonèmes /a, i, u/ (5 productions). Ces productions de parole ont toutes été enregistrées en une seule passation pour chaque sujet, à l'aide d'un enregistreur portatif ZOOM H5. Tous les participants atteints de la MP ont été enregistrés dans une pièce calme au sein de l'hôpital où ils étaient suivis, à l'occasion de visites de contrôle. Tous les participants ayant été enregistrés dans la même pièce, la même disposition du matériel de présentation et d'enregistrement a été instaurée. Les participants du groupe témoin ont quant à eux été enregistrés à leur domicile. Les sons de parole recueillis sont le résultat de demandes de production à partir de versions écrites des productions à réaliser présentées sur un écran d'ordinateur portable placé face aux participants, le support informatique nous permettant de proposer des instructions identiques à chacun des participants ainsi que d'ajuster la taille de l'écriture aux besoins des participants.

2.2 Analyses acoustiques

Les échantillons de parole recueillis auprès de chaque participant ont été analysés à l'aide du logiciel PRAAT. Les valeurs formantiques F1 et F2 ont été récoltées manuellement au moyen du tracking formantique et de l'examen spectrographique fournis par le logiciel. Ces valeurs ont été extraites au centre de la production. Afin d'apprécier l'étendue de l'espace vocalique et de suggérer des hypothèses quant à la possible configuration du tractus vocal, nous avons considéré chaque couple de mesures F1, F2 comme un objet dans le plan F1/F2. Ces valeurs nous ont permis de calculer différentes métriques témoignant de l'exploitation du champ vocalique : l'aire du triangle vocalique (tVSA), l'indice d'articulation des voyelles (VAI) (Roy, Nissen, Dromey, & Sapir, 2009) ainsi que l'indice d'organisation du système vocalique (Phi) et les deux métriques le composant, le CMinter et le CMintra (Huet & Harmegnies, 2000). Ces métriques nous semblent complémentaires grâce aux informations qu'elles délivrent (étendue vocalique, centralisation, variabilité intra- et inter-catégories vocaliques). Nous avons ainsi opté pour une diversification des indices afin d'apprécier l'étendue de l'espace vocalique tout en prenant en compte le possible manque de sensibilité évoqué par Neel (2008) ainsi que par Skodda, Visser, et Schlegel (2011) en ce qui concerne l'exploitation de l'aire du triangle vocalique auprès de locuteurs faiblement atteints sur le plan articulatoire. En effet, le VAI a notamment été développé afin de compléter les recherches sur l'altération des voyelles, en particulier dans la MP où le calcul de la métrique VSA peut être insensible aux formes légères à modérées de dysarthrie. Skodda et al. (2011) ont ainsi montré son potentiel informatif, jugé supérieur à celui du VSA, dans la détection de changements subtils lors de la production de voyelles. Enfin, la multiplicité de ces métriques pourrait nous permettre de mieux cerner les variations dans l'exploitation du système vocalique : diminution de l'espace acoustique, instabilité articulatoire suggérée par une variabilité intra-catégorielle importante, diminution des contrastes vocaliques découlant d'une centralisation des cibles et/ou d'une distance inter-catégorielle moindre.

3 Résultats

3.1 Espace vocalique

Nous avons tout d'abord questionné l'exploitation du champ vocalique des sujets en condition de production des trois voyelles brèves /a, i, u/ en calculant la surface des triangles vocaliques, sujet par sujet. Les valeurs de surface apparaissent comme significativement ([1]U = 1400 ; p = 0.027) plus importantes dans le groupe témoin (m = 363679 Hz²) que dans le groupe MP (m = 306501 Hz²). Ceci traduit donc une réduction de l'exploitation de l'espace vocalique chez les locuteurs atteints de la MP, suggérant une possible réduction de l'espace articulatoire.

Afin d'apprécier plus finement les variations apparaissant dans l'espace vocalique et de tester les hypothèses énoncées, la métrique de centralisation VAI a été calculée pour chacune des 5 productions. Ces calculs nous ont ensuite permis d'obtenir un indice VAI moyen par sujet. Pour rappel, la métrique VAI permet de mettre en évidence un phénomène de centralisation. Plus la valeur

[1] U de Mann-Whitney

obtenue est petite, plus la centralisation des voyelles est importante. Les valeurs obtenues témoignent d'une centralisation significativement plus marquée chez les locuteurs atteints de la MP ($U = 1519$, $p = 0.001$). Ceci suggère donc une possible réduction des contrastes vocaliques suite à une centralisation plus marquée des productions en ce qui concerne les locuteurs atteints de la MP.

Pour donner suite aux constats et aux hypothèses découlant des valeurs des métriques précédentes, nous avons exploité la métrique Phi ainsi que les deux métriques permettant son calcul, le CMintra et le CMinter. Pour rappel, l'indice Phi permet d'étudier le degré d'organisation du système vocalique à partir du rapport entre la dispersion inter-catégories vocaliques et la dispersion intra-catégorielle. Plus la valeur de l'indice Phi est élevée, plus le système vocalique apparait comme organisé (Huet & Harmegnies, 2000). L'indice Phi a été calculé pour chaque locuteur de nos deux groupes. Nous avons ensuite effectué la moyenne dans chaque groupe de locuteurs afin de les comparer. Les locuteurs du groupe MP se caractérisent, en moyenne, par un indice Phi significativement plus faible ($m = 150$) que les locuteurs du groupe témoin ($m = 1477$) ($U = 1960$, $p < 0.001$). La figure 1 illustre parfaitement ces différences de moyenne entre les deux groupes.

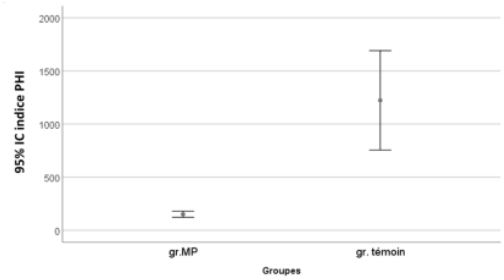


Figure 1 : valeurs moyennes et intervalle de confiance (à 95%) de l'indice Phi en fonction du groupe de sujets (intervalle de confiance à 95%)

De même, la dispersion intra-catégorielle (CMintra) est beaucoup plus importante chez les locuteurs atteints de la MP ($U = 278$, $p < 0.001$) que chez les locuteurs du groupe témoin (respectivement, CMintra $m = 31125$ et CMintra $m = 8751$). Cela laisse apparaitre une instabilité acoustique pour une même catégorie vocalique, suggérant une certaine instabilité articulatoire.

3.2 Regards cliniques

Cinq productions ont été produites par les participants. Afin d'apprécier plus finement les variations apparaissant dans l'espace vocalique, nous avons observé séparément chacune de ces productions au regard des indices précédemment décrits. En étudiant séparément les valeurs des aires des triangles vocaliques, nous avons ainsi constaté que, contrairement aux autres productions, la première production des phonèmes /a, i, u/ présente des caractéristiques similaires dans les deux groupes. Ainsi, la surface des triangles vocaliques calculée pour les premières productions de chaque sujet ne permet pas de mettre en évidence des différences significatives entre les deux groupes. Nous remarquons également une variabilité plus marquée dans le groupe MP lors de cette première production. Les quatre autres répétitions sont quant à elles significativement différentes entre les

deux groupes (respectivement, U = 1391, p = 0.032 ; U = 1368, p = 0.049 ; U = 1458, p = 0.008 ; U = 1485, p = 0.005).

Le même constat est mis en évidence à partir de la métrique de centralisation VAI : les valeurs du VAI, obtenues à partir des données extraites des premières productions de chaque sujet, ne permettent pas de mettre en évidence des différences significatives entre nos deux groupes de participants, contrairement aux valeurs obtenues à partir des quatre autres répétitions où des différences significatives apparaissent entre les deux groupes (respectivement, U = 1459, p = 0.008 ; U = 1526, p = 0.002 ; U = 1545, p = 0.001 ; U = 1564, p = 0.001).

La première production des locuteurs atteints de la MP se différencie donc moins de celle du groupe témoin, ce qui laisse à penser qu'un effort est réalisé en début de tâche afin de maintenir une production satisfaisante, le maintien de cet effort se dégradant au fil des productions. Un effet de fatigue pourrait également apparaitre à la suite du maintien des articulateurs dans des positions particulières, entrainant une imprécision plus marquée des productions au cours des répétitions, traduit par une centralisation des cibles vocaliques.

Nous pouvons également identifier qu'un de nos participants, le locuteur MP47, présente des indices de centralisation très marqués (VAI_{moyen} = 0.60)[2]. Il s'agit d'une femme, âgée de 69 ans et atteinte de la MP depuis 18 ans (diagnostic à 51 ans). Cette locutrice est au stade 5 selon la classification de Hoehn et Yahr (1967). C'est également la participante la plus sévèrement atteinte sur le plan de la dysarthrie hypokinétique si nous nous basons sur les scores obtenus à l'item « parole » de l'UPDRS III. Ceci pourrait suggérer un accroissement de la centralisation au cours de l'évolution de la MP.

Enfin, l'utilisation de la métrique PHI met en exergue des résultats attestant des différences d'organisation du système vocalique en fonction du groupe de sujets considérés, les locuteurs atteints de la MP présentant un degré d'organisation beaucoup plus faible que les sujets du groupe témoin. A titre illustratif, la figure 2 présente la dispersion dans l'espace acoustique des productions (5 répétitions /i, a, u/) du participant MP se caractérisant par la valeur d'indice Phi la plus basse (Phi = 43) et du participant du groupe témoin se caractérisant par une valeur d'indice Phi la plus haute (Phi = 5689).

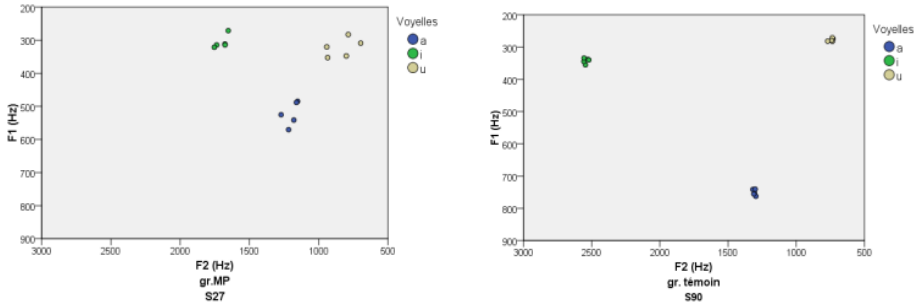


Figure 2 : Exploitation de l'espace acoustique (5 productions /i, a, u/) par un locuteur du gr. MP (à gauche) et par un locuteur du gr. témoin (à droite)

[2] Nous avons mené les mêmes analyses statistiques en neutralisant les données de ce locuteur afin de vérifier les différences entre nos deux groupes de participants. Les différences observées entre les locuteurs atteints de la MP et les locuteurs témoins demeurent significatives.

La figure 2 met en évidence une forte centralisation formantique chez le locuteur atteint de la MP, entrainant une différenciation inter-catégorielle plus faible que chez le locuteur du groupe témoin (CMinter m = 3017007). Les valeurs inter-catégorielles sont en effet plus importantes au sein du groupe témoin (CMinter m = 3716301 ; U = 1511, p = 0.001), témoignant d'une meilleure exploitation de l'espace vocalique.

3.3 Identification de phénomènes infracliniques

Lors d'une précédente étude (Roland *et al.*, 2016), nous avons mis au jour des phénomènes de nature infraclinique, les constats se faisant à partir de productions de parole recueillies auprès de locuteurs ne présentant que peu ou pas d'atteinte articulatoire. Ce faisant, nous avons souhaité interroger le caractère potentiellement informatif des productions recueillies dans cette deuxième étude auprès de nos participants ne présentant pas de troubles dysarthriques (pour rappel, N = 20). Pour ce faire, nous avons comparé leurs productions de voyelles en situation isolée et stable avec celles de notre groupe témoin.

Un bref examen descriptif permet à nouveau de constater une centralisation plus marquée chez les locuteurs atteints de la MP (Fig. 3), en particulier lors de la production du phonème /a/, suggérant une moins grande aperture.

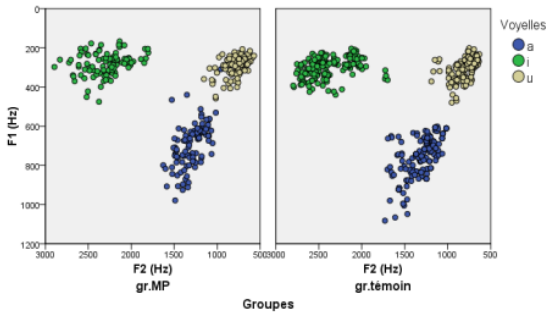


Figure 3 : dispersion des voyelles en isolation dans le plan F1/F2 (Hz) pour le groupe MP non dysarthrique (gauche) et témoin (droite).

Le calcul des différentes métriques à l'étude nous permet d'identifier des différences significatives dans le degré d'organisation du système vocalique - indice Phi (U = 639, p < 0.001), en particulier en ce qui concerne la dispersion intra-catégorielle (CMintra) (U = 86, p < 0.001).

Nous pouvons ainsi remarquer que le système vocalique apparait comme moins bien organisé chez les locuteurs atteints de la MP, la distance intra-catégorielle étant plus importante dans ce groupe. Nous constatons également que cette distance intra-catégorielle est plus variable chez les participants atteints de la MP ne présentant pas de dysarthrie que chez les locuteurs du groupe témoin. Les autres métriques ne mettent pas en évidence de différences significatives entre nos deux groupes de locuteurs, ce qui nous amène à considérer que le calcul de l'indice Phi permet de détecter des changements subtils lors de la production de voyelles par des locuteurs qui ne sont pas considérés comme atteints sur le plan articulatoire (Fig. 4).

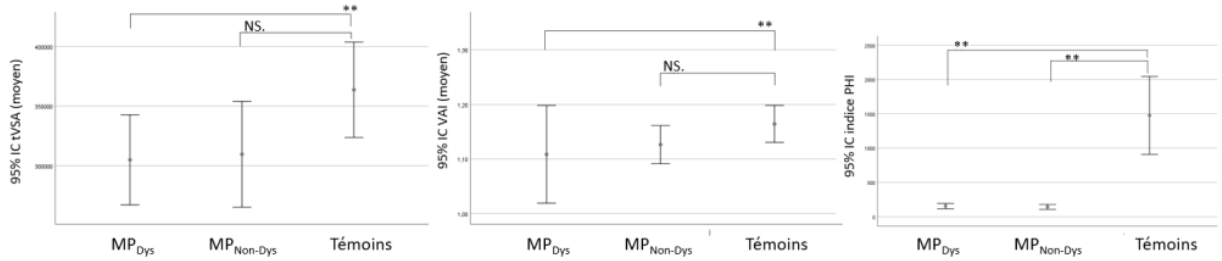


Figure 4 : valeurs moyennes et intervalle de confiance (à 95%) des métriques tVSA, VAI et Phi en fonction des groupes MP (dysarthriques et non-dysarthriques) et témoin.

4 Discussion

Cette étude a été motivée par les constats résultant d'une précédente étude menée auprès de participants atteints de la MP et faiblement atteints sur le plan articulatoire (absence de dysarthrie hypokinétique). En ce qui concerne l'étendue vocalique, nous avons voulu tester l'hypothèse selon laquelle une réduction de l'espace vocalique apparait chez les locuteurs atteints de la MP comparativement à des locuteurs sains tel qu'évoqué par de nombreux auteurs (par exemple, Mollaei, Shiller, Baum, & Gracco, 2016 ; Skodda et *al.*, 2011). L'étude de l'exploitation de l'espace vocalique suggère une réduction de celui-ci chez les locuteurs atteints de la MP, l'étendue des surfaces exploitées par ces derniers étant statistiquement plus faible que celle des locuteurs témoin. Profitant du nombre de productions par locuteur, nous avons constaté que les différences se font jour à partir des comparaisons de la deuxième à la cinquième répétition des monophtongues. Cela nous a amenés à constater que, même si en moyenne l'étendue vocalique est réduite chez les locuteurs atteints de la MP, elle l'est particulièrement en cours de répétition, ce qui nous amène à nous questionner sur les raisons de ces différences entre la première production et celles qui suivent. Nous évoquions précédemment un possible effort pour maintenir une production satisfaisante, avec une dégradation de ce maintien au cours des répétitions. Cela pourrait être rapproché des stratégies d'hyper-articulation évoquées par Lindblom (1990), stratégies devant permettre la production d'une parole claire et précise. En effet, Lam et Tjaden (2016) évoquent à ce sujet un ajustement acoustique constant en fonction de la situation de communication. Etant donné la situation d'enregistrement réalisée dans le cadre de cette étude, les participants ont pu tenter d'ajuster leurs productions afin de les rendre les plus précises possibles. Dès lors, les différences apparaissant entre les autres répétitions pourraient traduire une difficulté à maintenir sur un terme plus long les articulateurs dans des positions bien précises, entrainant ainsi une imprécision et une réduction de l'espace articulatoire. Cela correspondrait également à une situation de retour à l'état initial de leurs capacités, un phénomène d'hypoarticulation étant présent dans la dysarthrie hypokinétique. Ces hypothèses seraient toutefois à vérifier par le recours à des analyses articulatoires en complément des analyses acoustiques actuellement menées.

Afin d'affiner ces observations, nous avons étudié les caractéristiques de l'espace vocalique des locuteurs à la lumière de plusieurs métriques acoustiques, à l'instar de ce qui a été réalisé par Audibert et Fougeron (2012) à travers une comparaison inter-dysarthrique. Cette démarche, originale en ce sens que nous n'avons pas identifié d'autres études combinant plusieurs métriques acoustiques auprès d'un groupe de participants atteints de la MP présentant différents degrés de

dysarthrie hypokinétique (dont l'absence de dysarthrie hypokinétique), nous a permis d'observer les distorsions de l'espace vocalique en tenant compte de différents aspects (étendue, variabilité intra- et inter-catégorielle, centralisation). Qui plus est, certains auteurs (Neel, 2008; Skodda *et al.*, 2011) ont évoqué le manque de sensibilité de la métrique tVSA chez des locuteurs présentant la MP et faiblement atteints sur le plan articulatoire. Nous nous sommes dès lors également intéressés à l'apport de chaque métrique par rapport aux autres. Nous avons donc manipulé plusieurs métriques acoustiques nous fournissant des informations tant sur le maintien de la précision lors de productions stables (indice Phi, et plus particulièrement le CMintra) que sur un possible phénomène de centralisation des cibles vocaliques (VAI, CMinter). Il nous est apparu pertinent d'associer aux métriques fréquemment utilisées dans les recherches menées auprès de personnes atteintes de la MP, telles que les métriques tVSA ou VAI, une métrique telle que Phi puisqu'il s'agit d'un indice permettant d'étudier des phénomènes statiques – en l'occurrence des monophtongues – en tenant compte, dans son calcul, de plusieurs itérations pour un même sujet, et dès lors, de la variabilité inter- et intra-catégorielle, ce que ne nous permettent pas les autres indices. Or, nous savons que les difficultés dans la MP, et plus spécifiquement dans la production motrice de la parole dans la MP, touchent la dynamique des mouvements. Une métrique comme l'aire du triangle vocalique, très utilisée car porteuse d'informations sur l'étendue vocalique dans le plan F1/F2, nous apparait comme présentant une limite en ce sens qu'elle permet principalement d'observer des phénomènes périphériques, extrêmes avec un positionnement statique. L'utilisation de ces métriques nous permet d'établir deux constats : les productions des locuteurs atteints de la MP sont plus dispersées d'un point de vue intra-catégoriel et elles présentent également une forte centralisation. La dispersion intra-catégorielle suggère une instabilité des productions d'une même monophtongue, suggérant une difficulté à maintenir les articulateurs dans une position précise. Cette instabilité est renforcée par une centralisation des productions, entrainant une réduction des distances inter-catégorielles. Cette centralisation pourrait suggérer une diminution des contrastes vocaliques, renforcée par l'instabilité des productions. A nouveau, cette hypothèse se doit d'être vérifiée par des analyses articulatoires complémentaires.

Disposant d'informations à caractère individuel pour les locuteurs du groupe MP, nous avons pu nous interroger sur le lien entre ces aspects et les productions recueillies. Pour rappel, dans cette étude, les dossiers des participants atteints de la MP rapportaient différents degrés de dysarthrie. Pour certains, une absence de dysarthrie était constatée par les spécialistes (neurologue et orthophoniste, tous deux spécialisés dans l'évaluation et la prise en charge de personnes atteintes par la MP). Nous nous sommes dès lors interrogés sur les potentielles différences entre les productions orales de ce groupe de locuteurs réputés non-dysarthriques atteints de la MP et celles des locuteurs du groupe témoin. Les analyses effectuées nous permettent d'identifier, à partir de la métrique Phi, des différences déjà présentes entre nos deux groupes de locuteurs, ce qui nous suggère l'identification de phénomènes de nature infraclinique. Au vu de ces distorsions, l'utilisation de l'indice Phi nous confirme que l'organisation du système vocalique est beaucoup plus altérée chez les locuteurs atteints de la MP. Ces résultats nous confortent dès lors dans l'utilisation de la métrique Phi. Nous pensons qu'il est crucial de combiner ces différents regards si nous souhaitons obtenir une vision globale et réelle de l'espace vocalique et des multiples dimensions altérées et/ou préservées suite à la dysarthrie hypokinétique. De fait, à partir de cette expérimentation, il ressort que ces multiples regards sur des phénomènes statiques permettent de mettre en lumière des informations très riches et pertinentes dans la compréhension de phénomènes

résultant de la production motrice de la parole dans la MP. Etant donné le nombre de personnes qui connaitront au cours de leur maladie des troubles dysarthriques, identifier précocement ce symptôme dysarthrique nous semble revêtir un caractère crucial en vue d'une prise en charge précoce.

Références

AUDIBERT, N., & FOUGERON, C. (2012). Distorsions de l'espace vocalique : quelles mesures? Application à la dysarthrie. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, *1*, 217-224. Grenoble.

BRAAK, H., TREDICI, K. D., RÜB, U., DE VOS, R. A. I., JANSEN STEUR, E. N. H., & BRAAK, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. Neurobiology of Aging, 24(2), 197-211. DOI 10.1016/S0197-4580(02)00065-9

HOEHN, M. M., & YAHR, M. D. (1967). Parkinsonism: onset, progression, and mor-tality. *Neurology*, *17*(5), 427-442. DOI 10.1212/WNL.17.5.427

HUET, K., & HARMEGNIES, B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *XXIIIèmes Journées d'Etude sur la Parole*, *1*, 225-228. Aussois.

LAM, J., & TJADEN, K. (2016). Clear Speech Variants: An Acoustic Study in Parkinson's Disease. *Journal of Speech, Language, and Hearing Research*, *59*(4), 631-646. DOI 10.1044/2015_JSLHR-S-15-0216

LINDBLOM, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Éd.), *Speech Production and Speech Modeling* (p. 403-439). DOI 10.1007/978-94-009-2037-8_16

MOLLAEI, F., SHILLER, D. M., BAUM, S. R., & GRACCO, V. L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Research*, *1646*, 269-277. DOI 10.1016/j.brainres.2016.06.013

NEEL, A. T. (2008). Vowel Space Characteristics and Vowel Identification Accuracy. *Journal of Speech, Language, and Hearing Research*, *51*(3), 574-585. DOI 10.1044/1092-4388(2008/041)

ROLAND, V., DELVAUX, V., HUET, K., PICCALUGA, M., HAELEWYCK, M. C., & HARMEGNIES, B. (2016). Dynamique phonétique et contrôle moteur dans la maladie de Parkinson : analyse du contrôle de la production des glides. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, *1*, 211-219. Paris.

ROY, N., NISSEN, S. L., DROMEY, C., & SAPIR, S. (2009). Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy. *Journal of Communication Disorders*, *42*(2), 124-135. DOI 10.1016/j.jcomdis.2008.10.001

SAPIR, S. (2014). Multiple Factors Are Involved in the Dysarthria Associated With Parkinson's Disease: A Review With Implications for Clinical Practice and Research. Journal of Speech, Language, and Hearing Research, 57(4), 1330-1343. DOI 10.1044/2014_JSLHR-S-13-0039

SKODDA, S., VISSER, W., & SCHLEGEL, U. (2011). Vowel Articulation in Parkinson's Disease. Journal of Voice, 25(4), 467-472. DOI 10.1016/j.jvoice.2010.01.009

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Modèles de l'enrouement de la voix

Jean Schoentgen[1] Philipp Aichinger[2] Francis Grenez[1]
(1) B.E.A.M.S., Université Libre de Bruxelles, 50, Av. F.-D. Roosevelt, 1050 Bruxelles, Belgique
(2) Medical University of Vienna, Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Währinger Gürtel 18-20, 1090 Vienna, Austria
jschoent@ulb.ac.be, philipp.aichinger@meduniwien.ac.at, fgrenez@ulb.ac.be

RÉSUMÉ

L'objectif est l'étude des causes des dispériodicités des voix du type 1 qui sont pseudo-périodiques et monophoniques. Un modèle qui explique quantitativement les perturbations des durées de cycles glottiques fait appel aux fluctuations de la tension du muscle vocal. Or, ces fluctuations n'expliquent pas l'enrouement qui peut faire suite à une charge vocale ou une laryngite légère, par exemple. C'est pourquoi, nous discutons plusieurs modèles qui montrent qu'une redistribution des amplitudes vibratoires entre le corps et la couverture du pli module les perturbations qui trouvent leur origine au niveau du muscle vocal. Des simulations à l'aide d'un modèle corps-couverture suggèrent ainsi que les perturbations des durées des cycles glottiques augmentent avec une redistribution des amplitudes vibratoires de la couverture vers le muscle suite à une redistribution des masses vibrantes du muscle vers la couverture.

ABSTRACT

Models of vocal roughness

The objective is to model the origin of vocal dysperiodicities in type 1 voices that are pseudo-periodic and monophonic. An existing model that quantitatively explains the perturbations of the durations of the glottal cycles involves the fluctuations of the tension of the vocal muscle. However, these do not explain the increase of vocal jitter and roughness that may the consequence of vocal loading or light laryngitis, for instance. We therefore discuss several models that suggest that the relative amplitudes of vibration of the cover and body of the vocal folds modulate the perturbations that originate at the muscle. Simulations by means of a body-cover model of the folds show that the dysperiodicities of the durations of the glottal cycles increase with a shift of the amplitude of vibration from the cover to the muscle following a shift of vibratory mass from the muscle to the cover.

MOTS-CLÉS : jitter, flutter, tremblement vocal, raucité, qualité de voix, modèle corps-couverture.

KEYWORDS: jitter, flutter, vocal tremor, roughness, voice quality, body-cover model.

1 Introduction

L'objet de la présentation est l'étude de la relation entre les fluctuations de la tension du muscle vocal et les perturbations des durées des cycles glottiques qui sont des indices acoustiques de la qualité de voix de locuteurs normophoniques et dysphoniques. Lors de la production de sons de parole voisés, les perturbations de la périodicité stricte des voix du type 1 sont les suivantes, des plus lentes aux plus

Volume 1 : Journées d'Études sur la Parole, pages 570–578.
HAL : hal-02798588.

rapides (Kreiman & Sidtis, 2011; Buder & Strand, 2003; Cook, 1999).

— La dérive de la fréquence phonatoire F_o ainsi que le pleurage vocal (angl. "vow") $(< 2Hz)$,
— le tremblement vocal d'origine neurologique $(2Hz - 10Hz)$,
— le scintillement vocal (angl. "flutter") $(10Hz - 20Hz)$,
— le jitter vocal $(> 20Hz)$.

Les voix du type 1 incluent les voix normophoniques et sont produites lorsque les vrais plis vocaux vibrent de façon pseudo-périodique et monophonique dans un même mécanisme phonatoire (Behrman *et al.*, 1998).

Titze a proposé un modèle quantitatif des origines du jitter vocal des voix du type 1 qui fait appel aux fluctuations de la tension du muscle thyro-aryténoïdien *(TA)* (Titze, 1991). La tension d'un muscle squelettique est l'effet de l'activité conjointe de plusieurs unités motrices, chacune comprenant un neurone moteur qui innerve un groupe de fibres musculaires qui se contractent après l'arrivée d'un potentiel d'action, c.-à-d. une impulsion électrique émise par le neurone. Une mise à jour récente du modèle suggère que ces fluctuations sont à la fois à l'origine du jitter et flutter vocal et qu'elles contribuent au tremblement vocal sans l'expliquer (Schoentgen *et al.*, 2018; Schoentgen & Aichinger, 2019). En effet, l'origine du tremblement vocal neurologique est la lente fluctuation du taux d'émission des potentiels d'action des neurones moteurs (Kendall, 2013; Titze *et al.*, 2002) et l'origine du tremblement dit physiologique (perçu comme un pleurage vocal) est la circulation pulsée du sang ainsi que la respiration (Freund, 1987).

La contraction musculaire totale fluctue légèrement à cause de la superposition des secousses musculaires consécutives dans le temps et de la superposition spatiale de l'activité simultanée de beaucoup d'unités motrices. On appelle secousse musculaire la contraction simultanée de plusieurs fibres musculaires qui sont contrôlées par un même neurone moteur.

Les perturbations ΔT_m de la tension musculaire sont égales à la différence entre tension instantanée et tension moyenne T_m. Les perturbations relatives $\Delta f/f$ de la fréquence de vibration du muscle sont obtenues en les identifiant à la racine carrée des perturbations relatives de la tension, c.-à-d $\Delta f/f = \sqrt{\Delta T_m/T_m}$ (Titze, 1991).

L'objectif de la présentation est de montrer que les perturbations de la durée des cycles glottiques ne sont pas une copie exacte des perturbations de la fréquence de vibration du muscle *TA*. L'étude s'appuie sur des modèles "corps-couverture" du pli vocal qui représentent séparément la couche des tissus contractiles et la couche des tissus non-contractiles, appelées "corps" et "couverture" du pli respectivement. Le corps est essentiellement constitué du muscle et la couverture est mobile par rapport au corps.

La modélisation de la relation entre perturbations du corps et de la couverture offre la possibilité d'expliquer une large gamme de phénomènes sans postuler l'existence de nouvelles sources de perturbations chaque fois qu'une augmentation du jitter vocal ou un enrouement est constaté. En effet, la motivation de l'étude est l'expérience que, pour des voix du type 1, un chargement vocal ensemble avec (i) la consommation de boissons caféinées, alcoolisées ou gazeuses, (ii) le tabagisme, (iii) la (pré)menstruation chez les femmes, (iv) une laryngite légère ou (v) le chargement vocal en présence d'air sec peut augmenter le jitter vocal et éventuellement déclencher la perception d'une raucité lorsque le jitter des durées de cycles dépasse le seuil de 1%.

A. Fourcin avait proposé un modèle physiologique de ces observations qui exigeait l'injection d'atropine dans les plis vocaux (Abberton & Fourcin, 2006). La constatation que l'injection augmentait le jitter vocal tandis qu'un chargement vocal modéré le diminue en règle générale, suggère que les

propriétés constitutives ou dynamiques des plis vocaux peuvent altérer des perturbations vocales qui trouvent leur origine dans les fluctuations de la tension du muscle TA. Fourcin attribuait, d'ailleurs, l'amplification du jitter vocal à une augmentation de la viscosité des plis vocaux. Les simulations qui sont exposées ici ne confirment pas cette interprétation mais suggèrent une alternative qui est compatible avec l'injection d'un fluide.

L'étude de la relation entre les perturbations de la vibration du corps et de la couverture du pli fait appel à trois modèles existants, y compris deux versions d'un modèle dit à "trois masses" du pli vocal (Story & Titze, 1995). Une version qui inclut une perturbation aléatoire de la raideur du muscle est dédiée à la reproduction des prédictions des deux autres modèles tandis qu'une version déterministe en offre une interprétation physiologique. Le texte qui fait suite à (Schoentgen *et al.*, 2018; Schoentgen & Aichinger, 2019) développe les prédictions de trois modèles et présente les résultats de simulations numériques à l'aide du modèle à "trois masses" et en discute les implications pour la qualité de voix.

2 Modèles

Les sous-sections ci-après présentent trois modèles de la relation entre les perturbations de la fréquence de vibration du muscle vocal et les perturbations des durées de cycles de l'aire glottique. Les prédictions de deux modèles sont obtenues de façon algébrique ou analytique. Les solutions du modèle dit à "trois masses" sont obtenues numériquement avant d'être analysées statistiquement.

2.1 Modèle à superposition

On suppose que le muscle et la couverture vibrent à la même fréquence moyenne, mais que la phase instantanée du muscle $\phi_m = \phi_o + \theta_{pert}$ diffère légèrement de la phase instantanée de la couverture $\phi_c = \phi_o$ suite à une faible perturbation de la fréquence de vibration instantanée du muscle. Lorsque les plis ne sont pas en contact, la position du bord du pli x_{bord} est posée égale à la superposition des positions du muscle et de la couverture qui évoluent de façon sinusoïdale et d'une abduction x_{abd} constante. Les lettres A_m et A_c désignent les amplitudes vibratoires du muscle et de la couverture respectivement.

$$x_{bord} = x_{abd} + A_m \times \sin(\phi_o + \theta_{pert}) + A_c \times \sin(\phi_o) \tag{1}$$

En supposant que la perturbation θ_{pert} est petite, c.-à-d. $\cos\theta_{pert} \approx 1$, $\sin\theta_{pert} \approx \theta_{pert}$, $\tan^{-1}\theta_{pert} \approx \theta_{pert}$ et en appliquant une règle trigonométrique élémentaire, on obtient le modèle *à superposition* (2) qui montre que les perturbations du mouvement du bord des plis diminuent lorsque le rapport des amplitudes A_c/A_m augmente (Black, 1953).

$$x_{bord} = x_{abd} + (A_c + A_m) \times \sin(\phi_o - \frac{A_m}{A_m + A_c} \times \theta_{pert}) \tag{2}$$

2.2 Modèle à cordes

Titze a proposé une expression qui relie la fréquence de vibration (naturelle) F_o du pli vocal aux contraintes de traction passive et active ainsi qu'à la longueur L du pli, la densité des tissus ρ, l'amplitude de vibration latérale totale $A_c + A_m$ et l'amplitude de vibration du muscle A_m [cf. eq.

(3) de (Titze, 2011)]. La contrainte de traction active σ_m est due à la contraction du muscle *TA*. En calculant la dérivée de F_o par rapport à la contrainte de traction active, on obtient le modèle *à cordes* (3) qui montre que les perturbations ΔF_o de la fréquence instantanée diminuent lorsque le rapport des amplitudes A_c/A_m et la fréquence Fo augmentent. L'expression (3) est appelée modèle *à cordes* (au pluriel) car il repose sur l'hypothèse qu'un pli est formé de deux cordes qui ont la même longueur et qui vibrent en synchronie.

$$\Delta F_o = \frac{1}{8L^2\rho} \times \frac{1}{F_o} \times \frac{A_m}{A_m + A_c} \times \Delta\sigma_m \tag{3}$$

Les modèles *à superposition* et *à cordes* ont le terme $A_m/(A_m + A_c)$ en commun mais ne partagent pas le terme $1/F_o$. La raison en est que le modèle *à superposition* implique les perturbations de la phase instantanée et le modèle *à cordes* les perturbations de la contrainte de traction active.

Titze utilise le même symbole F_o pour désigner la fréquence phonatoire et la fréquence de vibration naturelle des plis et utilise ces termes comme synonymes, ce qui est au mieux une approximation (Titze, 2011). Une lecture attentive suggère que l'interprétation la plus plausible est que la fréquence F_o du modèle *à cordes* désigne la fréquence de vibration naturelle.

2.3 Modèle à "trois masses"

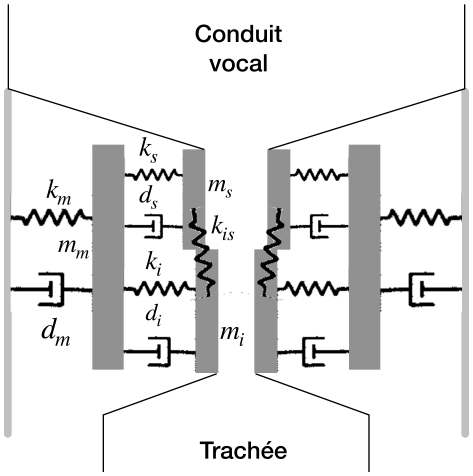


FIGURE 1 – Vue du modèle à "trois masses". Le modèle est supposé symétrique par rapport à la ligne médiane qui est l'origine des coordonnées. Les indices s, i et m ont trait aux masses correspondantes ; k désigne la raideur des ressorts, m les masses vibrantes et d l'amortissement. Les masses m_i et m_s sont identiques.

L'implémentation numérique du modèle à "trois masses" est inspirée du formalisme de (Story & Titze, 1995), mais omet l'interaction entre le conduit et la source, c.-à-d. la pression supra-glottique est posée égale à la pression atmosphérique. En effet, Story et Titze ont proposé un modèle à "trois masses" ($3M$) du pli vocal qui est une approximation de la structure "corps-couverture" par un modèle à éléments localisés. Ils combinent le modèle à "deux masses" conventionnel avec une troisième masse qui simule le muscle. La Figure (1) résume les éléments essentiels du modèle.

Paramètres	Unités	Intervalles	Références
(sur)pression pulmonaire	g/s^2cm	7500-12500	Birkholz et al. 2011
raideur k_{is}	g/s^2	1875-3125	Story et al. 1995
masse totale m_{tot}	g	0.075-0.125	Birkholz et al. 2011
m_m/m_{tot}	$s.o.$	0.525-0.875	
fréquence naturelle F_n	Hz	120-200	
adduction (i et s)	cm	0.015-0.025	Story et al. 1995
dist. corps→couverture	cm	0.21-0.35	Story et al. 1995
amortissement $\zeta_{i,s}$	$s.o$	0.3-0.5	Story et al. 1995
amortissement ζ_m	$s.o.$	0.15-0.25	Story et al. 1995

TABLE 1 – Intervalles dans lesquelles les valeurs des paramètres de contrôles sont choisies aléatoirement une fois par simulation. Les indices i et s désignent les masses inférieures et supérieures respectivement, m_m désigne la masse du muscle et k_{is} la raideur du couplage intra-couverture, cf. Fig. (1). Les coefficients d'amortissement ζ sont des grandeurs sans dimension à partir desquelles les amortissements d sont calculés : $d = 2 \times \zeta \times \sqrt{m \times k}$ (Story & Titze, 1995; Birkholz *et al.*, 2011).

3 Simulations numériques

Les simulations numériques font appel à une version "aléatoire" et une version "déterministe" du modèle $3M$. Le rôle de la première est de vérifier les prédictions des modèles *à superposition* et *à cordes* dans le cadre du modèle $3M$ en perturbant aléatoirement la raideur k_m du muscle et en enregistrant la taille des perturbations des durées des cycles de l'aire glottique. Le rôle de la version déterministe est de découvrir quels paramètres physiologiques déterminent des caractéristiques de la vibration telles que le rapport des amplitudes de vibration $A_m/(A_m + A_c)$, la fréquence phonatoire F_o et le quotient d'ouverture de la glotte Q_o.

Pour chacune des deux versions, 1000 solutions du type 1 ont été obtenues numériquement par la méthode de Runge-Kutta d'ordre quatre avec un pas de calcul de $5\mu s$. Pour chaque simulation, les valeurs des paramètres ont été fixées aléatoirement en puisant dans les intervalles indiqués dans le Tableau (1). La taille de chaque intervalle est égale à $(M - 0.25 \times M, M + 0.25 \times M)$ avec M une valeur plausible du paramètre. La fréquence naturelle $F_n = 1/2\pi\sqrt{k/m}$ est la fréquence de vibration des masses isolées. Elle permet de calculer la raideur k du ressort si la masse m est connue. Afin de favoriser la génération de solutions du type 1, la même fréquence naturelle a été affectée au muscle et à la masse inférieure de la couverture. La raideur du ressort de la masse supérieure a été fixée à 70% de la raideur inférieure (Story & Titze, 1995).

Les simulations à l'aide de la version dite "aléatoire" comprenaient deux passages, un passage "déterministe" afin de sélectionner les solutions du type 1, suivi d'un deuxième passage avec les mêmes valeurs de paramètres, mais incluant les perturbations de la raideur k_m. Les solutions déterministes dont les perturbations des durées des cycles dépassaient 0.001% ont été écartées.

Les perturbations de la raideur du muscle Δk_m ont été simulées par des échantillons de bruit uniforme filtrés passe-bas sélectionnés au préalable dans un intervalle fixe $(0, 6300g/s^2)$. La valeur $6300g/s^2$ correspond approximativement à 1% de la valeur typique de la raideur k_m et permet de générer des perturbations relatives des durées de cycles dans un intervalle $(0.1 - 1\%)$. La largeur de bande des

$A_m/(A_i+A_m)$	poids	F_o	poids	Q_o	poids
$m_i/(m_i+m_m)$	$+0.97$	F_n	$+0.83$	F_n	$+0.66$
F_n	-0.25	$m_i/(m_i+m_m)$	-0.53	$amort.\ \zeta_i$	$+0.53$
m_{tot}	$+0.13$	abd	$+0.10$	$(sur)pression$	-0.31
R^2_{ajst}	$+0.97$	R^2_{ajst}	$+0.96$	R^2_{ajst}	$+0.90$

TABLE 2 – Poids de régression des paramètres physiologiques du modèle $3M$ les plus importants pour le rapport $A_m/(A_m+A_i)$, la fréquence phonatoire F_o et le quotient d'ouverture Q_o. Le symbole m désigne les masses vibrantes, A les amplitudes de vibration et les indices i et m font référence à la couverture et au muscle respectivement (cf. Fig. 1). Les corrélations au carré R^2_{ajst} ont trait à la totalité des paramètres.

perturbations est de $50Hz$ et le filtre passe-bas est un filtre à réponse impulsionelle finie et symétrique.

Le modèle *à cordes* (3) suggère d'analyser les solutions numériques en régressant les perturbations $\log \Delta T$ des durées de cycles glottiques sur les log-variables $\log 1/F_n^3$ et $\log A_i/(A_i+A_m)$. L'amplitude A_i tient lieu d'amplitude vibratoire de la couverture. Le logarithme transforme le produit du modèle *à cordes* en une somme, et le cube de F_n transforme les perturbations ΔF de la fréquence en les perturbations ΔT des durées de cycles correspondantes. La section 4.2 discute l'approximation de la fréquence F_o du modèle *à cordes* par la fréquence naturelle F_n.

Les durées de cycles ont été obtenues à partir des passages à zéro de l'aire glottique après soustraction de la moyenne. La fréquence phonatoire F_o est l'inverse de la moyenne des durées de cycles. Les amplitudes A_i et A_m sont les écarts maximaux des masses m_i et m_m par rapport à leurs positions de repos. Les variables dépendantes et indépendantes ont été normalisées en les divisant par leurs maxima respectifs avant le calcul des logarithmes et l'analyse par régression linéaire.

Les simulations "déterministes" du modèle $3M$ sont analysées en régressant les variables $A_m/(A_m+A_i)$, F_o et Q_o sur les paramètres physiologiques affichés dans le Tableau (1). Toutes les quantités ont été z-normalisées afin de faciliter la comparaison des poids de régression.

4 Résultats

4.1 Simulations déterministes

L'objectif d'une série de 1000 simulations a été de découvrir une interprétation physiologique de la fréquence phonatoire F_o, du rapport des amplitudes $A_m/(A_m+A_i)$ et du quotient d'ouverture de la glotte Q_o. Les deux premiers font partie du modèle *à superposition* ou du modèle *à cordes* et le dernier permet de mettre en perspective la prédiction qu'amortissement des plis vocaux et jitter vocal augmentent ensemble. Les simulations sont déterministes, c.à-d. sans perturbation aléatoire de la raideur k_m du "muscle" du modèle $3M$. Le Tableau (2) indique les poids de régression des paramètres physiologiques *z-normalisés* du modèle $3M$. Le nombre de poids est limité aux trois les plus importants. On constate que la fréquence naturelle F_n et le rapport des masses $m_i/(m_i+m_m)$ sont les paramètres physiologiques explicatifs les plus proéminents. Les deuxième et troisième lignes du tableau montrent, en effet, que les variables cinétiques F_o et $A_m/(A_m+A_i)$ sont déterminées par

les variables physiologiques F_n et $m_i/(m_i + m_m)$ qui sont la fréquence naturelle et le rapport des masses vibrantes. L'influence partagée des deux derniers sur les deux premiers explique d'ailleurs la corrélation entre F_o et $A_m/(A_m + A_i)$ qui est égale à -0.63. Finalement, le quotient d'ouverture Q_o est la seule variable cinétique qui est fortement influencée par l'amortissement de la couverture du modèle $3M$.

Le nombre de poids dans le tableau est limité à trois. Les paramètres physiologiques restants contribuent de façon négligeable. En effet, tous les poids non publiés sont de l'ordre de 10^{-2}, c.-à-d. un ordre de grandeur plus petit que ceux de la fréquence naturelle et du rapport des masses. Les corrélations au carré ajustées R^2_{ajst} ont trait à la totalité des paramètres.

4.2 Simulations "aléatoires"

L'objectif d'une deuxième série de 1000 simulations a été de vérifier les prédictions du modèle *à superposition* ou *à cordes* (2) et (3) dans le cadre du modèle $3M$ en perturbant la raideur du "muscle" du modèle et en observant la relation entre les perturbations ΔT des durées des cycles et le rapport des amplitudes $A_m/(A_m + A_i)$ et la fréquence F_o.

La fréquence F_o du modèle *à cordes* désigne la fréquence naturelle de vibration du pli vocal.
On a le choix entre deux approximations de cette fréquence. L'une est la fréquence phonatoire F_o mesurée, l'autre la fréquence de vibration naturelle F_n du "muscle" et de la "couverture" du modèle $3M$. Lors d'une analyse par régression linéaire, un critère de sélection des variables explicatives est qu'elles soient aussi peu corrélées que possible. Ce critère favorise le choix de F_n (-0.18 versus -0.63, cf. section 4.1). L'équation de régression est la suivante après *normalisation* des variables dépendantes et indépendantes.

$$\log \Delta T = 0.73 \times [\log 1/F_n^3] + 1.27 \times [\log A_m/(A_m + A_i)] - 0.12, R^2_{ajst} = 0.81 \qquad (4)$$

Les poids de régression sont statistiquement significatifs ($p < 0.001$) et la valeur de la corrélation au carré $R^2_{ajst} = 0.81$ indique que les deux variables expliquent un large pourcentage de la variabilité des perturbations ΔT lors de 1000 simulations.

La deuxième colonne du Tableau (2) montre que le rapport $A_m/(A_m + A_i)$ est expliqué par F_n et le rapport des masses vibratoires $m_i/(m_i + m_m)$. On peut donc répéter l'analyse par régression des perturbations ΔT avec ces deux variables. Elles ne sont pas corrélées car les valeurs de F_n, m_i et m_m sont choisies aléatoirement. L'équation de régression est la suivante après *normalisation* des variables.

$$\log \Delta T = 0.83 \times [\log 1/F_n^3] + 0.53 \times [\log m_i/(m_i + m_m)] - 0.11, R^2_{ajst} = 0.81 \qquad (5)$$

Les poids sont statistiquement significatifs ($p < 0.001$) et l'équation de régression (5) offre une interprétation physiologique du modèles *à superposition* (2), du modèle *à cordes* (3) et de l'équation de régression (4) en substituant le rapport des masses vibrantes au rapport des amplitudes de vibration.

5 Discussion et conclusion

— L'attribution de valeurs typiques aux paramètres des modèles *à superposition*, *à cordes* ou $3M$ montre que les perturbations relatives des durées des cycles glottiques sont inférieures aux

perturbations relatives de la fréquence de vibration du muscle. Observer une augmentation des perturbations glottiques implique par conséquent que l'atténuation des perturbations du muscle a diminué lors du transfert à travers le pli. L'explication en est la redistribution des amplitudes vibratoires de la couverture vers le muscle ce qui, dans le cadre du modèle $3M$, correspond à une redistribution des masses vibratoires du muscle vers la couverture.

— Les valeurs élevées des carrés des corrélations R^2_{ajst} des équations de régression (4) et (5) indiquent que la fréquence de vibration naturelle F_n, le rapport des amplitudes vibratoires $A_m/(A_m + A_i)$ ou le rapport des masses vibrantes $m_i/(m_i + m_m)$ expliquent un large pourcentage de la variabilité des perturbations ΔT des durées de cycles dans le cadre du modèle $3M$. Ceci suggère que le modèle *à cordes* (3) est, en effet, une représentation compacte du transfert des perturbations du muscle à la glotte et que (3) peut guider l'interprétation des perturbations vocales de voix naturelles (du type 1).

— Les poids de régression et signes des paramètres du modèle $3M$ (Tableau 2) suggèrent qu'une augmentation de la masse vibrante de la couverture ou une diminution de la masse vibrante du muscle aggrave la voix et amplifie les perturbations des durées de cycles en modifiant le rapport des amplitudes vibratoires. En d'autres termes, un excès de sécrétions sur la surface des plis, un oedème de la couverture des plis ou une hypotrophie du muscle vocal peuvent amplifier les perturbations vocales et favoriser l'émergence d'une raucité. Une condition est que le rapport des amplitudes vibratoires $A_m/(A_m + A_c)$ augmente. En effet, les modèles *à superposition* et *à cordes* (2) et (3) suggèrent que la redistribution des amplitudes vibrantes entre la couverture et le corps est la cause de la modulation du transfert des perturbations à travers le pli. La redistribution des masses vibrantes en est une interprétation dans le cadre du modèle $3M$.

— A. Fourcin a exploré par injection d'atropine le rôle de la couverture des plis dans le transfert des perturbations musculaires (Abberton & Fourcin, 2006). Il a observé un accroissement du jitter vocal qu'il attribuait à une augmentation de la viscosité de la couverture. Les simulations à l'aide du modèle $3M$ ne confirment pas cette prédiction. Mais, elles offrent une explication alternative en termes d'une redistribution des masses vibrantes suite à l'injection d'un fluide.

— Cependant, même en restant dans le cadre du modèle $3M$, on ne peut pas exclure que la viscosité a une influence sur le jitter vocal car le modèle $3M$ n'inclut pas un amortissement intra-couverture, par exemple. De même, les modèles *à superposition* et *à cordes* reposent sur l'hypothèse que corps et couverture vibrent en phase, à des petites perturbations près, omettant ainsi l'influence de la phase relative entre corps et couverture sur l'amplitude de la vibration du pli. Aucun des modèles ne rend compte de l'influence possible de l'amortissement de la couverture sur la propagation de l'onde de surface.

— Les simulations à l'aide du modèle $3M$ accordent, par contre, un rôle à la viscosité de la couverture du pli dans l'augmentation du quotient d'ouverture de la glotte. La viscosité favoriserait donc un timbre soufflé ou asthénique plutôt qu'un timbre rauque (cf. Tableau 2, sixième colonne).

Remerciements

Le travail a été soutenu par le "Austria Science Fund (FWF) : KLI722-B30".

Références

ABBERTON E. & FOURCIN A. (2006). Electrolaryngography. In M. BALL & C. CODE, Éds., *Instrumental Clinical Phonetics* : Wiley. p. 119.

BEHRMAN A., AGRESTI C., BLUMSTEIN E. & LEE N. (1998). Microphone and electroglottographic data from dysphonic patients : Type 1, 2 and 3 signals. *J. of Voice*, **12**, 249–260. DOI : 10.1016/S0892-1997(98)80045-3.

BIRKHOLZ P., KRÖGER B. J. & NEUSCHAEFER-RUBE C. (2011). Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds. In *Proc. of the First International Workshop on Performative Speech and Singing Synthesis*, Vancouver, Canada.

BLACK H. (1953). *Modulation theory*. Van Nostrand. p. 221.

BUDER E. H. & STRAND E. A. (2003). Quantitative and graphic acoustic analysis of phonatory modulations : the modulogram. *J. Speech, Language and Hearing Res.*, **46**, 475–490.

COOK P. R. (1999). Pitch, periodicity, and noise in the voice. In P. R. COOK, Éd., *Music, Cognition and Computerized Sound : An Introduction to Psychoacoustics* : The MIT Press. page 199.

FREUND H.-J. (1987). Central rhythmicities in motor control and its perturbances. In *Temporal Disorder in Human Oscillatory Systems*, volume 36, p. 79–82 : Springer. DOI : doi.org/10.1007/978-3-642-72637-8_9.

KENDALL K. A. (2013). Vocal tremor. In G. GRIMALDI & M. MANTO, Éds., *Mechanisms and Emerging Therapies in Tremor Disorders* : Springer. p. 239, DOI : 10.1007/978-1-4614-4027-7.

KREIMAN J. & SIDTIS D. (2011). *Foundations of Voice Studies : An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell. page 55, DOI : 10.1002/9781444395068.

SCHOENTGEN J. & AICHINGER P. (2019). Analysis and synthesis of vocal flutter and vocal jitter. In *Proceedings INTERSPEECH 2019*, p. 2518 – 2522, Graz.

SCHOENTGEN J., DHOUHA R. & GRENEZ F. (2018). Simulation numérique des apériodicités vocales dues aux fluctuations de la tension musculaire. In *XXXIIe Journées d'Etudes sur la Parole*, p. 267–275, Aix-en-Provence : Association Francophone pour la Communication Parlée. DOI : 10.21437/JEP.2018-31.

STORY B. H. & TITZE I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *J. Acoust. Soc. Am.*, **97**, 1249–1260.

TITZE I. R. (1991). A model for neurologic sources of aperiodicity in vocal fold vibration. *J. Speech and Hearing Res.*, **34**, 460–472.

TITZE I. R. (2011). Vocal fold mass is not a useful quantity for describing f_o in vocalization. *J. Speech Lang. Hear. Res.*, **54**, 520–522. DOI : 10.1044/1092-4388(2010/09-0284).

TITZE I. R., STORY B., SMITH M. & LONG R. (2002). A reflex resonance model of vocal vibrato. *J. Acoust. Soc. Am.*, **111**, 2272–2282. DOI : 10.1121/1.1434945.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La « voyelle apicale » n'est pas une voyelle : étude acoustique et articulatoire de la voyelle apicale en chinois de Jixi

Bowei Shao & Rachid Ridouane
Laboratoire de Phonétique et Phonologie, CNRS/Sorbonne Nouvelle
19, rue des Bernardins, 75005 Paris, France
bowei.shao@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

RÉSUMÉ

Cette étude s'intéresse à la « voyelle apicale », notée /z/, telle qu'elle est attestée en chinois de Jixi. L'objectif est de déterminer sa nature phonétique sur la base de données acoustiques et articulatoires. Phonologiquement, ce segment est un phonème distinct qui s'oppose à /i/ dont il est issu diachroniquement. Il est exclusivement attesté en position noyau de syllabe où il constitue une unité porteuse de ton. Sur le plan articulatoire, les données ultrasoniques démontrent que, quand il est précédé de consonnes bilabiales /p, pʰ, m/, il présente un geste articulatoire semblable à celui de la fricative alvéolaire /s/. Ce geste est réalisé de manière anticipatoire durant la tenue des bilabiales. Une des conséquences de cette réalisation est que le relâchement de /pʰ/ présente les mêmes caractéristiques acoustiques que le bruit de friction de la fricative /s/, comme l'atteste la ressemblance de leur centre de gravité. Ces résultats montrent que la voyelle apicale en chinois de Jixi est mieux définie, au moins du point de vue phonétique, comme une fricative alvéolaire.

ABSTRACT

The 'apical vowel' is not a vowel: An acoustic and articulatory study of the apical vowel in Jixi-Hui Chinese
This study focuses on the 'apical vowel', noted /z/, as attested in Jixi-Hui Chinese. The objective is to determine its phonetic nature on the basis of acoustic and articulatory data. Phonologically, this segment is a distinct phoneme that opposes to /i/ from which it is diachronically derived. It is exclusively attested in the syllable nucleus position where it constitutes a tone-bearing unit. On the articulatory level, the ultrasonic data show that when it is preceded by the bilabial consonants /p, pʰ, m/, it presents an articulatory gesture which is virtually identical to that of the alveolar fricative /s/. This gesture is realised in an anticipatory way during the bilabials. One of the consequences of such a realisation is that the release phase of /pʰ/ exhibits the same acoustic characteristics as the friction noise of the fricative /s/, as evidenced by the resemblance of their centre of gravity. These results show that the apical vowel in Chinese Jixi is better defined, at least phonetically, as an alveolar fricative.

MOTS-CLÉS : Voyelles apicale, chinois de Jixi, acoustique, ultrason.
KEYWORDS: Apical vowel, Jixi-Hui Chinese, acoustics, ultrasound.

1 Introduction

Les langues chinoises sont connues pour avoir une série de segments atypiques, nommés « voyelles apicales » (Karlgren, 1915). Le cas le plus connu est attesté en chinois standard (CS), où ces segments n'apparaissent qu'après des sibilantes alvéolaires [s, ts, tsʰ] ou post-alvéolaires [ʃ, tʃ, tʃʰ], et sont considérés comme homorganiques avec celles-ci (Dell, 1994 ; Duanmu, 2007). En CS, cette série de

Volume 1 : Journées d'Études sur la Parole, pages 579–587.
HAL : hal-02798589.

segments est traditionnellement considérée comme variante allophonique en distribution complémentaire avec la voyelle [i] (Hartman, 1944). Ils occupent, comme les autres voyelles de la langue, la position noyau de syllabe et fonctionnent comme unités porteuses de ton (UPT). Il n'existe pas de consensus sur leur nature phonétique, mais l'homorganicité avec les attaques sibilantes et l'existence d'un bruit de friction durant cette série de segments sont souvent notées dans les différentes descriptions. Troubetzkoy (1949 : 198) remarque ainsi qu'ils sont « *une sorte de voyelle ayant un degré d'aperture très petit et un point d'articulation beaucoup plus en avant que celui de i par exemple, de sorte que dans son émission on entend un bruit fricatif semblable à un bourdonnement* ». Se basant sur l'homorganicité avec les attaques sibilantes et la présence de bruit de friction, Ladefoged & Maddieson (1996 : 314) les décrivent comme « voyelles fricatives » qui sont produites avec la langue la plupart du temps dans la même position que les sibilantes en attaques. Les données acoustiques et articulatoires de Lee-Kim (2014) montrent qu'il n'y a pas de bruit de friction chez la majorité des locuteurs qu'elle a enregistrés, mais indiquent en revanche que ces segments sont effectivement homorganiques avec les sibilantes qui les précèdent.

En considérant leur statut et fonction phonologique (*i.e.* allophonique avec [i], en position noyau d'une syllabe), cette série de segments est analysée comme voyelles apicales (Karlgren, 1915 ; Hartman, 1944) ou voyelles fricatives (Ladefoged & Maddieson, 1996). En se basant sur leurs caractéristiques phonétiques (*i.e.* présence de bruit de friction et homorganicité avec les attaques sibilantes), d'autres chercheurs ont analysé ces segments comme fricatives/sibilantes syllabiques (Dell, 1994 ; Yu, 1999 ; Duanmu, 2007), ou comme approximantes syllabiques (Lee & Zee, 2003 ; Lee-Kim, 2014). Ainsi, la particularité de ces segments est que leur statut et leurs fonctions phonologiques suggèrent qu'il s'agirait de voyelles, alors que leurs caractéristiques phonétiques suggèrent qu'il s'agirait plutôt de consonnes (fricatives ou approximantes). Pour arriver à une meilleure compréhension des voyelles apicales dans les langues chinoises, nous présentons ici une étude acoustique et articulatoire de la voyelle apicale que l'on trouve en chinois de Jixi (CJ). Cette nouvelle étude empirique montre que la voyelle apicale en CJ, qui sera notée ici /z/, s'apparente plutôt à une consonne fricative.

1.1 La voyelle apicale du chinois de jixi : phonologie et phonétique

La langue chinoise de Jixi 绩溪 (CJ) est une langue du groupe Hui 徽, parlée dans le district Jixi, au sud de la province de l'Anhui 安徽 (Zhao, 2003 ; Hirata, 1998). Les descriptions existantes de cette langue sont faites à partir de la variante parlée dans la ville de Jixi. Notre étude est aussi basée sur des données émanant de cette variante. La voyelle apicale /z/ en CJ, contrairement au même segment en CS, est un phonème qui s'oppose à la voyelle /i/, comme le montrent les paires minimales suivantes : [tsz↓] 'poule' *vs* [tsi↓] 'un nom de famille' ; [ts^{h}z↓] 'déception' *vs* [ts^{h}i↓] 'automne' ; [sz↓] 'soie' *vs* [si↓] 'réparer'. De plus, il a une distribution plus large qu'en CS. Il apparaît en effet dans les syllabes [pz, p^{h}z, mz, nz] en plus de [tsz, ts^{h}z, sz] et ces syllabes ne sont pas lexicalement rares, comptant pour 7,2% des entrées monosyllabiques du dictionnaire du CJ (Zhao, 2003). Comme ces syllabes le montrent, /z/ n'est pas systématiquement homorganique avec les consonnes précédentes, puisqu'il apparaît à la fois à la suite des consonnes bilabiales /p, p^{h}, m/ et des consonnes alvéolaires /s, ts, ts^{h}, n/.

Nos études précédentes sur ce segment ont permis de clarifier certaines de ces caractéristiques phonétiques qui nous ont permis de l'analyser comme une fricative alvéolaire syllabique. Au niveau

acoustique (Shao & Ridouane, 2018), la majorité des réalisations du segment /z/ en CJ contient du bruit de friction au début du segment (voir FIGURE 1), et le lieu d'articulation de la consonne en attaque n'a pas d'effet important sur la quantité de bruit observée. Quand le bruit de friction diminue, la structure formantique devient plus nette. Cette structure formantique est qualitativement différente de celle de la voyelle /i/, mais présente un F1 similaire à celui de /u/ et un F2 similaire à celui de /ʉ/.

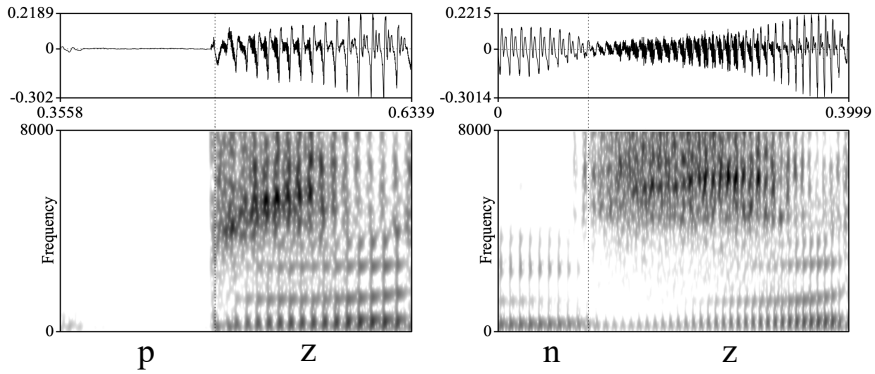


FIGURE 1 : Signaux acoustiques et spectrogrammes des formes /pz/ et /nz/ produites par MS3 (fenêtre temporelle : 5 ms) (Shao & Ridouane, 2018).

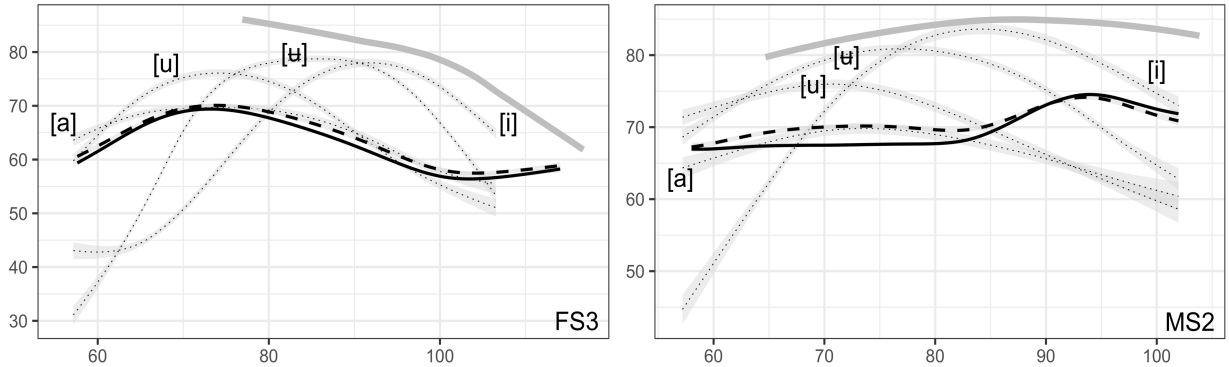


FIGURE 2 : Contours de langue de deux locuteurs en coupe sagittale, extraits en coordonnées (mm) x/y à l'image ultrason correspondant au milieu de chaque segment (en se basant sur le signal acoustique). Ces contours ont été généralisés en SS ANOVA avec l'intervalle de confiance bayésienne à 95%. Les tracés solides représentent /z/. Les tracés discontinus représentent la consonne /s/ en attaque de syllabe. Les tracés pointillés représentent les voyelles /i, ʉ, u, a/. Les tracés gris et pleins représentent le palais. Les noyaux /z, i, ʉ, u, a/ sont généralisés à partir des contextes consonantiques /p, p^h, m, n, ts, ts^h, s/, l'attaque /s/ est généralisée à partir des contextes nucléiques /z, i, ʉ, u, a/. La pointe de la langue est à droite (Shao & Ridouane, 2019).

Au niveau articulatoire, /z/ est produit avec une configuration linguale semblable à celle de la consonne /s/ (Shao & Ridouane, 2019). La FIGURE 2 montre les contours de la langue pour deux locuteurs, généralisés en SS ANOVA (smoothing-spline analysis of variance). Dans cette figure, les tracés des noyaux /z, i, u, ʉ, a/ sont généralisés à partir de sept contextes consonantiques différents. À titre comparatif, les contours de la langue pour la consonne attaque /s/ est aussi présentée. Nous observons deux stratégies articulatoires pour /z/. Pour le locuteur FS3, la pointe de la langue est élevée mais le dos de la langue est encore plus élevé que la pointe. Pour MS2, la pointe de la langue est élevée, mais le dos de la langue reste plat, laissant ainsi la pointe de la langue plus élevée que le dos. Cependant pour chacun des deux locuteurs, les contours de /z/ sont quasi-identiques aux contours de /s/. Nous avons par ailleurs montré que la nature de la consonne attaque (coronale ou labiale) n'a pas d'effet sur la configuration linguale de /z/.

2 Problématique

Les résultats des études acoustiques et articulatoires citées ci-dessus montrent que la voyelle apicale en CJ est mieux analysée comme une fricative alvéolaire, puisqu'elle présente un geste articulatoire semblable à celui de la fricative alvéolaire /s/, et contient du bruit de friction. Ces deux caractéristiques sont observées quand /z/ est précédé des consonnes alvéolaires et labiales. Dans cette étude, nous nous intéressons plus spécifiquement aux caractéristiques acoustiques et articulatoires de /z/ dans le contexte des consonnes labiales. L'idée est que ces consonnes, n'impliquant pas d'articulation linguale, doivent avoir un minimum d'impact sur la réalisation apicale de /z/. Les caractéristiques acoustiques et articulatoires observées dans ce contexte doivent ainsi être considérées comme propres à ce segment /z/.

3 Méthode

Les données acoustiques examinées dans cette étude émanent de dix locuteurs natifs, cinq hommes (MS1-5) et cinq femmes (FS1-5). Ils sont nés entre 1964 et 1974 (âge moyen : 49 ± 3,8). Tous les locuteurs ont grandi dans la ville de Jixi avec leurs parents qui sont eux aussi nés et ont grandi dans la même ville. Ils parlent la même variante du CJ dans leurs milieux familiaux et professionnels et se considèrent comme locuteurs natifs sans accent. Les sessions d'enregistrement acoustique ont été effectuées dans la ville de Jixi, dans un studio de télévision locale à l'aide d'un micro-casque hypercardioïde (AKG C520), d'une carte son Edirol UA25 et avec le logiciel Audacity (V2.1.0). Les données ultrasoniques sur le plan mid-sagittal ont été obtenues un an après les données acoustiques, dans les mêmes conditions et avec un sous ensemble des mêmes locuteurs (FS1, FS3, FS5 et MS2, MS3, MS5). L'enregistrement a été effectué à l'aide de l'*Ultrasound Stabilisation Headset* (Articulate Instruments Ltd., 2008) et du logiciel AAA (*Articulate Assistant Advanced*) (V217.03) (Articulate Instruments Ltd., 2012). Les données ont été enregistrées avec une probe d'ultrason microconvex à 40 mm de diamètre, le champ de vision a été fixé à 92°. En raison des spécificités morphologiques des locuteurs, la profondeur a été ajustée pour avoir une vue maximale de la langue, résultant en des fréquences d'images différentes (82,1 fps pour les femmes et 81,4 fps pour les hommes). La synchronisation des enregistrements échographiques et de l'audio correspondant a été effectuée automatiquement par le logiciel AAA. Pendant les sessions d'enregistrement, les locuteurs étaient assis dans un fauteuil, et il leur a été demandé de lire de manière naturelle une liste de mots en les incorporant dans une phrase porteuse. Pour l'enregistrement des données ultrason, la liste a été enregistrée phrase par phrase, avec une pose de quelques secondes entre chaque phrase porteuse prononcée.

Nous avons établi une liste de monosyllabes avec /a, i, u, ʉ, z/ comme noyaux, et /p, pʰ, m, n, ts, tsʰ, s/ comme attaques. Ces syllabes ont les tons /˥, ˧, ˨˩, ˥˩/ et forment des mots réels, inclus dans la phrase cadre /ki˧ ʂɔ˨˩ _ ʂɔ˨˩ sɔ˧ fa˥˩/ 'Il écrit _ trois fois'. Les phrases ont été répétées cinq fois pour l'enregistrement acoustique et trois fois pour l'enregistrement ultrasonique. Les syllabes cibles ont été segmentées manuellement en utilisant Praat et AAA. La frontière gauche des noyau /a, i, u, ʉ/ est basée sur le relâchement des attaques plosives et la fin de la friction des attaques sibilantes ; leur frontière droite est basée sur la fin de la structure formantique. La segmentation du noyau /z/ est basée sur la première « pulse » détectée par Praat pour la frontière gauche, en raison de la continuité de bruit de friction entre les attaques /pʰ, ts, tsʰ, s/ et le noyau /z/ (Shao & Ridouane, 2018). La frontière droite de /z/ a été délimitée par la fin de la structure formantique.

Deux paramètres acoustiques ont été mesurés : la durée de la voyelle apicale dans différents contextes consonantiques, et le centre de gravité (COG) de la phase d'aspiration de la consonne /pʰ/ que nous avons comparé avec le COG de la consonne /s/. Les contours de langues obtenus pour l'analyse articulatoire sont extraits en coordonnées x/y à l'image correspondant au milieu de chaque segment, sauf pour la consonne /pʰ/. Cette consonne a été divisée en deux phases, la phase d'occlusion et la phase d'aspiration, et deux images ont ainsi été obtenues correspondant au milieu de la phase d'occlusion et au milieu de la phase d'aspiration, respectivement. Les coordonnées en x/y sont généralisées en SS ANOVA (Davidson, 2006) à l'aide de R et le package gss (Gu, 2014). SS ANOVA est une procédure statistique qui permet d'étudier les similitudes et les différences de formes de courbes (Davidson, 2006 ; Lee-Kim, 2014). Cette méthode permet de générer une courbe lisse qui correspond le mieux aux différentes répétitions des segments cibles. Les palais sont obtenus et moyennés à partir de six taches séparées où chaque locuteur a été enregistré en train d'avaler de l'eau.

4 Résultats

4.1 La durée de la voyelle apicale dans les syllabes /pz, pʰz, mz/

Nous avons observé des différences importantes de durée entre les voyelles apicales précédées par /pʰ/ et les voyelles apicales précédées par /p, m/ (voir FIGURE 3). Cette différence systématique est spécifique à ce noyau, et n'a pas été observée dans les autres contextes. La durée de /z/ précédé par les attaques labiales est en distribution normale selon le test de Shapiro-Wilk (p=0,25). Le test ANOVA sur la durée de /z/ confirme qu'il y a une différence significative entre les trois contextes (F(2, 229)=21,12 ; p<0.001). Le test post-hoc TukeyHSD indique que cela est dû à la différence entre /pʰ/ d'un côté et /p, m/ de l'autre (/pʰ/ *vs* /p/ (p<0,001), /pʰ/ *vs* /m/ (p<0,001)). La différence de durée entre les contextes /m/ et /p/ n'est pas significative (p=0,74).

La durée des voyelles dans les langues chinoises est susceptible d'être influencée par les tons lexicaux. Ainsi un ton de type descendant-montant a tendance à allonger les voyelles porteuses de ce ton (Ho, 1976 ; Howie, 1976). Nous avons analysé ces différences de durée uniquement pour le ton /˨˩˦/ (le plus nombreux dans nos données), et nous avons observé que la différence reste significative (F(2, 134)=20,18 ; p<0,001). Le test post-hoc TukeyHSD indique là aussi que cela est dû à la différence entre les contextes /pʰ/ et /p/ (p<0,001), et entre les contextes /pʰ/ et /m/ (p<0,001). La différence de durée entre les contextes /m/ et /p/ n'est pas significative (p=0,81). Nous allons tenter d'expliquer plus loin cette différence de durée avec nos données acoustiques et articulatoires. Notre hypothèse est que cette différence de durée est due au chevauchement des gestes articulatoires entre les attaques bilabiales /p, pʰ, m/ et le noyau /z/. Notre critère de segmentation joue aussi un rôle important dans cette différence de durée.

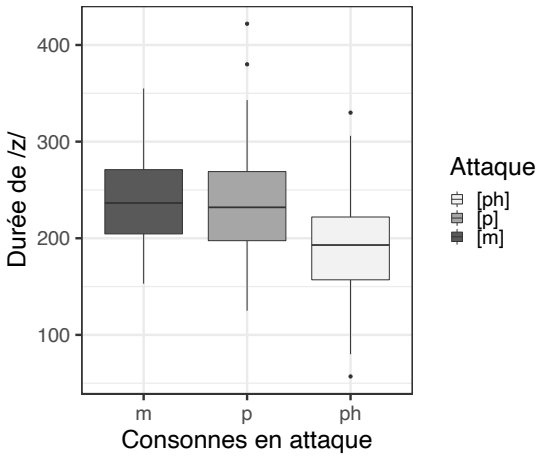


FIGURE 3 : Durée (ms) de la voyelle apicale /z/ dans 3 contextes consonantiques différents /p, pʰ, m/.

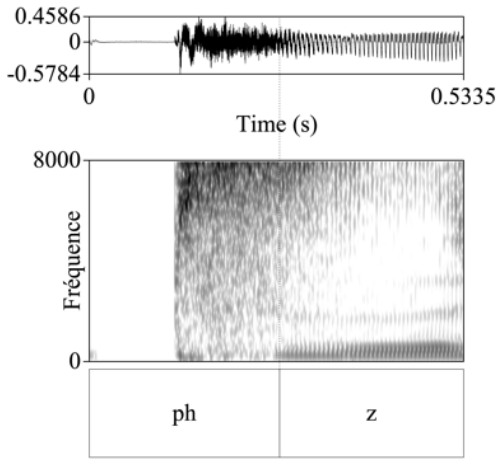


FIGURE 4 : Le signal acoustique et le spectrogramme de la forme /pʰzɹ̩/, produite par FS1.

4.2 Le chevauchement des gestes : données acoustiques et articulatoires

La voyelle apicale a le geste articulatoire d'une fricative alvéolaire (Shao & Ridouane, 2019) et cette cible serait atteinte de manière anticipatoire, pendant la tenue de la consonne en attaque. Pour /p/ et /m/, cette réalisation anticipatoire n'a pas d'indice acoustique majeur sur le signal, puisque la cavité orale est fermée pendant ces deux consonnes, et que le relâchement de /p/ est très court. Pour /pʰ/, cependant, ce chevauchement a des conséquences acoustiques visibles : la phase du relâchement de /pʰ/ est fortement perturbée par le bruit de friction généré par le geste fricatif alvéolaire. Nos données acoustiques et articulatoires présentées ci-après permettent de l'affirmer.

Trois phases sont généralement visibles sur le signal acoustique d'une occlusive sourde aspirée quand elle est suivie d'une voyelle : (i) la barre d'explosion, sous forme de barre verticale sur le spectrogramme, d'une durée très brève ; (ii) la phase de friction, où le bruit de turbulence généré au niveau de la constriction supraglottale excite principalement la cavité devant la constriction ; et (iii) l'aspiration proprement dite, où le bruit de turbulence généré au niveau glottal excite tout le conduit vocal (Fant, 1973 ; Stevens, 1998 : 457–465 ; Ridouane, Clements & Khatiwada, 2011). La phase d'aspiration est donc définie comme de la friction glottale qui se traduit acoustiquement sous forme de structure formantique plus ou moins masquée par du bruit, d'une durée supérieure à 30–40 ms. Dans nos données, nous avons observé que les phases (i) et (iii) sont bien présentes pour /pʰ/ dans les syllabes /pʰu, pʰʉ, pʰa/. La phase (ii) n'est pas observée car la consonne /pʰ/ a une constriction bilabiale, sans cavité devant la constriction. Une configuration différente a été observée pour /pʰ/ quand il est suivi de la voyelle apicale, avec un bruit de friction en hautes fréquences pendant la phase du relâchement de cette consonne. Ce bruit de friction continue jusqu'au noyau /z/ (voir FIGURE 4). La raison de la présence de ce bruit de friction est que la phase d'ouverture glottale pendant le relâchement de /pʰ/ coïncide avec une constriction supraglottale étroite correspondant à la tenue de /z/. La conséquence de ce chevauchement est que le bruit de friction supraglottale généré par /z/ domine le bruit glottal de /pʰ/, donnant ainsi lieu à une phase qui a des caractéristiques acoustiques ressemblant à celles d'une consonne /s/.

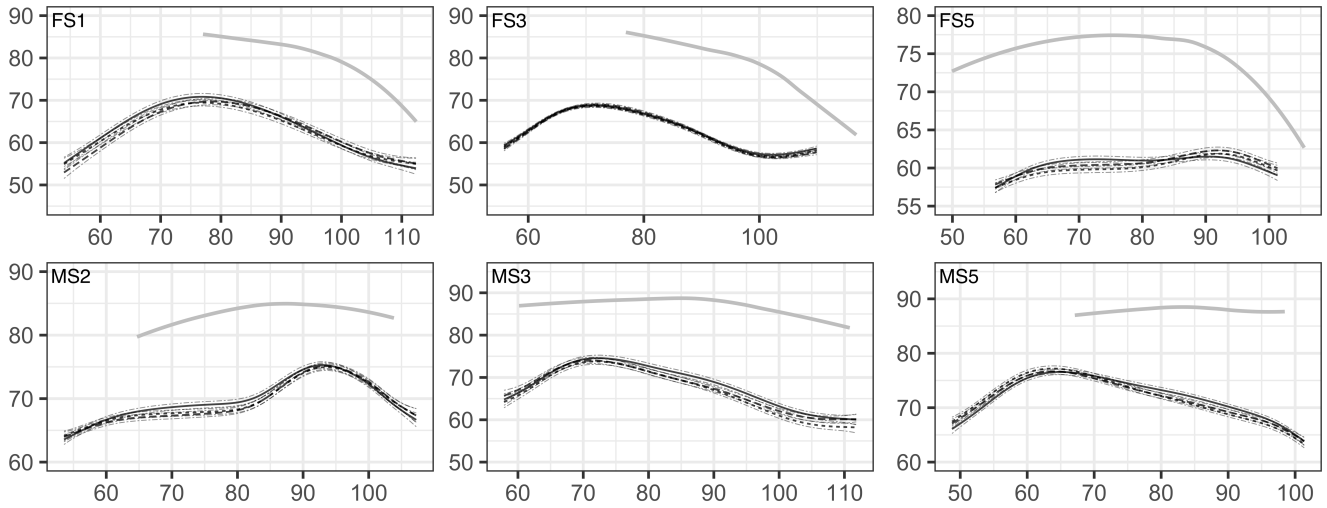


FIGURE 5 : Contours de langue pour les six locuteurs en coupe sagittale, extraits en coordonnées (mm) x/y à l'image d'ultrason correspondant au milieu de trois phases de la syllabe /pʰz/. Les tracés pleins, pointillés, et discontinus représentent le milieu de la phase d'occlusion de /pʰ/, le milieu du relâchement de /pʰ/ et le milieu de /z/, respectivement. Les contours ont été généralisés en SS ANOVA avec l'intervalle de confiance bayésienne à 95%. Les tracés gris et pleins représentent le palais. La pointe de la langue est à droite.

Les données articulatoires ont permis une observation directe de la nature du chevauchement des gestes entre les consonnes labiales et la voyelle apicale /z/. Nous présentons ici le cas de la syllabe /pʰz/ pour illustration. Comme le montre la FIGURE 5, la cible articulatoire pour la voyelle apicale /z/ est déjà atteinte pendant la phase d'occlusion de l'attaque /pʰ/. Une fois cette cible atteinte, la forme de la langue ne change plus jusqu'au milieu de la voyelle apicale. Le même phénomène a été observé pour les formes /pz/ et /mz/. La FIGURE 5 montre aussi que la configuration linguale pour /z/ ressemble quasi à l'identique à celle de la consonne fricative /s/, les locuteurs appliquant les mêmes stratégies articulatoires pour les deux segments. Une telle configuration dans le contexte des labiales montre que le geste de fricative alvéolaire est inhérent à la voyelle apicale /z/.

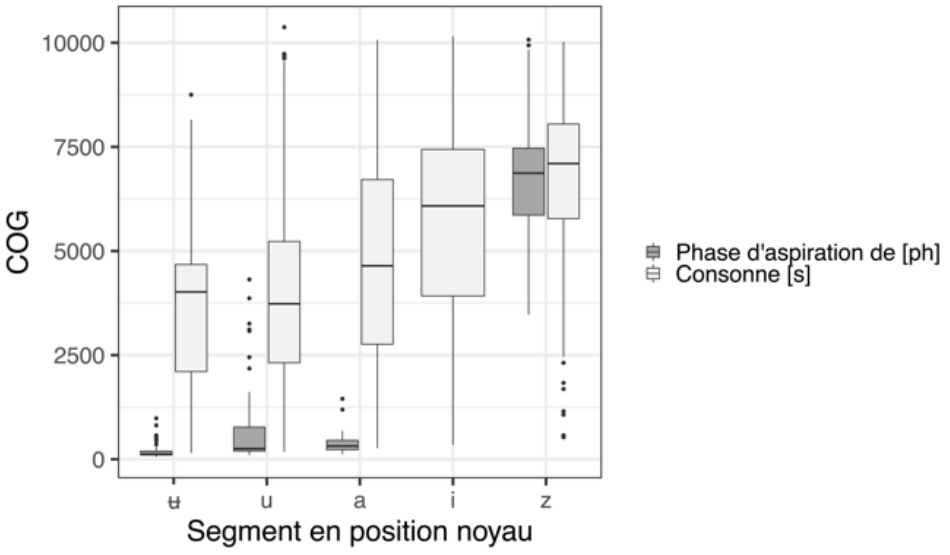


FIGURE 6 : Centre de gravité (Hz) de la consonne /s/ et de la phase de relâchement de la consonne /pʰ/ dans différents contextes vocaliques. Les données pour /pʰ/ sont obtenues avec les syllabes /pʰi, pʰu, pʰʉ, pʰa, pʰz/, et celles de /s/ sont obtenues avec les syllabes /si, su, sʉ, sa, sz/ (/pʰ/ n'apparait pas devant /i/).

Le chevauchement entre le geste lingual et le geste labial induit un effet sur les caractéristiques acoustiques du relâchement de /pʰ/. Pour confirmer qualitativement que la phase de relâchement de cette occlusive labiale a les caractéristiques acoustiques d'une fricative /s/, nous avons comparé le COG du relâchement de /pʰ/ à celui de /s/. Comme le montre la FIGURE 6, le COG de la consonne /s/ est toujours plus élevé que celui de la phase d'aspiration de /pʰ/ dans le même contexte, sauf quand ces deux consonnes sont suivies de la voyelle apicale /z/. Dans ce dernier cas, le COG de la consonne /s/ et la phase de relâchement de /pʰ/ ne présentent pas de différence significative (t(164,54)=0,19, p=0,85).

Le chevauchement entre le geste lingual et le geste labial, couplé avec le critère de segmentation que nous avions choisi, peut expliquer la durée plus courte de /z/ dans la syllabe /pʰz/. Pour rappel, nous avons choisi le début du voisement de /z/ comme onset de ce segment. Puisque la phase d'aspiration de /pʰ/ est réalisée comme un bruit de friction supraglottale et que ce bruit continue jusqu'au noyau /z/, il a été particulièrement difficile de déterminer la frontière exacte entre cette phase et la voyelle apicale /z/ (voir FIGURE 4). Les données articulatoires, montrant que la langue est déjà dans une configuration apicale pendant la tenue des consonnes labiales, suggèrent ainsi que l'onset du noyau /z/ dans la forme /pʰz/ pourrait avoir commencé bien avant le premier pulse détecté par Praat (*i.e.* pendant la phase de relâchement). Dans ce contexte, l'ouverture glottale caractéristique des occlusives aspirées est parmi les principales causes du dévoisement de /z/.

5 Conclusion

Cette étude, basée sur des données acoustiques et articulatoires, montre que la voyelle apicale en CJ a le geste articulatoire d'une fricative alvéolaire. La configuration linguale de /z/, observée dans un contexte bilabial minimisant les effets de la coarticulation, montre que le geste fricatif alvéolaire de /z/ fait bien partie de son articulation. Ce geste lingual est réalisé de manière anticipatoire durant la tenue des consonnes labiales dans les syllabes /pz, mz, pʰz/. Une conséquence de cette configuration est que la phase de relâchement de /pʰ/ est réalisée avec des caractéristiques acoustiques semblables à celles d'une fricative /s/.

Remerciements

Nous tenons à remercier tous les locuteurs qui ont participé à l'acquisition des données, ainsi que les trois relecteurs anonymes pour leurs nombreux commentaires et suggestions. Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083. Il contribue à l'IdEx Université de Paris - ANR-18-IDEX-0001.

Références

ARTICULATE INSTRUMENTS Ltd. (2008). *Ultrasound Stabilisation Headset User's Manual: Revision 1.5.* Edinburgh, UK: Articulate Instruments Ltd.

ARTICULATE INSTRUMENTS Ltd. (2012). *Articulate Assistant Advanced User Guide: Version 2.14.* Edinburgh, UK: Articulate Instruments Ltd.

DAVIDSON L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of the Acoustical Society of America* 120(1), 407–415.

DELL F. (1994). Consonnes à prolongement syllabique en Chine. *Cahiers de linguistique-Asie orientale* 23(1), 87-94.

DUANMU S. (2007). *The phonology of standard Chinese*. New York : Oxford University Press.

FANT G. (1973). Stops in CV syllables. In *Speech Sounds and Features*, 110–139. Cambridge, MA : MIT Press.

GU C. (2014). Smoothing Spline ANOVA Models: R Package gss. *Journal of Statistical Software* 58(5), 1–25. DOI : 10.18637/jss.v058.i05.

HARTMAN L. M. (1944). The segmental phonemes of the Peiping dialect. *Language* 20(1), 28–42.

HIRATA S. (1998). *Huizhou Fangyan Yanjiu* [Etude sur les dialectes du Huizhou]. Tokyo : Kohbun Press.

HO, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica* 33(5), 353–367.

HOWIE, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. New York: Cambridge University Press.

KARLGREN B. (1915). *Etudes sur la phonologie chinoise*. Uppsala : KW Appelberg.

LADEFOGED P. & MADDIESON I. (1996). *The sounds of the world's languages*. Oxford & Malden, MA : Blackwell.

LEE W.-S. & ZEE E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33(1), 109–112.

LEE-KIM S.-I. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association* 44(3), 261–282.

RIDOUANE R., CLEMENTS G. N. & KHATIWADA R. (2011). Language-independent bases of distinctive features. In J. A. Goldsmith, E. Hume, L. Wetzels, Éds., *Tones and Features: Phonetic and Phonological Perspectives,* 264–287.

SHAO B. & RIDOUANE R. (2018). La « voyelle apicale » en chinois de Jixi : caractéristiques acoustiques et comportement phonologique. In M. COOKE, B. BIGI & J. LAVAUD, Éds., *Actes de XXXIIe Journées d'Études sur la Parole*, p. 685–693. Aix-en-Provence, France.

SHAO B. & RIDOUANE R. (2019) Apical vowels in Jixi-Hui Chinese: an articulatory study. In S. Calhoun, P. Escudero, M. Tabain & P. Warren, Éds., *Proceedings of the 19th International Congress of Phonetic Sciences*, p. 2358–2362. Melbourne, Australia.

STEVENS K. N. (1998). *Acoustic Phonetics*. Cambridge, MA : MIT Press.

TROUBETZKOY N. S. (1949). *Principes de Phonologie*. Paris: Librairie C. Klincksieck.

YU A. (1999). Aerodynamic constraints on sound change: The case of syllabic sibilants. *Journal of the Acoustical Society of America* 105(2), 1096–1097.

ZHAO R. (2003). *Jixi Fangyan Cidian* [Dictionnaire du dialecte jixi]. Nanjing : Jiangsu Jiaoyu Chubanshe.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Symbolisme phonétique du genre dans les prénoms français

Alexandre Suire[1] Alba Bossoms Mesa[2] Michel Raymond[3] Melissa Barkat-Defradas[3]
(1) Université des Ryukyus, 903-0213 Okinawa, Japon
(2) Max Planck Institute for Evolutionary Anthroplogy, 04103 Leipzig, Allemagne
(3) Institut des Sciences de l'Évolution de Montpellier, 34095 Montpellier, France
alexandresuire@umontpellier.fr, alba.bossoms@gmail.com, michel.raymond@umontpellier.fr, melissa.barkat-defradas@umontpellier.fr

RESUME

Le symbolisme phonétique suggère un lien naturel entre les sons et la signification d'un mot. Les prénoms constituent d'excellents candidats afin d'étudier ces relations selon les prédictions de la théorie « *code-fréquence* », selon laquelle les sons de basses fréquences sont perceptivement associés à une large corpulence et par extension à la masculinité, tandis que les sons de hautes fréquences sont associés à la petitesse et à la féminité. En analysant les prénoms français attribués entre 1900 et 2009, nous avons confirmé ces prédictions en observant une différence significative de la qualité de la voyelle sur la syllabe perceptivement proéminente : les prénoms masculins exhibent plus fréquemment des voyelles de basses fréquences (e.g. /o/) tandis que les prénoms féminins attestent plus souvent des voyelles de hautes fréquences (e.g. /i/).

ABSTRACT

Sex-biased sound symbolism in French first names
Sound symbolism states that there is a natural link between the sound and the meaning of a word. First names are good candidates to study these relationships under the predictions of the "*frequency-code*" hypothesis, according to which sounds of low frequencies are associated to a large body-size and by extension to masculinity, while sounds of high frequencies are linked to smallness and therefore to femininity. Using a database of French first names from 1900 to 2009, we confirmed those predictions by observing a significant difference in vowel quality in the stressed syllable: male names are more likely to include lower-frequency vowels (e.g. /o/) whereas female names higher-frequency vowels (e.g. /i/).

MOTS-CLES : Symbolisme phonétique ; prénoms ; féminité ; masculinité ; voix.
KEYWORDS: Sound symbolism; first names; femininity; masculinity; voice.

Volume 1 : Journées d'Études sur la Parole, pages 588–596.
HAL : hal-02798590.

1 Introduction

Contrairement au principe d'arbitrarité du signe linguistique selon lequel il n'existe aucun rapport naturel entre le signifié (i.e. la représentation mentale de l'objet, ou le concept) et le signifiant (i.e. l'image acoustique des mots), la notion de symbolisme phonétique suggère qu'il existe une relation naturelle et motivée entre la forme sonore des signes linguistiques et leurs sens. Ce principe peut être illustré à travers l'expérience classique de Köhler (1929). Dans cette étude, l'auteur a montré que les sujets avaient tendance à associer les séquences sonores « *takete* » et « *baluba* » respectivement à une forme angulaire et circulaire. Bien plus tard, Ramachandran & Hubbard (2001) ont répliqué cette expérience avec la paire de non-mots « *bouba* » et « *kiki* » et ont montré que plus de 90% des sujets anglophones américains et tamouls (locuteurs du Tamil, langue dravidienne du sud de l'Inde) associaient la forme arrondie à la première séquence sonore et la forme étoilée à la seconde. Cette dernière décennie a vu un nombre croissant d'études mettre en évidence que ce principe d'association est effectif dans de très nombreuses langues naturelles, même lorsqu'elles sont phylogénétiquement très éloignées. En effet, dans une étude exploratoire sur plus de 4000 langues actuelles appartenant à plus de 350 familles linguistiques différentes, une importante similarité dans les associations sons-sens a été observée (Blasi et al., 2016). Par exemple, les mots renvoyant au concept de « *petitesse* » contiennent fréquemment la voyelle haute /i/ (e.g. « *petit* » en français, « *saghiir* » en arabe, « *maliit* » en phillipin), le mot correspondant au « *nez* » contient lui-même souvent la consonne nasale /n/ (e.g. « *nos* » en bosnien, « *nose* » en anglais, « *näsa* » en suédois), de même que « *langue* » (ici l'organe) contient souvent la consonne latérale alvéolaire /l/ (e.g. « *lleengua* » en catalan, « *liežuvis* » en lituanien, « *dil* » en Turc).

Plusieurs mécanismes peuvent expliquer ces associations sons-sens (pour une revue exhaustive, voir Sidhu & Pexman, 2018). L'un d'entre eux repose sur les propriétés articulatoires et acoustiques des voyelles et des consonnes. En effet, la théorie du « *code-fréquence* » telle que formulée par Ohala (1984) suggère que les mots renvoyant au concept ou à l'image mentale de « *petitesse* » présentent une incidence élevée de voyelles et/ou de consonnes caractérisées par des fréquences acoustiques élevées (i.e. sons aigus). A l'inverse, les mots désignant la « *grandeur* » et/ou la « *largeur* » contiennent plus souvent des voyelles et des consonnes caractérisées par des fréquences acoustiques basses (i.e. sons graves). Plusieurs exemples attestant de cet état de fait sont disponibles en anglais (« *teeny* », « *wee* », « *itsy bitsy* » vs. « *large* », « *humongous* »), en français (« *petit* », « *fin* », « *mince* » vs. « *grand* », « *large* », « *gros* »), en grec (« *mikros* » vs. « *makros* ») ou encore en japonais (« *tchiisai* » vs. « *ooki* »). En outre, Ohala (1984) suggère que les fondements du symbolisme phonétique proviennent des associations vocalisations-perceptions observées dans de nombreuses espèces (y compris l'humain) dans lesquelles l'agresseur produit des vocalisations de basses fréquences pour apparaître plus large, grand, menaçant et dominant, à l'inverse du subordonné qui produit des vocalisations de hautes fréquences afin de paraître plus petit et soumis. Ohala (1984) suggère ainsi que les mots incluant des sons de hautes vs. de basses fréquences pourraient également être utilisés dans les langues naturelles pour projeter ces mêmes impressions et traduire au plan phonétique des impressions de masculinité vs. féminité.

Les prénoms nous semblent ainsi constituer d'excellents candidats pour étudier le symbolisme phonétique à la lumière de ce cadre théorique. En effet, si les prédictions établies par la théorie du « *code-fréquence* » s'avèrent validées, nous pouvons avancer l'hypothèse que les caractéristiques sonores des prénoms attribués aux petits garçons vs. aux petites filles reflèteraient certaines caractéristiques morphologiques associées au dimorphisme homme/femme. Ainsi, les prénoms masculins pourraient attester une incidence plus élevée de sons produits dans les basses fréquences, du fait des impressions auditives de largeur, de masculinité et de dominance associées aux voix graves, caractéristiques attractives bien établies chez les hommes (e.g. Pisanski & Rendall, 2011). A l'inverse, les prénoms féminins pourraient faire état d'une incidence plus élevée de sons de hautes fréquences, via les liens observés entre voix aigues, petitesse, féminité et « *soumission* » chez les femmes (du moins tels que suggérés par une majorité des études, e.g. Borkowska & Pawlowski, 2011). Bien que les contraintes culturelles, la tradition familiale, la mode et l'euphonie jouent également un rôle primordial dans le choix des prénoms que les parents attribuent à leurs enfants, ces facteurs ne déterminent que le choix à l'intérieur d'un ensemble essentiellement marqué pour chaque sexe.

La présente étude se propose ainsi d'examiner ces relations dans les prénoms français. Enfin, nous souhaitons aussi souligner que le symbolisme phonétique est une thématique de recherche qui a été largement ignorée dans la linguistique française. De plus, les quelques rares études menées jusque-là ont été qualitatives plutôt que quantitatives (e.g. Chastaing, 1964; Bidaud, 2017)[1]. Actuellement, il existe donc une faible connaissance des sons potentiellement porteurs de symbolisme dans la langue française. Notre étude se propose ainsi de contribuer à l'étude du symbolisme phonétique et d'en encourager l'intérêt.

2 Matériels et Méthodes

2.1 Traitement des données

Les données ont été collectées en septembre 2014 auprès de l'Institut National de la Statistique et des Études Économiques. Nous avons sélectionné les 100 prénoms féminins et les 100 prénoms masculins les plus populaires pour chaque décennie, allant de 1900-1909 à 2000-2009[2]. Afin de contrôler par la taille de la population, la popularité a été estimée en calculant le classement annuel de chaque prénom et en les additionnant par décennie. Bien que cette approche exclue les prénoms rares, elle capture correctement les pratiques de dénomination pour une décennie donnée (Pitcher et al., 2013). Tous les prénoms récupérés ont ensuite été transcrits de manière indépendante par deux phonéticiens

1 A notre connaissance, hormis la nôtre, il n'existe qu'une seule autre étude quantitative ayant été menée sur un autre corpus onomastique dans la langue française, celui des noms vernaculaires d'animaux par De Carolis et al. (2017).

2 Nous avons choisi de travailler par décennie plutôt que par année afin d'obtenir une tendance plus générale des processus de dénomination.

francophones natifs, conformément aux principes de l'Alphabet Phonétique International. En l'absence d'agrément sur certaines transcriptions ou lorsque la prononciation était inconnue, différentes sources Web ont été utilisées (e.g. https://fr.wiktionary.org/wiki). Pour chaque syllabe d'un prénom, nous avons noté les caractéristiques articulatoires suivantes :

- Le point d'articulation. Il correspond à la position de la langue dans la cavité buccale. Nous avons ainsi distingué les voyelles antérieures (i.e. /i/, /y/, /e/, /ɛ/, /ø/), les voyelles centrales (i.e. /a/, /ə/) et les voyelles postérieures (i.e. /u/, /o/, /ɔ/). Le mouvement de la langue en position avant ou arrière entraîne une modification globale du tractus vocal et plus précisément de la taille de la cavité buccale. Sur le plan acoustique, le timbre d'un son dépend de la répartition des fréquences dans son spectre (représentation fréquence/amplitude d'un son). La position de la langue dans la cavité buccale a ainsi une incidence sur la valeur du deuxième formant (i.e. bandes de fréquences renforcées en fonction de la forme et du volume des résonateurs supraglottiques). Les voyelles antérieures sont ainsi caractérisées par un F2 élevé (e.g. /i/ ≈ 2000Hz), tandis que les voyelles postérieures laissent place à une cavité buccale plus ample et ont un F2 bas (e.g. /u/ ≈ 750Hz, Meunier, 2007). Ainsi, nous nous attendons à ce que les voyelles antérieures vs. postérieures soient respectivement associées aux prénoms féminins en raison de leurs F2 situés dans les hautes fréquences vs. masculins dont un maximum d'amplitude est observé dans les basses fréquences.
- La nasalité de la voyelle. Les voyelles nasales sont des voyelles articulées avec le voile du palais abaissé, à l'arrière de la cavité buccale, par opposition aux voyelles orales pour lesquelles le voile du palais est relevé. Le bourdonnement créé par les vibrations des plis vocaux entre en résonnance dans le conduit vocal, composé pour les voyelles nasales à la fois de la cavité orale et de la cavité nasale et uniquement de la cavité orale pour les voyelles orales. Le français en comporte quatre : /ɛ̃/, /œ̃/, /ɔ̃/, /ɑ̃/. Selon Passy (cité par Amelot, 2004 :22), « *le timbre de la voyelle nasale semble plus grave que celui de la voyelle orale correspondante* ». Les études perceptives conduites par Delvaux et al., (2002) ont en effet confirmé expérimentalement que les voyelles nasales sont perçues comme plus graves que les voyelles orales. Cette impression auditive a été confirmé à travers les mesures acoustiques réalisées par Delvaux (2012) qui a montré que pour les quatre voyelles nasales du français la balance spectrale est déplacée en faveur des basses fréquences (sous 2000 Hz) par rapport aux orales correspondantes. En conséquence, nous nous attendons donc à ce que les voyelles nasales aient plutôt tendance à être associées aux prénoms masculins.
- Le mode d'articulation de la consonne. Il est déterminé par la façon dont le flux d'air s'échappe du tractus vocal pendant l'articulation. Ici, nous nous sommes concentrés sur les occlusives (fermeture complète du flux d'air avant un relâchement soudain) vs. les fricatives (rétrécissement du tractus vocal en un point provoquant une perturbation de l'écoulement du flux d'air). Nous avons également noté si les consonnes étaient voisées ou non, respectivement pour les occlusives /b/, /d/, /g/ vs. /p/, /t/, /k/ et les fricatives /v/, /z/, /ʒ/ vs. /f/, /s/, /ʃ/. Sachant que l'articulation d'une occlusive voisée provoque plus de bruit dans les basses fréquences que celle d'une fricative non-voisée (Stevens, 1998), nous nous attendons à retrouver plus fréquemment des occlusives voisées dans les prénoms masculins.

La raison pour laquelle nous avons choisi d'analyser les phonèmes précédemment cités plutôt que d'autres (e.g. consonnes nasales, approximantes) est que la question de leur phono-symbolisme a été plutôt bien étudiée dans la littérature (Nielsen & Rendall, 2013). Enfin, notons que si des différences phonétiques sont observables entre les prénoms masculins et féminins, celles-ci doivent être situées sur la syllabe perceptivement proéminente puisque c'est celle-ci qui permet aux locuteurs d'inférer de manière efficace l'information lexicale. Dans ce contexte, en tant que langue syllabique, la prosodie française s'organise autour d'un accent rythmique régulièrement affecté à la dernière syllabe du mot (Di Cristo, 1998). Afin de vérifier cette hypothèse, nous avons donc analysé la première et la dernière syllabe de chaque prénom, avec la prédiction d'observer une différence phonétique sur cette dernière.

2.2 Analyses statistiques

Afin de tester nos prédictions, nous avons agrégé tous les prénoms s'étalant sur le siècle dernier afin d'établir une liste unique de prénoms (e.g. « *Marie* » est retrouvée dans plusieurs décennies). Pour chaque sexe, une seule version des prénoms phonétiquement équivalents a été récupérée (e.g. « *Danielle* » et « *Danièle* », homophones non homographes). Les prénoms composés tels que « *Jean-Marie* » et « *Marie-Pierre* » ont été écartés de l'analyse car ils représentent un ensemble particulier composé d'un prénom masculins associé à un prénom féminin. Les prénoms monosyllabiques ont également été écartés de l'analyse afin de pouvoir comparer la première et la dernière syllabe. Les prénoms épicènes homographes et non homographes ont été conservés. Au total, notre échantillon est constitué de 274 prénoms féminins et 197 prénoms masculins uniques répartis sur le siècle.

Un modèle linéaire généralisé a ensuite été utilisé pour étudier l'existence de schémas symboliques au plan phonétique en fonction du genre dans les prénoms masculins et féminins français. La variable de réponse « *genre* » étant binaire, une distribution binomiale avec une fonction de lien logit a été spécifiée. Les variables explicatives sont les caractéristiques articulatoires précédemment mentionnées. Étant donné notre prédiction concernant la syllabe accentuée, chacun des prédicteurs a été répété pour la première et la dernière syllabe. La significativité de chaque variable a été évaluée en comparant le modèle sans celle-ci vs. le modèle avec celle-ci (ANOVA type III). Des comparaisons post-hoc (test de Tukey) avec une correction de Bonferroni ont été effectuées pour le point d'articulation de la voyelle afin d'évaluer les différences entre les sexes. Enfin, un effet de taille a été calculé en utilisant le *f2* de Cohen.

3 Résultats

Un tableau décrivant la fréquence d'occurrence des différents phonèmes cibles est donné en matériel supplémentaire. Dans la dernière syllabe, la masculinité d'un prénom est significativement liée au point d'articulation de la voyelle (χ_2^2 = 11.82, p < .01), la nasalité (χ_1^2 = 65.41, p < .001) et les fricatives non-voisées (χ_1^2 = 13.23, p < .001). Précisément, les prénoms masculins exhibent plus souvent des voyelles postérieures telles que /o/ ou /ɔ/ (e.g. « *Enzo* », « *Léopold* »), au lieu de voyelles antérieures ou centrales telles que /i/, /y/ ou /a/ (respectivement t = 1.17, p < .01 ; t = 1,35, p < .01 ;

e.g. « *Jackie* », « *Auguste* », « *Bernard* »). Bien que les voyelles postérieures puissent être trouvées dans les prénoms féminins (e.g. « *Margot* », « *Nicole* »), les voyelles antérieures telles que /i/, /ɛ/ et la voyelle centrale /a/ sont plus courantes (e.g. « *Émilie* », « *Hélène* », « *Léa* »). Les prénoms masculins sont également plus susceptibles de contenir des voyelles nasales telles que /ã/ ou /ɔ̃/ (e.g. « *Roland* », « *Raymond* », contre-exemples féminins : « *Fernande* », « *Marion* ») et des fricatives non-voisées telles que /s/ ou /ʃ/ (e.g. « *Fabrice* », « *Michel* » ; contre-exemples féminins : « *Clémence* », « *Rachel* »). Les probabilités d'être un prénom masculin en fonction du point d'articulation de la voyelle et de la nasalité sont données dans la Figure 1.

De manière imprévue, dans la première syllabe, les prénoms masculins exhibent une plus grande incidence d'occlusives voisées (χ_1^2 = 12.59, p < .001) telles que /b/, /d/ ou /g/ (e.g. « *Bernard* », « *Dimitri* », « *Gustave* » ; contre-exemples féminins : « *Brigitte* », « *Deborah* », « *Gwenaëlle* »). Dans la première syllabe, le lieu d'articulation de la voyelle et la nasalité ne différaient pas entre les sexes, pas plus que le nombre de fricatives non-voisées (toutes p > .05). Les caractéristiques articulatoires expliquent 14% de la variation des différences entre les sexes et le *f2* de Cohen suggère un effet de taille modéré (0.17).

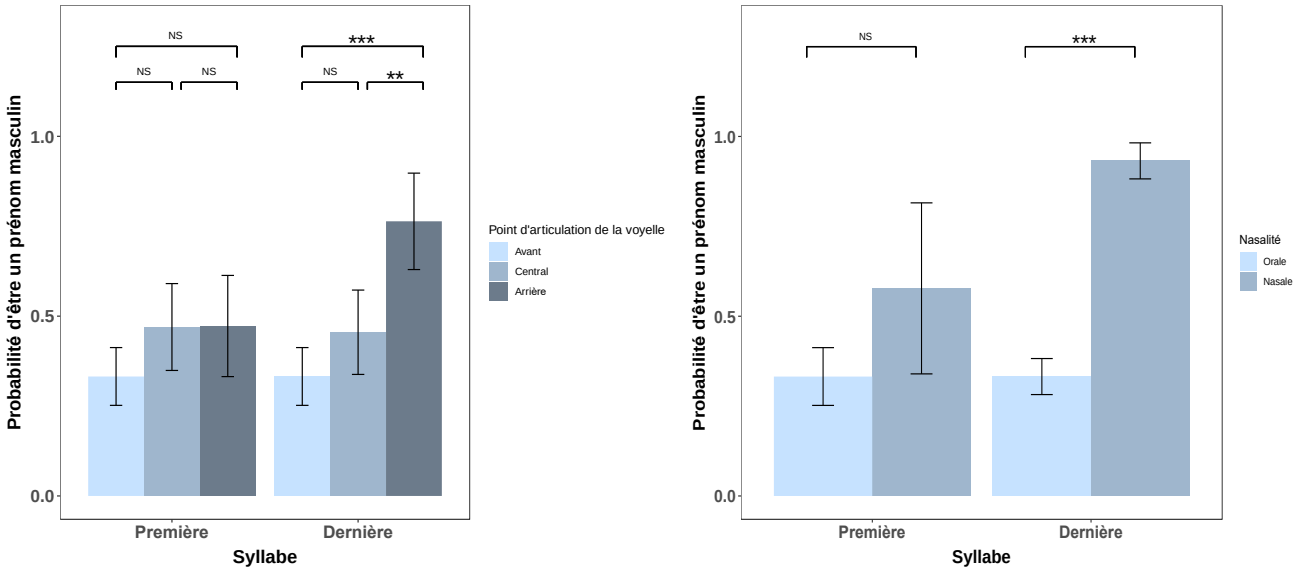


FIGURE 1 : Estimations du modèle linéaire généralisé, log-transformées pour obtenir les probabilités qu'un prénom soit masculin en fonction du lieu d'articulation de la voyelle orale (gauche) et de la nasalité (droite). Les histogrammes représentent la probabilité moyenne associée à des intervalles de confiance à 95%. Code de signification des comparaisons post-hoc : *** p < .001 ; ** p < .01; « NS » non significatif.

4 Discussion

Nos résultats mettent en évidence des différences significatives dans la composition sonore des prénoms masculins et féminins français.

Ces schémas phonétiquement symboliques peuvent être interprétés à la lumière de ceux qui ont été mis en évidence dans la composition sonore des prénoms anglophones. Par exemple, Cutler et al. (1990) ont montré que les prénoms féminins anglophones sont plus susceptibles de contenir la voyelle haute /i/ dans la première syllabe des prénoms dissyllabiques (e.g., « *Michelle* », « *Tina* ») et/ou sur la deuxième syllabe (e.g. « *Christine* », « *Elizabeth* », « *Patricia* »). Cette étude a également révélé que les voyelles de basses fréquences (voyelles postérieures arrondies comme /u/ ou /o/) sont beaucoup moins fréquentes dans les prénoms féminins. Ces résultats corroborent ceux de Pitcher et al. (2013) qui ont procédé à une analyse similaire sur un corpus de prénoms britanniques, australiens et américains les plus populaires entre 2001 et 2010. Les auteurs ont montré que les voyelles antérieures telles que /i/ ou /e/ (i.e. voyelles hautes) sont principalement attestées dans les prénoms féminins (e.g. « *Emily* ») et les voyelles postérieures telles que /u/ ou /o/ (i.e. voyelles basses) dans les prénoms masculins (e.g. « *Thomas* »). Dans ce contexte, bien que nous observons une distribution comparable du timbre des voyelles entre prénoms masculins vs. féminins en français et en anglais, notre étude révèle que, s'il existe une traduction phonétique du dimorphisme sexuel sur les prénoms, cette information est dépendante du système accentuel de la langue concernée. En effet, les voyelles « *sexuellement* » marquées (i.e. voyelles hautes-prénoms féminins vs. voyelles basses-prénoms masculins) sont régulièrement incluses dans la syllabe accentuée, perceptivement proéminente : la première syllabe des noms dissyllabiques ou sur la deuxième syllabe des trissylabiques en anglais vs. systématiquement la dernière en français.

De manière intéressante, nos résultats sur les occlusives voisées rejoignent ceux de Slepian & Galinsky (2016), qui ont établi qu'en anglais et en tamoul, les prénoms masculins ont plus souvent tendance à exhiber en position initiale une consonne voisée (i.e. par définition plus grave que leurs pendants non voisés en raison de la présence d'une activité glottique) comparativement aux prénoms féminins. Ils ont également montré que la présence d'un phonème voisé en position initiale du prénom augmente significativement la perception de masculinité chez les locuteurs de chaque langue, et ceci indépendamment du genre communément associé à celui-ci. Dans ce contexte, l'ensemble des différences phonétiques permettrait aux locuteurs d'une langue et/ou d'une culture donnée d'inférer rapidement et avec justesse le genre d'un prénom.

Au-delà des aspects phonétiques, les prénoms pourraient avoir une influence plus large sur la perception de leur porteur. En effet, quelques études suggèrent que la composition sonore d'un prénom peut influencer la perception physique que l'on se fait d'une personne. Par exemple, deux études ont examiné la façon dont certains phonèmes spécifiques aux prénoms allemands et anglais peuvent influencer le jugement de l'attractivité faciale. Dans le cas des prénoms allemands, les visages masculins sont perçus comme plus attractifs lorsqu'ils sont associés à des prénoms possédant des

voyelles postérieures (Klenovsak et al., 2018). A l'inverse, les prénoms féminins composés de voyelles antérieures augmentent l'attractivité perçue d'un visage féminin. De manière intéressante, des résultats contraires ont été trouvés pour les prénoms anglais (Perfors, 2004). Au-delà du visage, il a été montré que certains phonèmes peuvent aussi influencer la perception de la forme du corps. Par exemple, en considérant les consonnes /b/, /l/, /m/, /n/ et les voyelles /u/, /o/ et /ɒ/ comme des voyelles « *rondes* » (analogie liée à leur propriété articulatoire impliquant les lèvres, voyelles arrondies), et les consonnes /k/, /p/ et /t/ et les voyelles /i/, /e/, /ɛ/ et /ʌ/ comme « *pointues* », Sidhu & Pexman (2015) ont montré que des sujets canadiens sont plus susceptibles d'associer une silhouette ronde avec un prénom contenant des sons arrondis (i.e. labialisés ; e.g. « *Mo͟lly* », « *Bo͟b* »), et à l'inverse une silhouette plutôt fine et longue avec des prénoms composés de sons dits « *pointus* » (e.g. « *Kate* », « *Kirk* »). Enfin, en utilisant la même nomenclature de phonèmes, Barton & Halberstadt (2018) ont montré que des sujets américains considèrent les prénoms contenant des voyelles « *rondes* » et « *pointues* » comme convenant respectivement mieux à des visages ronds vs. anguleux. De manière plus intéressante encore, les auteurs ont également montré que les candidats aux élections sénatoriales américaines gagnaient 10% de vote en plus sur la seule base d'une congruence prénom-visage très forte.

En conclusion, notre étude est l'une des rares à entreprendre l'analyse quantitative du symbolisme phonétique en français. Elle ouvre des pistes de recherche originales en lien avec l'influence de la composition sonore des prénoms sur les perceptions de masculinité et de féminité. De plus, l'étude du symbolisme phonétique menée dans une perspective transculturelle permet d'aborder de manière séduisante la question des universaux vs. spécificités linguistiques. Enfin, nous souhaitons à travers ce premier travail encourager la recherche sur le symbolisme phonétique en français afin de questionner expérimentalement, et à partir de lexiques aussi divers que variés, la question de l'arbitrarité du signe.

Références

AMELOT. A., (2004). Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasalesdu français. *Thèse de doctorat nouveau régime en Linguistique*. Université de la Sorbonne nouvelle - Paris III.

BARTON, D. N., & HALBERSTADT, J. (2018). A social Bouba/Kiki effect : A bias for people whose names match their faces. *Psychonomic Bulletin & Review*, 25(3), 1013-020.
DOI : <10.3758/s13423-017-1304-x>

BIDAUD, S. (2017). Le phonosymbolisme des morphèmes du français. *Travaux de linguistique*, 75(2), 81-100. DOI : <10.3917/tl.075.0081>

BLASI, D. E., WICHMANN, S., HAMMARSTRÖM, H., STADLER, P. F., & CHRISTIANSEN, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 10818-10823. DOI : <10.1073/pnas.1605782113>

BORKOWSKA, B., & PAWLOWSKI, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1), 55-59.
DOI : <10.1016/j.anbehav.2011.03.024>

CHASTAING, M. (1964). *Nouvelles recherches sur le symbolisme des voyelles*. Presses universitaires de France: Paris.
CUTLER, A., MCQUEEN, J., & ROBINSON, K. (1990). Elizabeth and John : Sound patterns of men's and women's names. *Journal of Linguistics,* 26(02), 471.
DOI : <10.1017/S0022226700014754>
DE CAROLIS, L., MARSICO, E., & COUPE, C. (2017). Evolutionary roots of sound symbolism. Association tasks of animal properties with phonetic features. *Language & Communication*, *54*, 21 35.
DELVAUX, V., METENS, T., SOQUET, A. (2002). Propriétés acoustiques et articulatoires des voyelles nasales du français. *XXIVe Journées d'étude sur la parole*, Nancy, 1, 348-352.
DELVAUX,V. (2012). *Les voyelles nasales du français*. Bern, Suisse: Peter Lang B.
DI CRISTO, A. (1998). Intonation in French. In *Intonation systems : A survey of twenty languages*, 195-218, Cambridge University Press.
KLENOVSAK, D., HARTUNG, F., SANTIAGO, L., & ZAEFFERER, D. (2018). Preprint Are Tims Hot and Toms Not? In *Proceedings of the Annual Meeting of the Cognitive Science Society,* 31(31). DOI : <10.31219/osf.io/kj3b2>
KÖHLER, W. (1929). *Gestalt Psychology*. Liveright.
MEUNIER, C. (2007). Phonétique acoustique : Phonétique acoustique. Auzou P (Ed.). *Les dysarthries*, Solal, p.164-173.
NIELSEN, A. K. S., & RENDALL, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 67(2), 153-163. DOI : <10.1037/a0030553>
PERFORS, A. (2004). What's in a Name? The effect of sound symbolism on perception of facial attractiveness. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26. e-Scholarship : item/9bq5v5c7
PASSY, P. E., (1890). Etude sur les changements phonétiques et leurs caractères généraux. Paris: *Firmin-Didot*. Vol.1. 270p.
PISANSKI, K., & RENDALL, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, 129(4), 2201-2212. DOI : <10.1121/1.3552866>
PITCHER, B. J., MESOUDI, A., & MCELLIGOTT, A. G. (2013). Sex-Biased Sound Symbolism in English-Language First Names. *PLoS ONE*, 8(6), e64825.
DOI : <10.1371/journal.pone.0064825>
RAMACHANDRAN, V. S., & HUBBARD, E. M. (2001). Synaesthesia—A Window Into Perception, Thought and Language. *Journal of Consciousness Studies*, 8(12), 3-34.
SIDHU, D. M., & PEXMAN, P. M. (2015). What's in a Name? Sound Symbolism and Gender in First Names. *PLOS ONE,* 10(5), e0126809. DOI : <10.1371/journal.pone.0126809>
SIDHU, D. M., & PEXMAN, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, 25(5), 1619-1643. DOI : <10.3758/s13423-017-1361-1>
SLEPIAN, M. L., & GALINSKY, A. D. (2016). The voiced pronunciation of initial phonemes predicts the gender of names. *Journal of Personality and Social Psychology,* 110(4), 509-527.
DOI : <10.1037/pspa0000041>
STEVENS, K. N. (1998). *Acoustic phonetics.* MIT Press.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Caractérisation des plosives finales dans des langues d'Asie : une étude multilingue du non relâchement

Thi-Thuy-Hien Tran[1], Nathalie Vallée[1], Christophe Savariaux[1], Inyoung Kim[2], Sunhee Kim[3]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
*Institute of Engineering, Univ. Grenoble Alpes
(2) NAVER LABS Europe (3) College of Education, Seoul National University, South Korea
thi-thuy-hien.tran@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr,
christophe.savariaux@gipsa-lab.fr, inyoung.kim@naverlabs.com, sunhkim@snu.ac.kr

RÉSUMÉ

Cette étude propose de caractériser le non relâchement des plosives finales /p, t, k/ de deux langues d'Asie, tonale (vietnamien) et non tonale (coréen), du point de vue aérodynamique et glottographique. Le comportement glottique (ouverture et fermeture de la glotte, position verticale du larynx) a été examiné en synchronisation avec les valeurs de débits d'air (oral et nasal) pendant les phases de la réalisation consonantique. Les résultats mettent en évidence (1) l'absence de relâchement nasal après l'occlusion de la plosive finale pouvant entraîner une baisse de la pression intraorale, (2) que le larynx s'abaisse systématiquement durant la tenue de la consonne. Cette stratégie de réalisation va dans le sens de notre hypothèse selon laquelle les plosives non relâchées sont produites avec un mécanisme permettant de diminuer la pression intraorale de manière à minimiser le coût articulatoire de la tenue de la closion avec, pour conséquence acoustique, l'absence de burst.

ABSTRACT

Characterization of Stop Consonants in Asian Languages: A two-language Study of Unreleased Final Stops

Final stops are often produced without an audible release burst in many Asian languages of different families. This work aims to characterize the production of the final stops /p, t, k/ in two Asian languages, one tonal (Vietnamese) and one non-tonal (Korean), by using aerodynamic and glottographic data. We propose a study on laryngeal behavior (glottal opening and closing, vertical laryngeal position) in synchronization with measurements of simultaneous oral and nasal airflows during consonant production. Our results show that: (1) no nasal release able to reduce the intraoral pressure follows the stop closure; (2) the larynx systematically lowers during the closure phase. This laryngeal gesture is in line with our hypothesis that unreleased stops are produced with a mechanism providing a reduction in intraoral pressure resulting in weaker articulation and no acoustic burst.

MOTS-CLÉS : plosives, non relâchement, aérodynamique, EGG, vietnamien, coréen.

KEYWORDS: unreleased final stops, aerodynamics, EGG, Vietnamese, Korean.

1 Introduction

Toutes les langues du monde possèdent sans exception dans leur système phonologique des consonnes occlusives dont certaines constituent la catégorie des plosives. L'articulation de ces dernières est caractérisée par un blocage complet de l'écoulement de l'air au niveau du conduit vocal suivi de son relâchement brusque, générant un bruit court et audible (burst). Dans certaines langues, les plosives

Volume 1 : Journées d'Études sur la Parole, pages 597–605.
HAL : hal-02798592.

en finale de syllabe peuvent être réalisées sans burst. Bien que cette réalisation dite non relâchée ne crée pas une unité phonologique supplémentaire mais une variante allophonique, elle interroge cependant les capacités auditives et les processus de traitement des unités phonologiques.

Les recherches sur les caractéristiques phonétiques des plosives en position initiale sont abondantes dans la littérature. Celles consacrées aux consonnes finales sont toutefois beaucoup moins nombreuses, sans doute en raison de réalisations similaires entre positions pré et post-vocalique dans de nombreuses langues, au moins au niveau de leur structure acoustique. Cependant les consonnes finales sont, de manière générale, souvent sujettes à la lénition (affaiblissement articulatoire et acoustique, Kingston 2008). Certains processus de lénition aboutissent par exemple à la perte de voisement pouvant entraîner des neutralisations, c'est-à-dire la fusion de catégories phonologiques, avec parfois pour conséquence une forte restriction de l'inventaire consonantique en position finale. À cet égard, les langues de l'Asie de l'Est et du Sud-Est (entre autres le vietnamien, le thaï, le coréen, le cantonais) sont particulièrement intéressantes. En effet, plusieurs types de consonnes sont effectivement présents en initiale (par ex. la plosive bilabiale sonore /b/ *vs* sourde /p/ maintenant bien présente dans les emprunts en vietnamien ; sourde /p/ *vs* sonore / b/ *vs* aspirée /p^{h}/ en thai ; ou encore lenis /p/ *vs* fortis /p*/ *vs* aspirée /p^{h}/ en coréen). Que ces langues soient tonales ou pas et quelle que soit leur famille linguistique, les oppositions phonologiques entre plosives bilabiales citées ci-dessus, que l'on retrouve en position initiale de syllabe, sont neutralisées et réalisées sous une seule forme [p˺] en finale, décrite non voisée et non relâchée. Cette observation pose les questions suivantes : (1) Quels éléments physiques, articulatoires et acoustiques, caractérisent ces consonnes non relâchées ? (2) Quel est le lien entre les étages glottique et supra-glottique dans la mesure où le non relâchement accompagne la perte du voisement ? (3) Le non relâchement peut-il être associé à une fuite nasale accompagnant la fermeture du conduit oral qui entrainerait une baisse de la pression intra-buccale ? Ce dernier phénomène est relevé par Ladefoged et Maddieson (1996 : 129) « (…) *nasal release occurs in some of the languages which are usually described as having unreleased final stops. A good example is Vietnamese. In this language, word-final stops are usually released, but the release is by lowering the velum while the oral closure is maintained, so that a short voiceless nasal is produced* ». À notre connaissance, seules quelques études ont porté sur le comportement glottique des plosives sourdes non relâchées en finale de syllabe dans les langues de l'Asie de l'Est et du Sud-Est. Ces études aboutissent à des résultats différents quant à la présence ou non de phénomènes de glottalisation accompagnant la production de ces consonnes. Iwata et collègues (1990) ont attesté la présence d'un geste de renforcement (adduction) des bandes ventriculaires et des replis ary-épiglottiques juste après la fermeture orale des plosives en cantonais et en thaï. Le même comportement a été observé pour le taiwanais et le haka, un dialecte chinois (Edmondson et al., 2011). En coréen, il a été relevé qu'un très faible degré d'ouverture de la glotte accompagnait les plosives finales (Iwata et al., 1990). Quant au vietnamien de Hanoi, Michaud (2004) a montré que /p, t, k/ ne sont pas glottalisés en finale et qu'un accolement des plis vocaux fermant la glotte n'est pas observé. Cependant, Edmondson et collègues (2010) ont relevé une importante variabilité au niveau glottique en fonction de dialectes du vietnamien. Notons que ces études sur des langues ou dialectes différents n'ont pas procédé avec les mêmes protocoles expérimentaux (utilisant l'électroglottographie (EGG), la fibroscopie ou la laryngoscopie) et n'ont pas examiné et mesuré les mêmes paramètres physiques.

2 Objectifs et hypothèses

Notre travail vise à réexaminer le **comportement glottique** *simultanément* avec le timing de **l'écoulement d'air nasal**, lors de la réalisation des plosives sourdes. Deux langues sont sélectionnées pour le présent travail. Le vietnamien a le type de structure syllabique dominant CVC et possède un système tonal. C'est une langue monosyllabique sur le plan phonologique et en partie polysyllabique

sur le plan lexical (Tran et al., 2019). Le coréen, langue non tonale, majoritairement polysyllabique, possède des syllabes de structure V, CV, VC, CVC (Iwata et al, 1990). Dans ces deux langues, les consonnes /p, t, k/ sont les seules plosives permises en finale de syllabe et les coda branchantes sont illicites. Notre étude cherche à tester les deux hypothèses suivantes. (H1) La syllabe à plosive finale est suivie par un relâchement nasal, observation faite par Ladefoged et Maddieson (1996). L'examen de cette hypothèse devra s'appuyer sur les données concernant le débit d'air nasal. La production d'une plosive nécessite que le port vélo-pharyngé soit fermé (velum relevé) durant l'obstruction buccale pour que la pression intra-orale atteigne un niveau suffisant afin de générer un relâchement audible (Ohala, 1975). S'il y a relâchement nasal, soit par affaiblissement articulatoire, soit provoqué par un mécanisme laryngé ayant pour effet de diminuer la pression intrabuccale, l'abaissement du velum devra avoir lieu après la fermeture orale. (H2) Le larynx s'abaisse pendant l'occlusion de la plosive finale. Pour vérifier cette hypothèse, il s'agira de mesurer l'ouverture et la fermeture des plis vocaux, couplés à une détermination de la position verticale du larynx (PVL). Il a effectivement été montré que la PVL a le pouvoir d'influencer les propriétés acoustiques du signal de parole ainsi que les caractéristiques physiologiques de sa production (Guzman et al., 2013). La PVL décroit systématiquement lors de la production d'une consonne sourde alors que les consonnes voisées et les voyelles sont produites avec une PVL haute (Shipp et al., 1987). Nous faisons l'hypothèse complémentaire que, dans le cas d'une consonne sourde non relâchée, la PVL devrait s'abaisser davantage que ce qui est observé pour les plosives avec burst. Si un tel phénomène est observé ici, il devrait contribuer à valider la baisse de pression intra-orale qui faciliterait la production des plosives non relâchées. Plusieurs études antérieures montrent que l'articulation d'une consonne est influencée par la position qu'elle occupe dans la syllabe (parmi d'autres Keating et al., 1999). Il s'agit aussi dans ce travail d'examiner les caractéristiques du flux d'air et les mécanismes laryngés en regard de la réalisation acoustique de ces consonnes en comparant leur production lorsqu'elles sont en finale de syllabe *vs* finale de mot. Il a été montré effectivement que le type de frontière syllabique (CV<u>C</u>#CVC *vs* CV<u>C</u>.CVC) influence la réalisation acoustique des consonnes en coda (Tran et al., 2019).

3 Méthodologie

3.1 Matériel

Les données aérodynamiques et glottographiques ont été obtenues simultanément, via le logiciel Phonedit, avec le dispositif EVA2™ (Evaluation Vocale Assistée, Société SQLab) (Ghio et Teston, 2002). Nous avons relevé simultanément les paramètres aérodynamiques (débits d'air oral et nasal) en fonction des mouvements des articulateurs ainsi que leurs corrélations avec le signal de la parole. Les oscillations des plis vocaux (mouvements d'ouverture et de fermeture de la glotte) et la position verticale du larynx (PVL) ont été enregistrées par électroglottographie (matériel EGG EG2-PCX2 avec larynx tracking, Glottal Enterprises Inc), à l'aide de deux électrodes placées de part et d'autre du larynx, le tout en synchronisation avec les mesures des débits d'air.

3.2 Corpus

Pour chaque langue, deux types de mots, monosyllabiques $CVC_{2\#}$ avec C_2 = /p, t, k/ et composés dissyllabiques CVC_2.CVC (avec syllabe CVC_2 identique pour les deux types), ont été sélectionnés dans le lexique en fonction de plusieurs paramètres : ton (selon la langue), voisement de la consonne suivante, noyau vocalique. Les mots cibles étaient ensuite insérés dans une phrase porteuse élaborée pour faciliter la segmentation des paramètres acoustiques des plosives sourdes et neutraliser au mieux des effets de contexte, au niveau segmental comme suprasegmental. Ainsi, pour le vietnamien, les

syllabes cibles CVC_2 ont été choisies telles que V = /a/, ton = montant D1 (*sắc*). Ce ton ne comporte pas de glottalisation dans le parler du Nord (Michaud, 2004). La deuxième syllabe des mots composés était sous un ton du registre soit haut soit bas. La syllabe après la syllabe cible (correspondant au mot suivant de la phrase porteuse ou à la syllabe 2 du mot composé) commençait par une consonne voisée.

Le choix d'un contexte environnant neutre qui favorise la segmentation acoustique des plosives finales en coréen est moins évident du fait que cette langue ne possède aucune plosive voisée dans son système phonologique. Quelques éléments sonores disponibles (nasales, glides, liquides) sont inadaptés à la position du segment qui suit immédiatement les plosives cibles. Les nasales /m, n, ŋ/ sont écartées pour éviter l'anticipation de l'abaissement du velum durant la réalisation des plosives (processus systématique bien connu dans cette langue, comme /sok.nɛ/ *intention caché* > [soŋ.nɛ]). Les glides /w, j, ɯ/ subissent la resyllabification qui transforme /p, t, k/ en attaque de la syllabe suivante. Quant aux liquides [l] et [ɾ] qui sont en distribution complémentaire de /l/ dans cette langue (Sung, 2005), si /l/ suit une plosive en intersyllabique, toute la séquence plosive-liquide se réalisera comme une suite de deux nasales (ex. /tɕap.lju/ *personnes vulgaires* > [tɕam.nju]). Aucun élément voisé n'est donc possible pour le segment qui suit les consonnes cibles en coréen. Nous avons par conséquent opté pour une fricative disponible dans le système (/s/ ou /h/), certes, qui est sourde, mais qui permet au moins de segmenter, même en cas de non relâchement des plosives, la partie du silence (occlusion) de la partie de la turbulence (friction). La fricative /h/ n'est pas adéquate car /p, t, k/ lenis deviendront attaques (resyllabification) et se réaliseront respectivement /p^h, t^h, k^h/ aspirés. Avec ces différentes contraintes phonologiques et phonotactiques, la fricative /s/ est le dernier candidat possible pour le segment suivant les plosives cibles en coréen. Bien que plusieurs critères soient contrôlés, nous sommes conscients que ce choix très restreint de /s/ pourrait faire émerger certains phénomènes de coarticulation avec les plosives cibles. Faute de trouver suffisamment de mots répondant à nos critères dans le lexique du coréen, nous avons élargi le corpus en ajoutant une deuxième voyelle (/i/) pour cette langue.

Au total, 36 mots vietnamiens (12 monosyllabes (4*/p/, 4*/t/, 4*/k/) et 24 composés dont la seconde syllabe est sous 2 contextes tonals), 34 mots du coréen (17 monosyllabes (5*/p/, 5*/t/, 7*/k/) et 17 composés) ont été sélectionnés pour les deux corpus. Ils ont ensuite été insérés dans de courtes phrases porteuses : en vietnamien [dɤ̌j la tɯ _ **d**ɤ̌j] (*Voici le mot* …) et en coréen [jʌ.ki.sʌ _ **s**ak.ʧe.hɛ] (*Supprime le mot ... ici*), chacune répétée trois fois, et mises pour chaque langue en ordre aléatoire. Les locuteurs étaient installés face à un écran de 24 pouces sur lequel étaient présentés les énoncés un par un. Plusieurs pauses étaient insérées au cours des séances d'enregistrement. Le corpus du vietnamien a été enregistré à MICA, Institut Polytechnique de Hanoi (Vietnam) et celui du coréen à SNU, Université nationale de Séoul (Corée du Sud), dans leurs chambres sourdes pendant l'été 2019.

3.3 Locuteurs

Vingt locuteurs par langue (10 hommes, 10 femmes) ont été recrutés pour l'expérience. Tous sont natifs de la même variété dialectale (parler du Nord pour le vietnamien, parler standard de Séoul pour le coréen). La moitié des locuteurs recrutés n'est pas ou n'a pas été au contact d'autres langues étrangères contenant des plosives finales relâchées allophoniques (groupe sans L2). L'autre moitié a déjà des expériences linguistiques autres que leur langue maternelle (groupe avec L2). Il a été montré qu'au contact de personnes plurilingues, la parole varie d'une langue à l'autre et implique des ajustements biomécaniques et aérodynamiques au niveau de l'appareil phonatoire (parmi d'autres Wagner & Braun, 2003). Les résultats préliminaires présentés ci-après portent sur les monosyllabes $CVC_{2\#}$, produits par 4 sujets par langue (homme *vs* femme, avec L2 *vs* sans L2). D'autres stimuli et sujets sont en cours d'analyse.

3.4 Mesures et analyses

La segmentation et l'étiquetage des phonèmes cibles ont été manuellement effectués à partir du signal acoustique synchronisé avec les données aérodynamiques. Les mesures des débits et de la trajectoire du larynx ont été extraites avec un logiciel interne du GIPSA-lab (TRAP) développé dans l'environnement Matlab. Les mesures relevées sont les suivantes : (1) valeurs du débit d'air nasal pendant la durée de la plosive cible, (2) valeurs des minima et maxima de la courbe qui représente le déplacement vertical du larynx, d'une part entre le début et la fin acoustique des plosives finales et, d'autre part, entre le début et la fin acoustique des plosives sourdes en initiale de mot. Les mesures du déplacement vertical du larynx (ΔPVL) ont été calibrées en divisant l'amplitude du mouvement vertical (Min-Max) mesurée entre le début et la fin acoustique de la plosive par la valeur maximale absolue de la PVL pour l'ensemble des sujets dans chaque langue. Le pourcentage d'abaissement du larynx par rapport à la position maximale absolue du larynx pour chaque langue est donc calculé avec cette formule : 100*(Min PVL – Max PVL) / Max abs PVL.

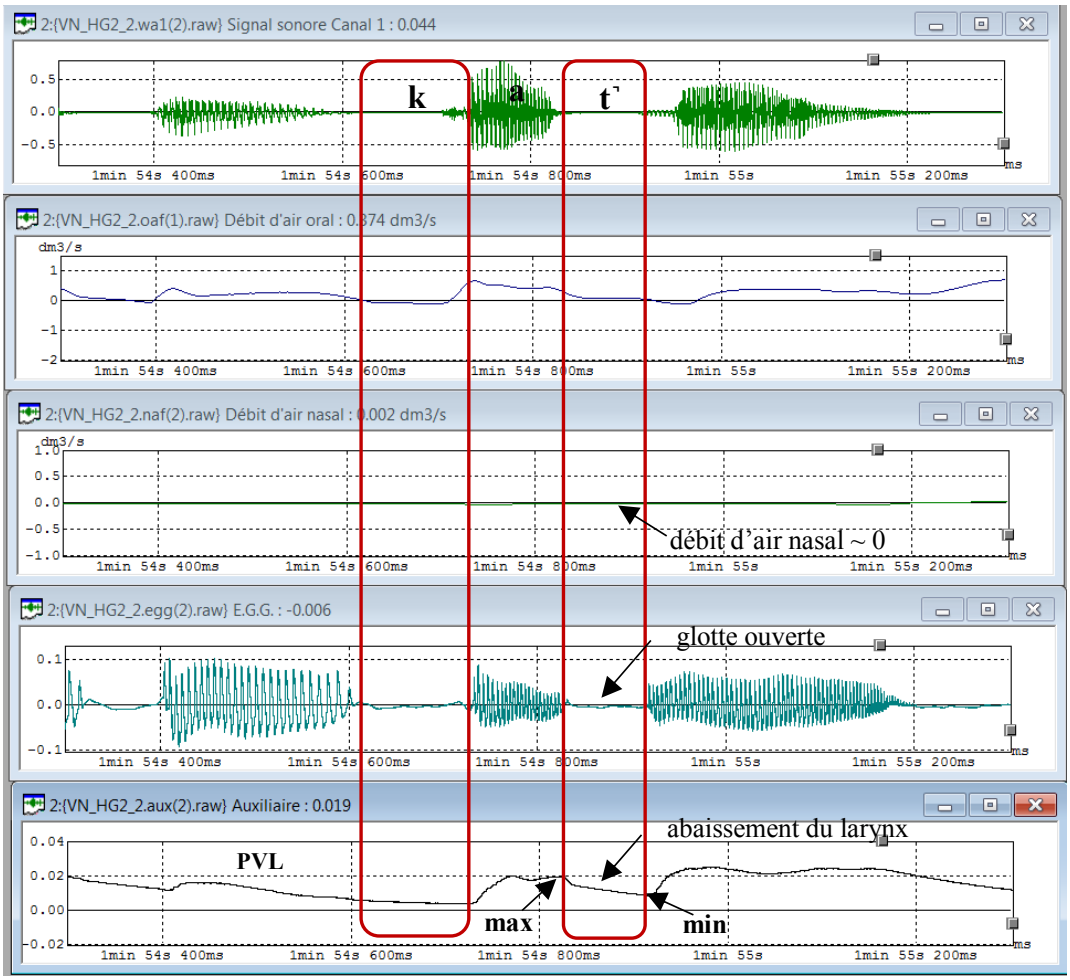


FIGURE 1 : Paramètres mesurés pour le mot /kat/ ('sable') réalisé [kat˺] par un locuteur vietnamien (de haut en bas : signal acoustique, débit d'air oral, débit d'air nasal, signal EGG, déplacement vertical du larynx)

4 Résultats

Rappelons que les plosives finales /p, t, k/ en vietnamien et en coréen sont généralement décrites comme non relâchées en raison de l'absence de burst après la phase d'occlusion (entre autres Doan, 1999 ; Kim, 1998). L'analyse acoustique confirme la caractéristique non relâchée de ces plosives

finales en vietnamien déjà remarquée dans les travaux antérieurs (Tran, 2011 ; Tran et al, 2019). Une expertise auditive et visuelle à partir de l'analyse spectrographique du signal de parole n'a montré pour le vietnamien aucune présence de bruit d'explosion lors de la production des plosives finales et ce quel que soit le locuteur. Les productions des consonnes finales sont donc à 100 % non relâchées. Pour le coréen, les réalisations des plosives finales sont moins homogènes. Rappelons que ces consonnes sont suivies de /s/, initiale de la syllabe suivante dans la phrase porteuse. Dans ce contexte, les réalisations de /k/ ont été trouvées avec burst. L'alvéolaire /t/ a été assimilée au mode articulatoire de la fricative de même lieu d'articulation. Aucune interruption du signal n'a été constatée pendant la réalisation de cette plosive. Pour tous les locuteurs, /t/ > [s] est relevé. Seule la bilabiale /p/ a été observée sans burst, et ce chez les 3 locuteurs sur 4 analysés à présent. Le sujet KO_HG, venant de la province de Gumi (dans l'Est du pays), est le seul locuteur qui n'est pas originaire de Séoul parmi les sujets sans L2. Ce locuteur prononce la plupart du temps les plosives finales avec un relâchement bruité. Les résultats présentés ci-après ne prennent donc en compte que les plosives finales effectivement réalisées non relâchées. Dans notre analyse, ont été considérés les monosyllabes suivants : 12 en vietnamien /fap/, /ʔap/, /dap/, /zap/, /kat/, /lat/, /fat/, /nat/, /ɣak/, /sak/, /ʔak/, /lak/ * 4 locuteurs (VN_FE, VN_HE, VN_FG, VN_HG) et 5 en coréen /kap/, /tap/, /pap/, /ip/, /tʃip/ * 3 locuteurs (KO_FE, KO_HE, KO_FG).

4.1 Débits d'air nasal et oral

Pour aucun locuteur dans chacune des deux langues, la présence d'un débit d'air nasal n'a été observée, quel que soit le lieu d'articulation de la consonne non relâchée et la langue du locuteur (cf. Figure 1). Ce résultat infirme l'observation de Ladefoged et Maddieson (1996) en indiquant clairement que le relâchement nasal n'a pas été utilisé comme mécanisme compensatoire pour faire baisser la pression intra-buccale lors de la réalisation des plosives non relâchées en finale de syllabe. Il est important de préciser que toutes les consonnes cibles ont été réalisées comme des plosives avec obstruction totale de la cavité buccale, donc sans fuite labiale. Ce résultat est validé par le débit d'air oral nul pendant toute la tenue de l'occlusion. La chute du débit d'air oral est observée dès le début de la consonne (cf. Figure 1).

4.2 Déplacement vertical du larynx

Un mouvement systématique d'abaissement du larynx a été relevé pendant toute la durée de la phase d'occlusion des plosives finales non relâchées dans les deux langues, quel que soit le locuteur (homme (H) ou femme (F), étudiant avec L2 (E) ou sans L2 (G)) (cf. Figure 2).

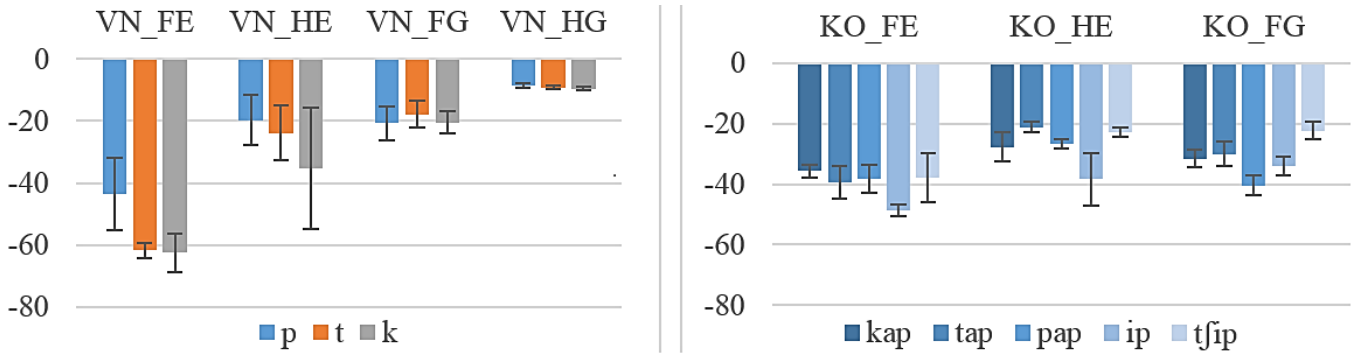


FIGURE 2 : Variation de la descente du larynx (ΔPVL) par rapport à la position maximale absolue (exprimée en %), pour les 4 locuteurs vietnamiens et /p, t, k/ à gauche, et les 3 locuteurs coréens pour /p/ à droite.

Pour être en mesure d'observer les effets en inter-sujet, l'amplitude du déplacement vertical du larynx est estimée par rapport au point de référence correspondant spécifiquement à la valeur de la position la plus haute (Max abs PVL) relevée pour l'ensemble des locuteurs d'une langue donnée (maximum absolu). Les mêmes tendances sont observées en intra-sujet si le point de référence correspond à la position la plus haute chez chaque locuteur dans les deux langues. Pour le vietnamien, un abaissement de la PVL plus important est observé chez la locutrice avec L2 (VN_FE). La valeur maximale absolue relevée chez les 4 sujets vietnamiens appartient à cette locutrice (0.11). En moyenne, un abaissement du larynx de 43.64 % par rapport à cette valeur de référence est observé chez cette locutrice lors de la production de /p/ et encore davantage pendant celles de /t/ et /k/ (61.75 % et 62.53 % respectivement). Concernant le locuteur VN_HE, il est observé que le larynx s'abaisse moins pour /p/ (19.65 %) et /t/ (23.84 %) que pour /k/ (35.20 %). On relève une grande variation dans les valeurs de ΔPVL chez ce locuteur, notamment pour /k/. Pour les deux autres locuteurs (groupe sans L2), le mouvement d'abaissement du larynx est nettement moins important par rapport à la locutrice VN_FE, et un peu moins par rapport à VN_HE. On remarque une valeur d'abaissement assez stable entre les 3 plosives finales /p, t, k/, respectivement 20.74 %, 17.86 %, 20.55 % chez la locutrice VN_FG et 8.37 %, 9.08 %, 9.53 % chez le locuteur VN_HG. On trouve chez ce dernier l'abaissement vertical du larynx le moins important de tous les locuteurs vietnamiens. Quant au coréen, pour les mots à finale /p/, la valeur maximale de référence est relevée chez la locutrice KO_FE (0.067). On observe un abaissement du larynx un peu plus important lors de la réalisation de /p/ chez ce sujet (39.90 %) que chez le locuteur KO_HE (27.34 %) et chez la locutrice sans L2 KO_FG (31.65 %). En moyenne chez la locutrice KO_FE, le larynx descend plus pendant la réalisation de la plosive finale du mot /ip/ (48.57 %). La même observation est faite chez le locuteur KO_HE (38.37 %) pour ce mot. Quant au sujet KO_FG, le larynx descend davantage pour la plosive finale dans /pap/ (40.37 %) que dans /ip/ (34.04 %).

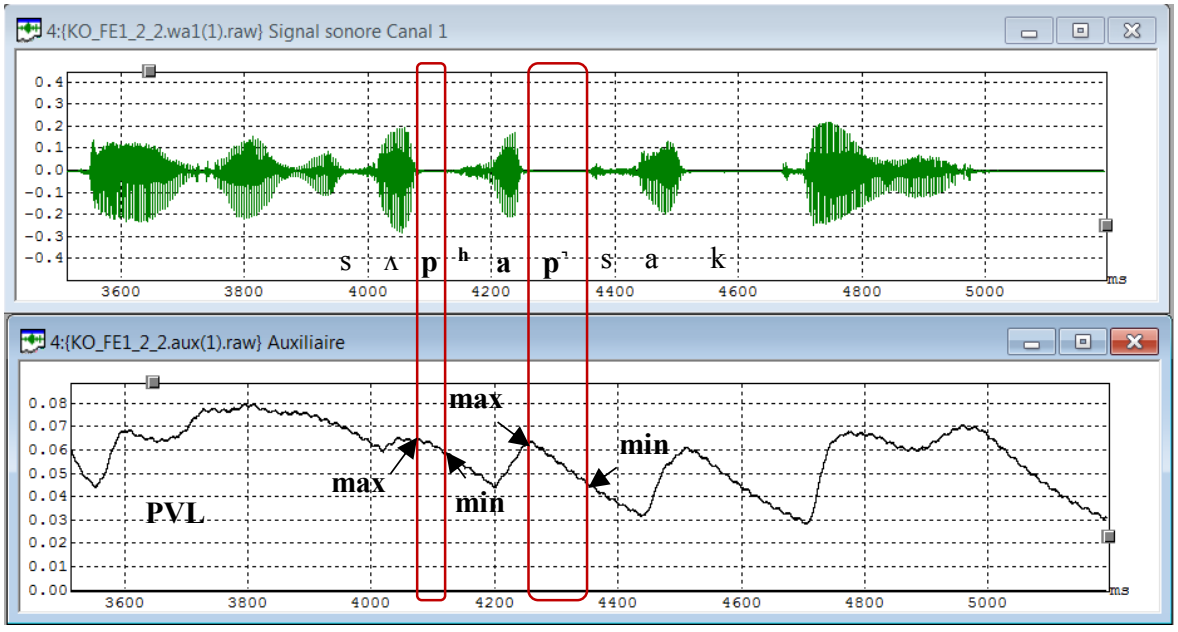


FIGURE 3 : Descente du larynx pendant la réalisation de /p/ (initiale *vs* finale) du mot /pap/ [p^{h}ap˺] ici chez la locutrice KO_FE.

La PVL a aussi été examinée en fonction de la position des plosives dans la syllabe en regard de la réalisation acoustique (avec *vs* sans burst) en fonction de leur position (initiale *vs* finale). Les mots /kat/ (vietnamien), /kap/, /tap/, /pap/ (coréen) ont donc été pris en compte dans cette analyse. Un effet de la position de la plosive dans la syllabe sur la PVL a été observé chez la plupart des sujets dans les deux langues. La figure 3 montre l'abaissement du larynx lors de la production de /p/ à la fois en

initiale (max = - 0.064, min = - 0.058) et en finale (max = - 0.056, min = - 0.043) du mot /pap/ par la locutrice coréenne KO_FE. On observe que le larynx descend davantage comparé au point de référence (0.067) lorsque /p/ se trouve en coda (26.70 %) plutôt qu'en initiale de la syllabe (8.90 %). Il est à préciser que /p/ lenis est réalisé légèrement aspiré en initiale (Hardcastle, 1973). On observe que le larynx descend lors de la production d'une consonne sourde, ce qui est conforme aux résultats de Shipp et al. (1987). Le maximum de la PVL au début de la plosive finale est bien contrasté en raison de la voyelle précédente ([a]). On relève d'ailleurs une remontée du larynx après /s/ pour la réalisation de la voyelle suivante /a/ (cf. Figure 3). Chez les vietnamiens, la différence de valeurs de ΔPVL en fonction de la position de la plosive dans le mot /kat/ est clairement observée chez les locuteurs VN_FE, VN_HE, VN_HG où le larynx descend davantage pour /t/ finale que pour /k/ initiale (voir Figure 1). Respectivement, pour ces trois sujets, l'abaissement du larynx en fonction de la position de la plosive (finale *vs* initiale) est respectivement de 64.01 % *vs* 24.43 %, 29.10 % *vs* 10.27 %, 8.67 % *vs* 4.26 %. Chez le sujet VN_FG, la différence est peu marquée (17.54 % *vs* 17.01 %).

5 Discussion et conclusion

Ce travail porte sur une étude préliminaire du comportement glottique en synchronisation avec le timing de l'écoulement d'air nasal, lors de la réalisation de /p, t, k/ en vietnamien et en coréen. Les résultats à l'issue de ce travail fournissent des éléments permettant de considérer nos deux hypothèses. Concernant l'hypothèse H1 portant sur la présence d'un débit d'air nasal, nos données montrent que la syllabe à plosive finale n'est jamais suivie d'un relâchement nasal. Ce résultat va donc à l'encontre de ce qui est relevé par Ladefoged et Maddieson (1996) selon lesquels « *the release is by lowering the velum while the oral closure is maintained, so that a short voiceless nasal is produced* ». Aucune fuite nasale n'est observée dans notre étude lors de la production des plosives non relâchées ; le port vélo-pharyngé est donc fermé durant et après l'obstruction buccale. À propos de l'hypothèse H2 portant sur la PVL, il est observé que, quelle que soit la langue étudiée (tonale ou non tonale), quel que soit le sujet (homme ou femme, avec ou sans L2), le larynx descend systématiquement lors de la production des plosives finales sourdes non relâchées. L'abaissement du larynx est d'ailleurs observé pour toutes les consonnes sourdes, comme relevé par Shipp et al (1987). Mais une différence nette de ΔPVL est aussi remarquée entre plosive sourde relâchée en initiale et plosive sourde non relâchée en finale pour la plupart des sujets des deux langues : l'abaissement du larynx est plus important pour les plosives finales non relâchées. Ce processus pourrait donc bien être une « manœuvre de compensation » planifiée du non relâchement observé dans ces langues car provoquant, en l'absence de fuite nasale ou labiale, une baisse de la pression intra-buccale. Il reste à vérifier ce résultat statistiquement avec l'ensemble des locuteurs déjà enregistrés et avec des stimuli de type CVC où C_1 et C_2 sont identiques ([pap], [tat], [kak]). Chez l'ensemble de nos locuteurs, l'abaissement du larynx n'est accompagné d'aucun signal EGG, ce qui signifie que la glotte reste ouverte pendant la descente du larynx. Sans que l'on puisse déterminer la position des plis vocaux qui renseignerait sur le degré d'ouverture à la glotte, l'absence de signal EGG indique que ces plosives ne sont pas glottalisées, ce qui confirme les résultats de travaux antérieurs (Iwata et al., 1990 ; Michaud, 2004). Quelle que soit l'expérience linguistique (au contact d'une L2 ou non), les locuteurs vietnamiens produisent les plosives finales sans burst visible sur le signal, en conformité avec notre étude antérieure (Tran, 2011). Il a été montré en effet que lors de l'apprentissage du FLE, même à un niveau avancé bénéficiant de longs contacts avec la langue cible, les vietnamiens conservent toujours le schéma articulatoire de leur langue maternelle qui consiste à ne pas prononcer les plosives finales avec un relâchement audible lorsqu'ils parlent en français. Pour avancer sur les questions soulevées par la présente étude, dans le prolongement de ce travail, une analyse d'autres locuteurs, d'autres mots (composés) et d'une autre langue tonale (le thaï) déjà enregistrés est en cours. Il est aussi prévu de compléter cette étude avec des données acoustiques quantitatives et perceptives déjà récoltées auprès des mêmes locuteurs.

Remerciements

Cette étude est financée par le projet SCALA (ANR-15-IDEX-02). Nous remercions vivement Alain Ghio, Yohann Meynardier, Nathalie Henrich, Sylvain Geranton, Alain Arnal pour leur aide et assistance précieuse. Grand merci à Yeo Eun Jung, Do Hee Kim, Seo Yoon Lee (SNU) et à Viet-Son Nguyen, Ngoc-Diep Do, Thanh-Hai Tran, Thi-Lan Le (MICA) ainsi que tous les locuteurs qui ont participé aux enregistrements.

Références

DOAN T. T. (1999). *Ngữ âm tiếng Việt (Vietnamese Phonetics)*. Hanoi National University Publishing

EDMONDSON J. A., CHANG C., HUANG J., HSIEH F. & PENG Y. (2010). Reinforcing voiceless stop codas in Taiwanese, Vietnamese and other East and Southeast Asian languages: Laryngoscopic case studies. *Labphon*, 13, Albuquerque, New Mexico.

EDMONDSON J. A., CHANG Y., HSIEH F. & HUANG H. J. (2011). Reinforcing voiceless finals in Taiwanese and Hakka: Laryngoscopic case studies. In W. S. LEE & E. ZEE, Éds., *Proceedings of the 17th International Congress of Phonetic Sciences,* 627-630, Hong Kong.

GHIO A. & TESTON B. (2002). Caractéristiques de la dynamique d'un pneumotachographe pour l'étude de la production de la parole : aspects acoustique et aérodynamique. *24èmes Journées d'Etudes sur la Parole (JEP)*, 337-340, Nancy, France.

GUZMAN M., CASTRO C., TESTART A., MUÑOZ D. & GERHARD J. (2013). Laryngeal and pharyngeal activity during semioccluded vocal tract postures in subjects diagnosed with hyperfunctional dysphonia. *Journal of Voice,* 27(6), 709–716. DOI: 10.1016/j.jvoice.2013.05.007

HARDCASTLE W. J. (1973). Some observations on the Tense-Lax distinction in initial stops in Korean. *Journal of Phonetics,* 1, 263-271. DOI: 10.1016/S0095-4470(19)31390-7

IWATA R., HIROSE H., NIIMI S., SAWASHIMA M. & HORIGUCHI S. (1990). Syllable final stops in East Asian languages: Southern Chinese, Thai, and Korean. *Proceedings of the 1990 International Conference on Spoken Language Processing*, 621-624, Kobe, Japan.

KEATING P. A., WRIGHT R. & ZHANG J. (1999). Word-level asymmetries in consonant articulation. *University of California Working Papers in Phonetics*, 97, 157-173.

KIM H. (1998). A Phonetic Characterization of Release and Nonrelease: The Case of Korean and English. *Language Research*, 34(2), 347-368.

KINGSTON J. (2008). Lenition. In L. COLANTONI, J. STEELE, Éds., *Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology*, 1-31, Somerville, MA: Cascadilla Press.

LADEFOGED P. & MADDIESON I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

MICHAUD A. (2004). Final consonants and glottalization: New perspectives from Hanoi Vietnamese. *Phonetica*, 61, 119-146. DOI: 10.1159/000082560

OHALA J. J. (1975). A mathematical model of speech aerodynamics. In: G. FANT, Éds., *Speech Communication*. *Vol. 2: Speech production and synthesis by rule*, 65-72, Stockholm: Almqvist & Wiksell.

SHIPP T., GUINN L., SUNDBERG J. & TITZE I. R. (1987). Vertical laryngeal position. Research findings and their relationship to singing. *Journal of Voice*, 1(3), 220-222.

SUNG E.-K. (2005). Perception of Flaps in American English and Korean, *Proceedings of the 4th International Symposium on Bilingualism*, 2197-2221. Somerville, MA: Cascadilla Press.

TRAN T. T. H. (2011). *Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde : de l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE*. Thèse de Doctorat, Université de Grenoble.

TRAN T. T. H., VALLÉE N. & GRANJON L. (2019). Effects of word position on the acoustic realization of Vietnamese final consonants. *Phonetica*, 76 (1), 1-30, Karger. DOI: 10.1159/000485103

WAGNER A. & BRAUN A. (2003). Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. *Proceedings of the 15th International Congress of Phonetic Sciences*, 6, 651-654, Barcelona, Spain.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Capacités d'apprentissage phonétique chez des patients aphasiques francophones : étude de cas.

Clémence Verhaegen[1], Véronique Delvaux[1,2], Kathy Huet[1], Sophie Fagniart[1], Myriam Piccaluga[1], Bernard Harmegnies[1]
(1) Unité de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fond National de la Recherche Scientifique, Belgique
clemence.verhaegen@umons.ac.be, veronique.delvaux@umons.ac.be,
kathy.huet@umons.ac.be, sophie.fagniart@umons.ac.be,
myriam.piccaluga@umons.ac.be, bernard.harmegnies@umons.ac.be

RÉSUMÉ

Cette étude explore les capacités de patients aphasiques, présentant des troubles phonologico-phonétiques, notamment des difficultés de coordination temporelle entre les articulateurs, à acquérir une variante phonétique, non familière dans leur langue, nécessitant l'adoption de nouveaux schèmes articulatoires. 4 patients aphasiques, de langue maternelle française, ont participé à la présente étude, ainsi que 36 participants contrôles. Au cours du paradigme d'apprentissage, la tâche principale consistait à répéter des non-mots C[t]V[a], dont le VOT est de 60ms et ce à 3 reprises : avant toute intervention, puis après un « entraînement » en perception-tâche de discrimination de 5 non-mots CV dont le VOT variait entre 20 et 100ms, enfin après un « entraînement » en production-tâche de répétition de ces 5 non-mots. Les participants étaient par ailleurs amenés à effectuer une tâche de calibration, destinée à évaluer leur VOT en français. Les patients présentent une plus grande variabilité des durées de VOT. Trois patients sur quatre montrent des valeurs de VOT plus longues entre la calibration et les tâches de répétition de VOT60ms indiquant des capacités d'apprentissage phonétique. Les liens entre ces observations et les profils des patients, ainsi que les implications pour la rééducation du langage, seront discutés.

ABSTRACT

Phonetic learning abilities in French-Speaking aphasic patients : a case study

This study investigates the capacities of aphasic patients with phonological-phonetic disorders, and coordination difficulties between articulators, to acquire a non-usual phonetic variant in their mother tongue that requires new articulatory schemes. Four French-speaking aphasic patients participated to the present study, as well as 36 control participants. The experimental paradigm consisted in repetition sessions of 5 C[t]V[a] non-words of 60 ms VOT (target), 3 times: before any training, after a perceptual training (AX discrimination involving CV pseudo-words of 20-ms to 100-ms VOT), and after a training in production (repetition task of the non-words). Participants were also presented a calibration task, what aim was to analyze their VOT values in French. The patients show a larger variability of the VOT values. 3 participants also show longer VOT values in the repetition task of non-words with VOT of 60 ms than in the calibration task, indicating phonetic learning capacities. The relations between these results and the profiles of our patients, as well as the implication for their language rehabilitation, are discussed.

MOTS-CLÉS : VOT, apprentissage phonétique, aphasie, troubles phonologico-phonétiques

KEYWORDS: Voice Onset Time, phonetic learning, aphasia, phonological-phonetic disorders

Volume 1 : Journées d'Études sur la Parole, pages 606–616.
HAL : hal-02798593.

1 Introduction

Les troubles phonologico-phonétiques sont fréquents dans l'aphasie. Bien qu'encore débattus actuellement, tant en ce qui concerne leur description que leur évaluation, les auteurs s'accordent fréquemment sur le fait que ces troubles consistent en une atteinte de la sélection des phonèmes au sein du système phonologique (troubles plus « phonologiques ») ou en une atteinte de la programmation ou de l'exécution motrice des mouvements nécessaires à la réalisation des phonèmes (troubles plus « phonétiques »). Au niveau de la production orale, ces atteintes résultent en des paraphasies ou erreurs consistant en des ajouts, omissions, permutations ou substitutions de phonèmes au sein du mot (en cas de trouble à tendance plus « phonologique ») ou en des distorsions de la réalisation des phonèmes (en cas de trouble à tendance plus « phonétique ») (Galluzzi, Bureca, Guariglia, & Romani, 2015; Laganaro, 2015; Nespoulous, Baqué, Rosas, Marczyk, & Estrada, 2013).

Ces deux dernières décennies principalement, en vue d'apporter une description plus précise et plus objective des troubles phonologico-phonétiques dans l'aphasie, certains auteurs ont analysé les troubles de la production de la parole dans l'aphasie à l'aide d'analyses acoustiques des signaux de parole (Auzou et al., 2000; Baqué, Marczyk, Rosas, & Estrada, 2015; Blumstein, Cooper, Goodglass, Statlender, & Gottlieb, 1980; Galluzzi et al., 2015; Marczyk, Baqué, Rosas, & Nespoulous, 2011; Nespoulous et al., 2013; Ryalls, Provost, & Arsenault, 1995; Verhaegen et al., 2019). La plupart de ces travaux se sont centrés sur les contoïdes plosives et ont, pour la majorité, recouru à l'analyse du Voice Onset Time (VOT), qui mesure le délai entre le relâchement de l'occlusion supra-glottique et l'apparition des vibrations laryngées (Lisker & Abramson, 1964). Les études ayant procédé à des analyses acoustiques de la production des plosives chez les patients aphasiques ont montré qu'en cas d'atteinte phonologico-phonétique, les patients présentaient une plus grande variabilité au niveau des valeurs de VOT, engendrant des recouvrements inter-catégoriels entre les voisées et non-voisées, de même que des difficultés d'initiation et de maintien du voisement des plosives voisées dans des langues telles que le français, qui a la particularité de présenter un prévoisement avec un VOT (négatif) long. Les auteurs ont principalement attribué ces irrégularités du VOT à des difficultés de coordination temporelle entre les articulateurs glottiques et supra-glottiques (Auzou et al., 2000; Baqué et al., 2015; Laganaro, 2015; Nespoulous et al., 2013; Verhaegen et al., 2019).

Dans cette étude exploratoire, nous nous intéressons aux capacités de patients aphasiques francophones présentant des troubles phonologico-phonétiques, et par conséquent des difficultés de coordination temporelle entre les articulateurs, à acquérir une variante phonétique non familière dans leur langue (ici: une occlusive initiale non voisée avec un VOT positif long), nécessitant l'adoption de nouveaux schèmes de timing entre les articulateurs. L'objectif de cette étude est double : (1) si les patients sont incités à produire en adoptant de nouveaux schèmes articulatoires, il est possible que les résultats obtenus montrent un accroissement des phénomènes déjà observés en langue française (par exemple, augmentation de la variabilité des durées des VOT chez les patients aphasiques) ou encore l'apparition de phénomènes non favorisés par les structures de la langue maternelle (par exemple, raccourcissement de la durée des VOT positifs longs); (2) si les patients aphasiques présentent toujours des capacités d'acquisition de nouveaux schèmes articulatoires, cela constitue un indice intéressant pour la rééducation orthophonique des troubles phonologico-phonétiques chez les patients aphasiques, fréquemment basée sur le réapprentissage des schémas articulatoires des phonèmes atteints dans leur langue.

2 Matériel et méthode

2.1 Participants

Quatre patients aphasiques, IJ, CL, TM et BD, de langue maternelle française, ont participé à la présente étude. Trois patients (IJ, CL, TM) ont été diagnostiqués par des orthophonistes comme présentant une aphasie de Broca, et le patient BD a été diagnostiqué comme présentant une aphasie de Wernicke. Tous les patients se caractérisaient par une vue non altérée ou corrigée et une absence d'atteinte auditive. La Table 1 présente leurs principales caractéristiques. Les performances des patients ont été comparées à 36 participants contrôles, appariés en âge (44-54 ans, 55-65 ans et 66-75 ans, *N*=12 dans chaque groupe de participants). Les participants ne présentaient pas d'atteinte psychologique, neurologique ou langagière. Leurs capacités visuelles et auditives étaient dans les normes ou corrigées.

Patient	Âge	Genre	Type d'aphasie	Temps post-onset	Etiologie	Lésion	Groupe contrôle
IJ	44	F	Broca	18 mois	AVC	Fronto-pariétale	44-54 ans
CL	65	M	Broca	2 ans	AVC	Fronto-pariétale	55-65 ans
TM	62	M	Broca	11 ans	AVC	Fronto-temporale	55-65 ans
BD	72	M	Wernicke	18 mois	AVC	Pariétale	66-75 ans

TABLE 1 : Résumé des informations relatives aux participants de notre étude.

Le bilan orthophonique des patients aphasiques a montré qu'ils présentaient tous une atteinte de la dénomination (*Lexis*, Bilocq et al., 2001) et de la répétition de mots et de non-mots (*Examen Long du Langage*, UCL-ULg) et commettaient de nombreux ajouts, omissions, permutations ou substitutions ainsi que des distorsions de phonèmes. Par contre, ils ne présentaient pas d'altération importante de la compréhension du langage, évaluée à l'aide de tâches de désignations de mots (*Examen Long du Langage*, UCL-ULg) ou de phrases (*Montréal-Toulouse*, Joanette et al., 1998). En outre, tous les patients montraient une atteinte de la mémoire à court terme ainsi que des fonctions exécutives de mise à jour, flexibilité et inhibition[1]. L'ensemble des patients a été également évalué précédemment à l'aide d'une tâche de répétition de pseudo-mots CVCV, contenant les 6 occlusives voisées et non voisées du français, accompagnées de la voyelle /a/ (ex. /pata/). Cette tâche nous a permis d'examiner leurs VOT en langue française. Ces résultats ayant été publiés précédemment (Verhaegen et al., 2019), pour des raisons de place, ils ne seront pas décrits en détail dans cet article. Les résultats ont montré la présence d'un grand nombre de dévoisements de consonnes voisées ainsi que des valeurs de VOT négatifs (prévoisement) des consonnes voisées plus courtes chez les patients CL et TM, traduisant des difficultés de tenue du

[1] Le lecteur intéressé par les détails de l'évaluation langagière et neuropsychologique des patients est invité à consulter l'étude de Verhaegen et al. (2019) dans laquelle les analyses des patients présentés dans cette étude sont décrites.

voisement. Ces difficultés ont été interprétées comme révélatrices d'un déficit de coordination des articulateurs en raison de troubles plus « phonétiques ». Chez les patients BD et IJ par contre, on notait un grand nombre de changements de lieux et de modes d'articulation, ainsi que la présence d'un grand nombre de voisements et de dévoisements complets, traduisant des difficultés de sélection du phonème adéquat et dès lors des troubles plus « phonologiques ». Cependant, l'ensemble des patients montrait des valeurs de VOT des voisées et non voisées plus variables, traduisant probablement des difficultés de précision dans le timing entre les articulateurs chez tous les patients. La Table 2 résume les valeurs des VOT des consonnes voisées et non voisées du français chez les patients aphasiques, en comparaison avec les sujets contrôles.

	Patients		Contrôles	
	Voisées	Non voisées	Voisées	Non voisées
IJ	+ 5 (26) *	+16 (18)*	- 122 (50)	+ 30 (16)
CL	- 46 (80) *	+ 14 (32)	- 44 (56)	+ 24 (21)
TM	- 56(52) *	+ 4 (37)*	-144 (56)	+ 24 (21)
BD	- 61 (80)	+ 8 (64)	- 75 (74)	+ 23 (29)

TABLE 2 : Durées de VOT (ms) des patients aphasiques en langue française dans la tâche de répétition de non-mots CVCV.
*= Performance significativement différente des participants contrôles (Crawford et al., 2010).

2.2 Paradigme expérimental

Dans la présente étude, nous proposons un paradigme adapté de précédents travaux visant initialement à entraîner des adultes francophones afin qu'ils acquièrent une nouvelle variante phonétique, non familière en langue maternelle, à savoir une occlusive initiale non voisée avec un VOT long, typique de l'anglais (Delvaux, Cano-Chervel, Huet, Piccaluga, & Harmegnies, 2011; Delvaux, Huet, Piccaluga, & Harmegnies, 2014). Concrètement, ce paradigme consistait en cinq répétitions de syllabes C[t]V[a], dont le VOT était de 60 ms, proposées trois fois. Entre chaque répétition, les participants étaient amenés effectuer des tâches de production (répétition «la plus fidèle possible») et de perception (discrimination AX), de pseudo-mots CV se différenciant uniquement par le VOT de la consonne initiale (respectivement de 20, 40, 60, 80 et 100 ms ; toutes les autres propriétés étant strictement invariantes). Ces stimuli ont été construits à partir d'un pseudo-mot [t^ha] produit en parole naturelle par une locutrice anglophone, qui se composait d'un burst de 20 ms, suivi d'une "aspiration" de 20 ms (soit d'un VOT de 40 ms), et enfin d'une voyelle de 210 ms. De cette production naturelle, le burst et la voyelle ont été conservés dans tous les stimuli. Seule a été manipulée la durée de l'aspiration, qui varie selon les stimuli entre 0 et 80ms, par pas de 20 ms. Le paradigme expérimental se déroulait dans l'ordre suivant : (1) Production de la 'cible' de l'apprentissage: répétition 'la plus fidèle possible' du stimulus VOT 60 ms (5 répétitions); (2) Perception: discrimination AX de 30 paires de stimuli différents (Intervalle inter-stimuli (ISI): 1000ms), la différence de VOT entre les stimuli est de 40 ms (ex. 20-60 ms) ; (3) Production: répétition du stimulus VOT 60 ms (5 répétitions) (idem qu'à l'étape 1); (4) (Re)production: répétition 'la plus fidèle possible' des stimuli présentés un par un, en ordre 'montant': VOT 20, 40, 60, 80, 100 ms (4 blocs) ; (5) Production: répétition du stimulus VOT 60

ms (5 répétitions) (idem qu'à l'étape 1). Une épreuve de calibration avait également lieu un autre jour. Elle consistait en une épreuve de dénomination (pour les participants contrôles) ou de répétition (pour les patients aphasiques, en raison de leurs importants troubles de dénomination), des mots 'pas', 'tas', 'k' (5 répétitions). Ce paradigme expérimental permet d'évaluer la capacité des participants à produire des VOT typiques de leur L1 (calibration), à produire des VOT typiques d'une L2 avant expérimentation (flexibilité phonétique) (1), à discriminer entre des stimuli dont les valeurs de VOT sont, dans leur majorité, non familières en L1 (2), à produire des VOT typiques d'une L2 après 'entraînement' perceptuel uniquement (3) à reproduire des VOT non familiers juste après écoute du modèle (4) enfin à produire des VOT typiques d'une L2 après un 'entraînement' en perception et en production (apprentissage phonétique) (5).

2.3 Mesures

Le VOT a été mesuré manuellement (en ms) sur l'oscillogramme entre le début du burst et le début du voisement (défini comme le passage par zéro précédant le premier cycle glottique). Les performances en perception ont été évaluées via le calcul du pourcentage de réponses correctes au test de discrimination.

2.4 Procédure générale

Les participants ont été évalués individuellement à leur domicile dans un local calme. Nous leur avons présenté l'ensemble des tâches (paradigme expérimental d'apprentissage phonétique et tâches destinées à évaluer le profil langagier des patients) sur 4 jours différents. Chaque séance durait entre 45 et 60 minutes. L'ordre des tâches était: Jour 1 : (1) Anamnèse, (2) Paradigme d'apprentissage phonétique (paradigme expérimental décrit ci-dessus); Jour 2 : (1) Calibration ; (2) Dénomination d'images (40 premiers items), (3) Désignation de phrases, (4) Tâche d'évaluation des VOT en langue française; Jour 3 : (1) Dénomination d'images (40 derniers items), (3) Désignation de mots ; Jour 4 : (1) Évaluation des fonctions exécutives, (2) Audiométrie.

3 Résultats et discussion

3.1 Discrimination (perception)

La Table 3 indique les résultats des participants dans la tâche de discrimination (tâche de perception, exprimée en pourcentages de réponses correctes). L'ensemble des patients aphasiques présente un résultat de 0% de réponses correctes et affichent par conséquent des résultats significativement inférieurs à ceux des participants contrôles (statistiques adaptées aux cas uniques, Crawford et al., 2010). Étant donné qu'il est peu probable que les capacités de discrimination auditive soient totalement réduites à zéro pour les patients aphasiques, nous attribuons ces résultats à des potentielles difficultés de compréhension des consignes de la tâche de la part des participants, de même qu'à des persévérations de la réponse « même », en raison de troubles des fonctions exécutives.

	Score patient	Contrôles
IJ	0 (0) *	52.14 (26.20)
CL	0 (0) *	62.78 (9.64)
TM	0 (0) *	62.78 (9.64)
BD	0 (0) *	54.04 (24.87)

TABLE 3 : Pourcentages moyens de réponses correctes dans la tâche de discrimination, et écarts-types, entre parenthèses.*= Performance significativement différente des participants contrôles (statistiques adaptées aux cas uniques Crawford et al., 2010).

3.2 Tâches de répétition de syllabes [t^{h}a] avec un VOT de 60 ms (production de la cible de l'apprentissage)

La figure 1 indique les performances des sujets dans les tâches de reproduction des syllabes cibles [t^{h}a] avec un VOT de 60 ms (production de la cible de l'apprentissage) en comparaison avec leurs valeurs de VOT en L1, dans la tâche de calibration. On note tout d'abord que les valeurs de VOT sont beaucoup plus variables pour l'ensemble des patients aphasiques, avec une variabilité très importante chez CL et BD. Ceci reproduit ce que nous observons déjà en L1, dans la tâche de répétition de pseudo-mots décrite ci-dessus. Au niveau des capacités d'acquisition d'un nouveau schème articulatoire, nous remarquons des différences importantes entre les patients. Les différences en termes de durées de VOT entre les valeurs dans la tâche de répétition de syllabes [t^{h}a] avec un VOT de 60 ms et celles dans la tâche de calibration ont été évaluées à l'aide de tests non paramétriques U de Mann-Whitney. Trois patients, TM, CL, et BD montrent des capacités de flexibilité et/ou d'apprentissage phonétique. En effet, leurs durées de VOT sont plus importantes, et plus proches de la valeur cible de 60 ms, dans les tâches de répétitions de VOT de 60 ms que dans la tâche de calibration. Parmi les trois patients, TM est celui qui semble le plus bénéficier du paradigme d'apprentissage. En effet, les différences en termes de valeurs de VOT entre la tâche de calibration et les tâches de répétition sont les plus importantes. La différence en termes de durées de VOT est statistiquement significative entre la tâche de calibration et la première tâche de répétition de VOT de 60 ms, U=135.00, p=.05, entre la tâche de calibration et les deuxièmes et troisièmes tâches de répétitions de VOT de 60 ms, U=125.00, p=.002 et U=123.00, p=.001, respectivement, et entre la première et la deuxième tâche de répétition de VOT de 60 ms, U=15.00, p=.008. Ces résultats indiquent des capacités d'ajustement des schèmes articulatoires dès la première tâche de répétition, ainsi que d'apprentissage phonétique (deuxième et troisième tâches de répétition) chez ce patient. D'ailleurs, TM ne présente pas de valeurs de VOT statistiquement différentes des contrôles (différences également évaluées à l'aide de tests U de Mann-Whitney). Chez CL, la différence est significative entre la tâche de calibration et la tâche de répétition de VOT 60 ms, première répétition, U=9.00, p=.004, mais plus entre les différentes tâches de répétitions de VOT de 60 ms ou entre les deuxième et troisième tâches de répétitions de VOT 60 ms et la calibration. En fait, pour CL, les valeurs de VOT pendant le paradigme expérimental sont surtout caractérisées par une variabilité importante, témoignant de tentatives plus ou moins réussies d'ajustement des schèmes articulatoires chez ce patient également. Chez CL, les valeurs de VOT atteintes en fin d'apprentissage sont significativement inférieures à celles des contrôles : pour la troisième répétition de VOT de 60 ms, U=9.00, p=.004. Les autres différences ne sont pas significatives, en raison, entre autres, des grandes valeurs d'écarts-types chez CL. BD présente un

profil plus particulier. En effet, le patient voise un grand nombre des consonnes dans la calibration, engendrant une moyenne de valeurs de VOT négative, alors que les cibles sont des phonèmes non voisés de sa L1. En outre, il commet également d'autres erreurs de changement de lieux d'articulation. Par contre, il ne présente pas ce phénomène dans les tâches de répétitions de VOT de 60 ms, dans lesquelles il ne commet aucune erreur de voisement ou de substitution de lieu d'articulation, indiquant une possible modification de ses schèmes articulateurs et un apprentissage. Il est également probable qu'une répétition identique du même non-mot, [t^ha] ait engendré moins de difficultés exécutives chez BD, qui présente des troubles exécutifs, en comparaison avec une répétition de non-mots différents (tâche de calibration). Le fait que les stimuli soient nouveaux dans la tâche d'apprentissage d'une variante en L2 peut également avoir focalisé son attention. Ainsi, pour BD, la différence entre la calibration et la première répétition de VOT 60 ms est significative, U=167.00, p=.03, de même qu'entre la tâche de calibration et la troisième tâche de répétition de VOT de 60 ms U=170.00, p=.05. Les autres différences ne sont pas significatives. Cependant, le patient ne semble pas ensuite bénéficier de l'apprentissage phonétique car ces valeurs de VOT n'augmentent significativement pas entre la première tâche de répétition et des deux autres, même si sa variabilité diminue entre la deuxième et la troisième répétition, ce qui pourrait être le signe d'un apprentissage phonétique. Les différences entre les valeurs de VOT de BD et celles des contrôles approchent de la significativité au niveau de la calibration, U=1355.00, p=.06, ce qui est expliqué par le fait que les valeurs de VOT des participants contrôles sont bien positives. Enfin, IJ semble le moins bénéficier du paradigme d'apprentissage phonétique. Comme on le note dans la Figure 1, les valeurs de VOT de la patiente augmentent peu entre la calibration et les différentes répétitions, ou entre les différentes épreuves de répétitions entre elles, et les différences ne sont pas significatives. Les différences avec les participants contrôles sont toutes significatives (calibration : U=806.00, p<.001; première répétition : U=43.00, p=.003; deuxième répétition : U=47.00, p=.006; troisième répétition : U=5.00, p<.001). Notons cependant que, tout comme BD, IJ commet des erreurs de changement de lieu d'articulation dans la tâche de calibration alors que ce n'est pas le cas dans la tâche d'apprentissage.

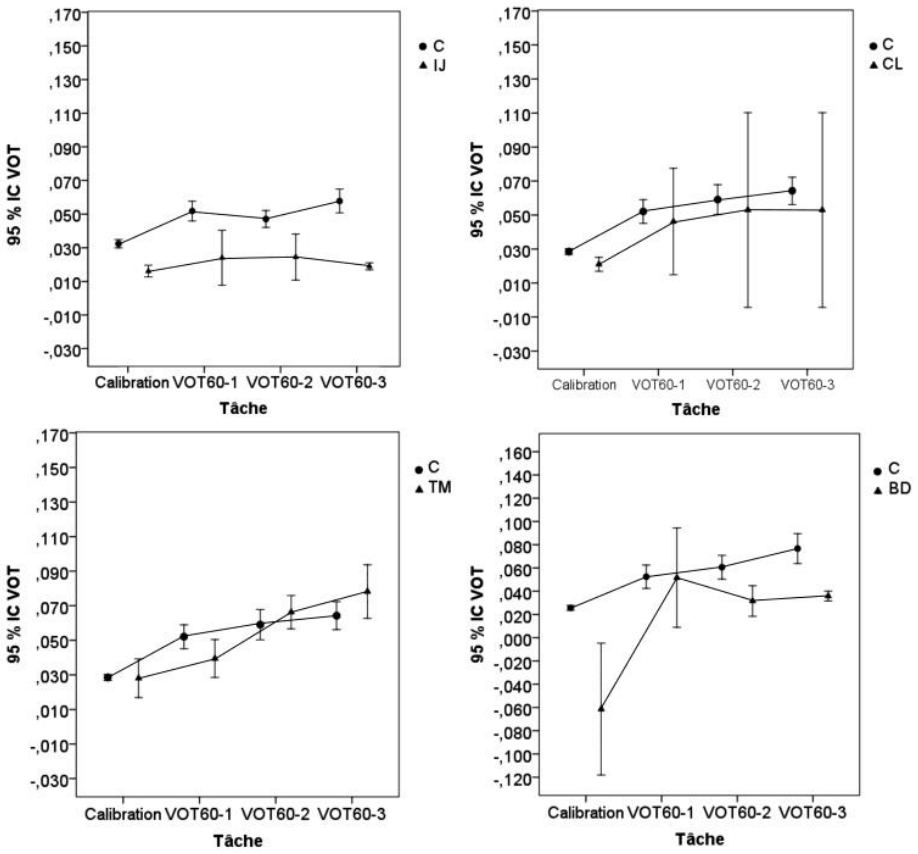


FIGURE 1: Valeurs de VOT (ms) dans la tâche de calibration et dans les trois épreuves de répétitions de VOT de 60 ms. (C= contrôles)

3.3 Production (stimuli avec VOT de VOT 20, 40, 60, 80, 100 ms)

La figure 2 indique les résultats dans la tâche de production en fonction de la valeur de VOT attendue, 20, 40, 60, 80 ou 100 ms. Les résultats montrent que l'ensemble des participants, sujets aphasiques et contrôles, montrent une faible augmentation des valeurs des VOT. En effet, les différences ne sont pas significatives entre les différentes valeurs de VOT obtenues en fonction des valeurs de VOT attendues pour l'ensemble des participants (tests *U* de de Mann-Whitney). Parmi les patients aphasiques, IJ et BD semblent présenter le plus de difficultés. En effet, leurs valeurs de VOT n'augmentent pas en fonction du VOT attendu. Chez TM, sa courbe suit celle des contrôles dans l'ensemble. CL par contre, semble présenter des capacités de flexibilité phonétique. En effet, bien que le patient présente une grande variabilité, les valeurs de VOT augmentent en fonction du VOT attendu, surtout entre 40 et 80 ms.

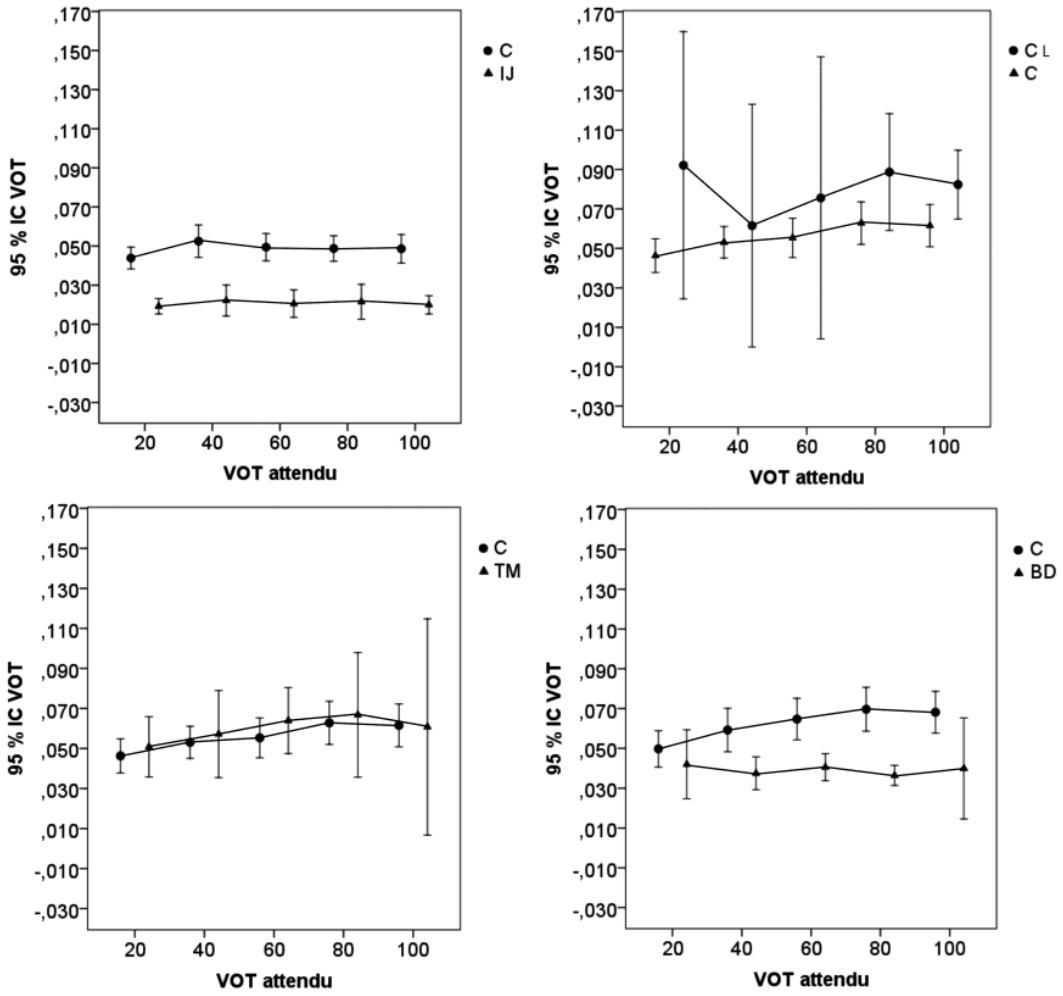


FIGURE 2: Valeurs de VOT dans la tâche de production, en fonction du VOT attendu (ms). (C= contrôles)

4 Conclusions

Dans cette étude, nous nous sommes intéressés aux capacités de flexibilité et d'apprentissage phonétique de patients aphasiques, présentant des troubles phonologico-phonétiques, notamment des difficultés de coordination entre les articulateurs. Nous leur avons plus spécifiquement proposé un paradigme destiné à examiner l'apprentissage d'une variante non familière en langue français, à savoir un non mot C[t]V[a], dont le VOT de la consonne était de 60 ms.

Les résultats montrent que, tout comme en langue française, les participants aphasiques présentent une plus grande variabilité des valeurs de VOT dans les tâches de répétitions de VOT de 60 ms. Par contre, contrairement à la tâche de répétition en français, dans laquelle les participants BD et IJ présentaient un grand nombre d'erreurs de changement de lieux d'articulation, ces erreurs ne se retrouvent plus dans la tâche de répétitions de VOT de 60 ms. Nous attribuons ces modifications au fait que la focalisation sur un seul et même non-mot, nouveau de surcroît, ait engendré moins de difficultés exécutives chez les patients, qui présentent des syndromes dysexécutifs. Parmi les patients aphasiques, trois d'entre-deux semblent présenter des capacités de flexibilité et/ou d'apprentissage phonétique. En effet, on note des valeurs de VOT plus importantes, et plus proches de la valeur cible de 60 ms, dans la tâche de répétition de la variante non familière que dans la tâche de calibration chez CL, TM et BD. Chez CL et TM, la différence a principalement lieu entre la tâche de calibration et la première répétition chez ces patients, témoignant avant tout de

capacités préservées de flexibilité phonétique. Cependant, les valeurs de VOT augmentent encore ensuite, quoique plus légèrement, entre les différentes répétitions de VOT de 60ms, indiquant un effet positif des deux tâches d'entraînement : discrimination et production. Chez BD, le résultat est différent. En effet, le patient voise les plosives non voisées dans la tâche de calibration engendrant des valeurs de VOT négatives. Ce phénomène n'apparait plus dans la tâche d'apprentissage phonétique, indiquant la possibilité d'un phénomène de modification de ses schèmes articulatoires. Cependant, le patient ne semble pas ensuite bénéficier de l'apprentissage phonétique car ces valeurs de VOT n'augmentent significativement pas entre la première tâche de répétition et des deux autres. Ces résultats indiquent que même en cas de difficultés de coordination temporelle entre les articulateurs glottiques et supra-glottiques, montrées dans la tâche de répétition de non-mots CVCV en français, les patients aphasiques sont toujours capables d'acquérir de nouveaux schèmes articulatoires.

Ces observations sont intéressantes pour la rééducation des patients aphasiques qui consiste principalement en des tâches de répétition et d'imitation destinées à réacquérir les phonèmes identifiés comme altérés et donc à se réapproprier des schèmes articulatoires. Le fait que les patients soient toujours capables d'acquérir de nouveaux schèmes malgré des difficultés de coordination entre les articulateurs constitue un élément encourageant face aux interrogations concernant la possibilité d'une rééducation potentiellement couronnée d'effets.

Références

AUZOU, P., OZSANCAK, C., MORRIS, R. J., MARY JAN, F. E., HANNEQUIN, D., JAN, M., … HANNEQUIN, D. (2000). Voice onset time in aphasia , apraxia of speech and dysarthria : a review. *Clinical Linguistics & Phonetics*, *14*(2), 131–150. DOI : 10.1080/026992000298878

BAQUÉ, L., MARCZYK, A., ROSAS, A., & ESTRADA, M. (2015). Disability, repair strategies and communicative effectiveness at the phonic level: evidence from a multiple-case study. *Neuropsycholinguistic Perspectives on Language Cognition*, (May), 144–165.

BLUMSTEIN, S. E., COOPER, W. E., GOODGLASS, H., STATLENDER, S., & GOTTLIEB, J. (1980). Production deficits in aphasia: A voice-onset time analysis. *Brain and Language*, *9*(2), 153–170. DOI : 10.1016/0093-934X(80)90137-6

CRAWFORD, J. R., GARTHWAITE, P. H., WOOD, L. T. (2010). The case controls design in neuropsychology: Inferential methods for comparing two single cases. *Cognitive Neuropsychology*, 27, 377-400.

DELVAUX, V., CANO-CHERVEL, J., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (2011). *Capacités d'apprentissage Phonétique Chez Les Sujets Âgés Francophones*, 385–393. Le Mans: XXXe édition des Journées d'Études sur la Parole (JEP 2014).

DELVAUX, V., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (2014). Phonetic compliance: a proof-of-concept study. *Frontiers in Psychology*, *5*. DOI : 10.3389/fpsyg.2014.01375

GALLUZZI, C., BURECA, I., GUARIGLIA, C., & ROMANI, C. (2015). Phonological simplifications, apraxia of speech and the interaction between phonological and phonetic processing. *Neuropsychologia*, *71*, 64–83. doi.org/10.1016/j.neuropsychologia.2015.03.007

LAGANARO, M. (2015). Paraphasies phonémiques et/ou phonétiques ? Des raisons et des difficultés de cette distinction. *Revue de Neuropsychologie*, *7*(1), 27. https://www.cairn.info/revue-de-neuropsychologie-2015-1-page-27.htm

LISKER, L., & ABRAMSON, A S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word Journal Of The International Linguistic Association*, Vol. 20, pp. 384–422. DOI : 10.1080/00437956.1964.11659830

MARCZYK, A., BAQUÉ, L., ROSAS, A., & NESPOULOUS, J. L. (2011). On the nature of speech errors in aphasia: Acoustic analysis of the speech output of 8 native speakers of spanish with aphasia. *Procedia - Social and Behavioral Sciences*, *23*(September 2015), 84–85. DOI : 10.1016/j.sbspro.2011.09.181

NESPOULOUS, J. L., BAQUÉ, L., ROSAS, A., MARCZYK, A., & ESTRADA, M. (2013). Aphasia, phonological and phonetic voicing within the consonantal system: preservation of phonological oppositions and compensatory strategies. *Language Sciences*, *39*(1), 117–125. DOI : 10.1016/j.langsci.2013.02.015

RYALLS, J., PROVOST, H., & ARSENAULT, N. (1995). Voice onset time production in French-speaking aphasics. *Journal of Communication Disorders*, *28*(1), 205–215. DOI: 10.1016/0021-9924(94)00009-O

VERHAEGEN, C., DELVAUX, V., FAGNIART, S., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (2019). Phonological and phonetic impairment in aphasic speech: an acoustic study of the voice onset time of six French-speaking aphasic patients. *Clinical Linguistics & Phonetics*, 1–21. DOI : 10.1080/02699206.2019.1619095

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Qualité vocale dans l'acquisition d'une langue étrangère : le cas des apprenants sinophones en FLE

Dongjun Wei & Mohamed Embarki
ELLIADD EA 4661
Université de Franche-Comté, Besançon – France
dongjun.wei@edu.univ-fcomte.fr, mohamed.embarki@univ-fcomte.fr

RÉSUMÉ

L'étude porte sur les configurations de la qualité vocale de huit apprenants sinophones qui parlent en mandarin dans une tâche de production de La bise et le soleil en L1 chinois et L2 français. Une comparaison est faite avec la lecture en français de quatre locuteurs natifs du français. Les corpus chinois/français sont utilisés pour recueillir les impressions d'auditeurs français sur la qualité vocale des apprenants sinophones. Des enregistrements vidéo ont été également réalisés en L1 chinois et L2 français par les mêmes apprenants. Les données subjectives récoltées conformément à la littérature indiquent des variations de configurations de la qualité vocale dans les deux langues. Les mesures acoustiques, Fo moyenne du texte lu et Fo moyenne de la voyelle [a], présentent dans les deux langues des variations ordonnées intra- et interindividuelles, entre lecture en L1 chinois et lecture en L2 français, et entre locuteurs L1 français et apprenants L2 français.

ABSTRACT

Voice quality in the second language acquisition: The case of Chinese learners of French as Foreign Language.
The study focuses on the voice quality settings of eight Chinese learners in a production task of La bise et le soleil in L1 Chinese and L2 French. A comparison is made with a production task in French of four French native speakers. Chinese / French corpora are used to collect the impressions of French listeners on the voice quality of Chinese learners. Video recordings were also made in L1 Chinese and L2 French by the same learners. The subjective data collected in accordance with the literature indicates variations in the settings of voice quality in the two languages. The acoustic measurements, mean Fo of the text reading and mean Fo of the vowel [a], present in the two languages ordered intra- and inter speaker variations, between reading in L1 Chinese and L2 French, and between L1 French speakers and L2 French learners.

MOTS-CLÉS : Qualité vocale, acquisition d'une langue étrangère, L1 chinois, L2 français, fréquence fondamentale.
KEYWORDS: Voice quality, SLA, L1 Chinese, L2 French, fundamental frequency.

1 Introduction

La voix humaine a toujours intéressé les spécialistes, son étude est très ancienne. De nos jours, différents aspects ou fonctions de la voix sont étudiés et constituent même des domaines particuliers ou des disciplines académiques reconnues. Dans le domaine de la musique et du chant par exemple, le registre vocal et l'expression artistique de la voix chantée connaissent de nombreuses études, des

Volume 1 : Journées d'Études sur la Parole, pages 617–626.
HAL : hal-02798594.

plus anciennes qui datent de l'époque romaine, aux plus récentes. Dans le domaine de l'Intelligence artificielle, le monde connaît actuellement une profusion de produits basés sur la connaissance minutieuse de la voix humaine et les applications les plus variées sont proposées. Au niveau linguistique, le message produit par le locuteur est encodé par de nombreuses informations vocales personnelles qui vont influencer l'auditeur aussi bien dans le décodage du contenu linguistique que dans l'interprétation d'éléments paralinguistiques, comme l'identité du locuteur (âge et sexe), son appartenance sociale et régionale, ses états physique et mental, ses émotions, etc. Depuis quelques années, la qualité vocale est explorée dans le domaine de l'acquisition des langues étrangères, la présente étude s'inscrit dans ce cadre.

2 État de l'art

Depuis les années 1960, les études spécifiques sur la voix humaine ont pris plus de place dans les préoccupations des phonéticiens et d'autres spécialistes, comme les psychologues, les orthophonistes. Nous pouvons citer les travaux de Honikman (1964), d'Abercrombie (1964, 1967), et de nombreux travaux de Laver (1968, 1975, 1980, 1991, 1994). Abercrombie (1967) estime que les traits de la qualité vocale ont des caractéristiques quasi permanentes qui se distinguent dans le temps des traits segmentaux à déplacement rapide et des traits vocaux dynamiques à fluctuation lente. Selon Laver (1968), la qualité vocale renvoie à la qualité quasi permanente de la voix du locuteur et peut être considérée comme provenant de la base anatomique et physiologique des organes vocaux, et des ajustements musculaires à long terme, ou configurations, du larynx et du tractus vocal supralaryngal, acquis de manière idiosyncratique, ou par imitation sociale, devenus par la suite inconscients.

Selon Nolan (1982), la qualité vocale doit être considérée comme la combinaison de trois facteurs : 1) le mécanisme de la production ; 2) le degré de contrôle ; et 3) le laps de temps. Du point de vue du mécanisme de la production, la qualité vocale comporte aussi de nombreuses significations possibles. Au sens étroit, elle fait référence à la qualité vocale produite par l'activité du larynx, en particulier l'effet perceptuel produit par le mode de vibration des plis vocaux, appelé mode phonatoire, ou l'effet auditif produit par le type de phonation. Au sens général, la qualité vocale désigne toutes les couleurs auditives uniques de la voix du locuteur, y compris la qualité produite au niveau du larynx et des résonateurs supralaryngés (Laver, 1980). En ce qui concerne le degré de contrôle du locuteur, Laver (1968) estime que la qualité vocale provient de deux sources différentes, à savoir une source interne et une source externe, qui correspondent respectivement aux caractéristiques anatomiques intrinsèques et aux configurations externes (ou settings). En ce qui concerne le laps de temps, Lhote (1982 : 258) indique que « chaque locuteur tend à utiliser ses propres ajustements musculaires à long terme […] ou "setting", et ces settings caractérisent de façon quasi permanente la couleur de sa voix ». Laver et Trudgill (1979) ont distingué en fonction du laps de temps trois aspects de la performance vocale dans la parole : l'aspect à long terme, à moyen terme et à court terme. Selon Esling & Wong (1983), les paramètres de la qualité vocale correspondent aux postures, sur le long terme, du larynx, du pharynx, de la langue, du système vélopharyngé, des lèvres ainsi que les configurations laryngales qui se manifestent dans les divers types de phonation.

3. Catégorisation de la qualité vocale

Une grande partie de la variabilité de la parole relève du domaine de la qualité vocale, car elle provient principalement de sources organiques, telles que la longueur et la forme du tractus vocal, la taille de la mâchoire et de la langue, ainsi que la longueur, la masse et la viscosité des plis vocaux (Goldstein, 1980 ; Fitch & Giedd, 1999 ; Titze, 1989). Ces différences anatomiques et

physiologiques ont une influence directe sur la fréquence fondamentale *(Fo*) (Holmberg & al., 1988 ; Titze, 1989) et sur les fréquences centrales des premiers formants (Peterson & Barney, 1952 ; Hillebrand & al., 1995), qui sont généralement plus élevées chez les femmes que chez les hommes. Une deuxième source de variabilité provient de la manière dont le système de production de la parole est utilisé par le locuteur. Des configurations assez fines et subtiles du tractus vocal, utilisées habituellement ou de manière ponctuelle et intentionnelle, vont produire une qualité vocale particulière à laquelle les auditeurs attribuent des catégories, comme grave vs aiguë, rauque vs cristalline, forte vs fluette, rieuse vs morne. Certaines de ces catégories caractérisent le volume ou le timbre de la voix, d'autres l'état physique ou moral.

Les liens entre la voix prise comme un tout et le décodage de la parole sont très nombreux. Bradlow & Pisoni (1999) ont montré le lien entre habituation de l'auditeur à la qualité vocale du locuteur et la réussite dans l'identification des mots difficiles, même avec un débit de parole rapide. Certaines études ont montré que les voix masculines, avec une *Fo* basse, étaient perçues plus agréables, attractives et persuasives (Brown & al., 1973 ; Zuckerman & Miyake, 1993), tandis que l'augmentation de la *Fo* provoque la perception du locuteur comme moins compétent, moins bienveillant, moins sincère, moins persuasif, plus faible et plus nerveux. Depuis, les corrélats acoustiques d'une voix attractive ont été identifiés dans plusieurs langues. Pour des auditeurs anglophones, une voix féminine attractive doit présenter une *Fo* élevée, une dispersion formantique large, et elle doit être soufflée. Bruckert et Sanguin-Bruckert (2013) ont sélectionné parmi 105 voix d'une base de données 12 voix différentes lisant un texte d'environ 1 mn, 6 voix jugées agréables et 6 voix jugées désagréables (6 féminines et 6 masculines). Le choix a été fait par plus de 120 auditeurs. Ces derniers devaient accomplir trois tâches de compréhension. Les résultats montrent que les auditeurs qui ont écouté une voix jugée agréable ont réalisé des performances en compréhension plus importantes que les auditeurs qui ont écouté une voix jugée désagréable.

Ces multiples recherches montrent que la qualité vocale joue un rôle important dans la communication. Cependant, les travaux sur la qualité vocale dans l'acquisition d'une langue étrangère sont moins nombreux. Oh (2011) a mené une étude sur 20 sujets coréens avec un bon niveau en L2 chinois. L'expérimentation a porté sur la voyelle /a/ dans les deux langues, coréen et chinois. Les résultats montrent que la Fo moyenne passe chez les 10 femmes de 219 Hz en coréen L1 à 251 Hz en chinois L2 et chez les 10 hommes de 119 Hz en coréen à 131 Hz en chinois. Cette étude montre que le même locuteur change de Fo moyenne : l'apprenant d'une L2 augmente automatiquement sa Fo moyenne de L1 pour atteindre une qualité vocale spécifique à la L2.

Dans le domaine du Français Langue Etrangère (FLE), Pillot-Loiseau et al. (2012) ont mené une étude sur l'évaluation des ressentis des changements vocaux entre le français et la LM, auprès de 114 apprenants de FLE, en complément d'une analyse déjà effectuée auprès de 312 sujets bilingues dont 20% d'enseignants de FLE. 80,7% des apprenants ressentent une *Fo* différente entre leur production du français et celle de leur LM, et 71,6% ressentent une différence analogue concernant l'intensité. Quand les apprenants parlent en français : 1) 40% se plaignent d'une douleur à la gorge ou à la mâchoire ; 2) 38% ressentent une voix fatiguée et cassée ; et 3) 22% déclarent forcer leur voix, et 58 % déclarent forcer leur articulation. Dans une autre étude, Pillot-Loiseau (2013 : 19) a conclu que « 63% de sinophones ressentent une voix plus haute en français ». L'auteure a conclu qu'une « fréquence fondamentale significativement plus grave en chinois mandarin qu'en français en lecture (p=0.01) … Trois femmes bilingues mandarin/ français ont effectivement ressenti cette différence ».

Dans ce travail, nous allons explorer objectivement deux paramètres, la *Fo* et la vitesse d'articulation dans l'apprentissage du FLE par des apprenants d'origine chinoise. Nous examinerons en particulier chez les mêmes sujets des productions en français comparées d'une part à leurs

productions en langue maternelle, et d'autre part à la production du même corpus en français par des locuteurs natifs du français. Nous explorerons également l'impression que les auditeurs français natifs ont de ces productions.

4 Méthodologie

Les données présentées sont de deux natures : des données objectives fondées sur des mesures acoustiques et des données subjectives, car issues des seules impressions des auditeurs. Les premières portent sur les différences de qualité vocale chez huit apprenants chinois produisant un texte en chinois puis en français. Il s'agira de présenter des comparaisons intra- et inter-individuelles internes au groupe de langue maternelle chinoise. Les données en français des locuteurs chinois sont comparées à celle de quatre locuteurs français natifs produisant le même corpus. Ces mesures objectives sont complétées par des mesures subjectives sur la perception de la qualité vocale par un groupe d'auditeurs français natifs.

4.1 Public

Nos sujets ont été divisés en trois groupes : 1) le groupe expérimental comprend des Chinois qui apprennent le français au Centre linguistique Appliqué (CLA) de Besançon. Ils sont au nombre de huit, trois hommes et cinq femmes, leurs niveaux de langue certifiés sont différents selon les échelles du CECR (2005) : deux apprenants féminins avec un niveau A2, un apprenant masculin et un apprenant féminin de niveau B1, deux apprenants masculins de niveau B2 et deux apprenants féminins de niveau C1 ; 2) le groupe témoin : il comprend quatre sujets de langue maternelle française (deux hommes et deux femmes) ; 3) le groupe d'auditeurs composé de sujets français (trois femmes et deux hommes). Tous les groupes ont un niveau d'études supérieures, âgés de 20 à 25 ans, et ils ne présentent aucun problème apparent ni d'élocution, ni d'audition.

4.2 Corpus et matériel de mesure

Le texte utilisé correspond aux versions française et chinoise de « La bise et le soleil ». Nous voulions un texte court, cohérent et complet d'un point de vue de la signification et comprenant peu de difficultés lexicales. La version chinoise comporte 141 mots, la version française comporte 125 mots, toutefois le nombre de syllabes dans chaque langue est quasi similaire (le chinois a tendance à avoir des mots monosyllabiques). Les Chinois lisent le texte respectivement en chinois et en français et les Français le lisent uniquement en français. Nous avons utilisé le logiciel Praat pour l'analyse des données acoustiques. Pour les données subjectives, nous avons eu recours à un questionnaire portant sur le profil linguistique des apprenants chinois, ainsi qu'à un questionnaire permettant de rendre compte du jugement porté sur l'impression de la voix des sujets.

4.2.1 Questionnaire sur le profil linguistique des apprenants chinois

Nous avons choisi comme base le questionnaire du profil linguistique de Pillot-Loiseau & al. (2012 : 47-49). Notre questionnaire porte principalement sur des informations personnelles des apprenants chinois et sur leur expérience d'apprentissage du français. Il comprend dix éléments : les quatre premiers éléments concernent les informations personnelles (sexe, âge, niveau d'étude, temps passé en France) et les six derniers éléments concernent leur expérience d'apprentissage du français, qui inclut le niveau de cours de langue actuel, la durée de l'apprentissage du français en Chine, s'il y avait un cours de prononciation en Chine, la durée de l'apprentissage du français en France, s'il y a un cours de prononciation et s'il est fréquent de pratiquer le français en dehors des cours de langue.

4.2.2 Questionnaire d'auto-évaluation

Sur la base du questionnaire de ressenti de Pillot-Loiseau et al. (2012), nous en avons adapté une partie portant principalement sur les modifications de la hauteur, de l'intensité, du débit et des organes vocaux. Le questionnaire comprend cinq aspects : 1) l'impression de changement de hauteur ; 2) le changement d'intensité ; 3) le changement de débit ; et 4) et 5) les jugements portés sur les organes vocaux.

4.2.3 Tableau de comparaison de la configuration de la voix

Le modèle VPA (Vocal Profile Analysis) de Laver et al. (1981) pour l'orthophonie qui n'est pas utilisé par les orthophonistes français est très complexe et certaines mesures nécessitent beaucoup d'expériences pratiques. Afin d'appliquer au mieux ce modèle à notre public, nous n'avons retenu que des traits simples à mesurer d'un point de vue expérimental, notamment les traits labiaux que nous avons estimés qualitativement, les traits de la voix dynamique et les traits de l'organisation temporelle. Ce tableau est utilisé pour comparer le degré de changement des configurations lorsque les apprenants chinois parlent français par rapport à leur production en chinois ou lorsque les Français parlent français. Le résultat de la comparaison est divisé en trois types : la note « -1 » signifie que le changement du trait est plus faible, bas ou lent, « 0 » signifie que le changement du trait n'est pas évident, « 1 » signifie que le changement du trait est plus fort, haut ou rapide.

4.2.4 Procédure

Pour les huit apprenants chinois, nous avons réalisé deux enregistrements : la lecture en chinois et la lecture en français. Pour les quatre Français, nous avons réalisé un seul enregistrement, *i.e.* la lecture en français. Les enregistrements audio ont été complétés dans un second temps par des enregistrements vidéo du même corpus réalisés dans les mêmes conditions pour les deux groupes. Nous avons suivi exactement le même ordre, lecture en chinois, puis ensuite en français pour le groupe chinois, lecture en français uniquement pour les quatre Français. Les données subjectives composées de jugements des locuteurs chinois, d'auditeurs français et des jugements personnels des auteurs de cette étude, sur les productions sont prises ici à titre indicatif.

5 Résultats

5.1 Données subjectives

5.1.1 Impressions des Chinois sur leur propre voix

Pour le jugement de hauteur, 7 apprenants chinois sur 8 ressentent un léger changement de hauteur lorsqu'ils parlent français : 2 sujets féminins de niveau A2 trouvent que la hauteur est légèrement élevée en passant de L1 à L2, 5 apprenants trouvent au contraire qu'elle baisse (1 sujet masculin de niveau B1, 2 sujets masculins de niveau B2 et 2 sujets féminins de niveau C1), 1 sujet féminin de niveau B1 ne ressent pas de changement entre les deux langues. Pour l'intensité, 5 sujets sur 8 (3 de niveaux B2 et C1, 2 de niveaux A2 et B1) ont l'impression que leur voix est légèrement plus faible lorsqu'ils parlent français. Il y a 3 sujets sur 8 (de niveaux A2, B1 et B2) qui ne ressentent pas de changements d'intensité. Pour le débit, tous les apprenants (8 sur 8) ont la sensation de s'exprimer plus lentement en français. Parallèlment, 2 sujets sur 8 ont déclaré ressentir de la fatigue en parlant français, contre 6 sur 8 des apprenants qui ne ressentent pas de fatigue en s'exprimant en français. En ce qui concerne le forçage articulatoire, 8 sujets sur 8 ont indiqué que, dans la plupart des cas, ils ne forçaient pas intentionnellement leur articulation pour être mieux compris par leurs interlocuteurs.

Toutefois, 4 apprenants chinois ont concédé forcer certains gestes articulatoires sur des mots spécifiques pour mieux se faire comprendre.

5.1.2 Impressions des Français sur la voix des apprenants chinois

Pour les impressions des auditeurs français, nous avons sélectionné parmi les 8 apprenants chinois les enregistrements de 4 apprenants : une femme de niveau A2, un homme de niveau B1, un homme de niveau B2, une femme de C1. Ce choix a été fait d'une part pour la comparabilité entre sexes et d'autre part pour éviter le caractère répétitif des mêmes commentaires concernant deux sujets de même niveau de langue. Les cinq auditeurs français ont jugé de manière qualitative la lecture en français de chaque apprenant chinois comparée à la lecture en chinois du même locuteur. L'ordre de présentation des enregistrements aux auditeurs était aléatoire, sans relation avec le niveau de langue des locuteurs. Selon les réponses, les différences entre les quatre apprenants chinois se reflètent principalement dans cinq aspects : la hauteur, la tonalité, l'intonation, le rythme et le débit. Nous rappelons ici que le rythme et le débit ont été intégrés dans l'étude de la qualité vocale conformément aux premiers travaux de Laver (1968, 1980, 1991) et aux travaux d'Abercrombie (1964, 1967).

Dans l'ensemble, les auditeurs français estiment que les locuteurs chinois utilisent en français moins de variation mélodique, ou tonalité. Ils estiment que la voix des sujets féminins devient plus grave, alors que la voix des apprenants masculins devient plus aiguë. Les auditeurs trouvent que la qualité vocale est relativement homogène pour les apprenants les moins avancés des niveaux A2 et B1 ; de même ils jugent de manière uniforme et positive les apprenants les plus avancés des niveaux B2 et C1. Les impressions des auditeurs français vis-à-vis de la qualité vocale des apprenants chinois des niveaux B2 et C1 semble meilleure que celle des apprenants des niveaux A2 et B1, surtout au niveau du rythme et du débit. Par conséquent, nous pouvons en déduire que plus le niveau de compétence en français augmente, plus les auditeurs natifs du français estiment que l'apprenant chinois semble garder la même voix dans les deux langues.

5.1.3 Impressions des deux auteurs sur la qualité vocale des apprenants chinois

Concernant la hauteur, nous avons écouté les enregistrements plusieurs fois, sans recourir à un outil d'analyse de la parole. Nos impressions sont aussi naïves et intuitives que celles des deux autres groupes présentés ci-dessus. Nos impressions rejoignent celles des locuteurs chinois sur leurs propres productions, *i.e.* un sujet avec une hauteur identique dans les deux langues, 2 apprenants sur 8 présentant un niveau de hauteur plus élevé en français et 5 apprenants sur 8 avec un niveau plus bas lors de la lecture en français. Selon nos impressions, le niveau de hauteur plus bas en français qu'en chinois concerne davantage les sujets féminins, et l'inverse pour les sujets masculins.

Nous avons utilisé les enregistrements vidéo pour estimer de manière qualitative l'activité labiale des sujets. Considérant qu'il y a plus de voyelles labialisées en français qu'en chinois, nous devrions logiquement y observer une activité labiale plus conséquente. Or, selon nos propres relevés, la protrusion labiale est moins importante en lecture en français chez 5 apprenants sur 8, parmi eux 4 sont des apprenants des niveaux A2 et B1. Nous estimons que l'activité labiale est identique dans les deux langues chez 2 apprenants sur 8, et elle n'est plus importante en français que chez un seul apprenant. Et pour l'étirement des lèvres, aucun apprenant n'a changé de manière significative en lisant le texte en français. Ils ont gardé la même configuration dans les deux langues.

Pour la configuration de l'intensité, nous n'avons pas perçu de changement significatif pour l'ensemble des apprenants chinois. Pour les configurations de continuité et de débit, tous les apprenants semblent présenter en français une lecture plus saccadée et hachée et un débit de parole

plus lent. Cependant, la situation des apprenants des niveaux B2 et C1 est meilleure que celle des apprenants des niveaux A2 et B1, ils font moins de pauses et ils présentent un débit plus rapide. Certains parmi ces derniers présentent des configurations presque similaires dans les deux langues.

5.2 Mesures acoustiques

Nous avons segmenté et étiqueté sous Praat tout le corpus, en français et en chinois. Nous avons ensuite mesuré la *Fo* moyenne des locuteurs chinois en lecture dans les deux langues et celle des locuteurs français en LM. Conformément à la littérature, nous avons choisi d'estimer la *Fo* de la voyelle [a], en prenant les mesures au centre de la voyelle, toujours sur la partie stable. Nous avons limité les mesures de la *Fo* de la voyelle à quatre sujets chinois sur huit, deux de chaque sexe. Pour le groupe témoin, nous avons utilisé les enregistrements des quatre sujets, que ce soit pour la *Fo* moyenne du texte ou de la voyelle [a]. Concernant les mesures de la *Fo* moyenne de la voyelle, nous avons retenu cinq occurrences de [a] par locuteur et par langue, ce qui donne un total de 20 occurrences de [a] par langue chez les apprenants chinois de FLE et 20 occurrences en français pour le groupe témoin. Pour la voyelle [a] en chinois, nous n'avons pas fait de distinction en fonction de la nature du ton affectant la syllabe, bien que nous soyons conscients du fait que le contraste tonal se fait essentiellement par des variations de la Fo.

Les résultats montrent que les apprenants de sexe masculin ne présentent pas de différence sur la *Fo* entre la lecture du texte en chinois et en français, avec une moyenne de 126 Hz (Δ=22) et 125 Hz (Δ=7) respectivement. Pour les mesures de Fo prises au centre de la voyelle [a], les sujets masculins présentent une Fo moyenne plus élevée en chinois (124 Hz ; Δ=6) qu'en français (115 Hz ; Δ=3). Comparativement, les deux locuteurs français présentent une Fo moyenne pour le texte de 168 Hz (Δ=30) pour le Loc. 1 et 134 Hz (Δ=20) pour le Loc. 2. Leur Fo moyenne pour la voyelle [a] est de 157 Hz (Δ=18) pour le premier et 93 Hz (Δ=11) pour le second. Chez les apprenants de sexe féminin, la moyenne globale de la lecture en français est légèrement supérieure à la moyenne en chinois, 218 Hz (Δ=25) et 213 Hz (Δ=26) respectivement. Cependant, la Fo moyenne de la voyelle [a] en français est inférieure à celle de la même voyelle en chinois, 222 Hz (Δ=12), contre 259 Hz (Δ=16) respectivement. Comparativement, les deux locutrices françaises présentent une Fo moyenne pour le texte de 251 Hz (Δ=46) pour la locutrice 1 et 205 Hz (Δ=16) pour la locutrice 2. Leur Fo moyenne de la voyelle [a] est de 207 Hz (Δ=68) et 189 Hz (Δ=55) respectivement.

Nos résultats, qui ne prétendent en aucun cas à une quelconque généralisation, laissent apparaître des différences de trois natures : 1) des différences parmi les locuteurs chinois intra-individuelles *vs* interindividuelles ; 2) des différences inter langues chinois *vs* français ; et 3) des différences entre locuteurs natifs du français *vs* apprenants de FLE. Concernant les premières différences, la variabilité intra-individuelle est très limitée en chinois comparées aux mêmes productions en français ; la variabilité interindividuelle est également limitée, pour preuve les valeurs de l'écart-type qui sont très faibles. Concernant les différences entre lecture en chinois et lecture en français par les mêmes sujets, nos valeurs divergent entre locuteurs masculins et féminins : la baisse de la *Fo* moyenne pour le texte chez les locutrices et la stabilité chez les locuteurs en passant de L1 à L2. Concernant les troisièmes différences, globalement, la *Fo* moyenne pour le texte en français chez les locutrices chinoises est intercalée entre les deux moyennes des locutrices françaises tandis que la *Fo* moyenne de la voyelle [a] est supérieure chez les Chinoises, comparées aux deux Françaises. La *Fo* moyenne pour le texte en français chez les locuteurs chinois de sexe masculin est inférieure à celle de leurs homologues français. Concernant la *Fo* moyenne de la voyelle [a], le pattern est inversé, les valeurs chez les Chinois sont intermédiaires, comparées à celles des deux Français.

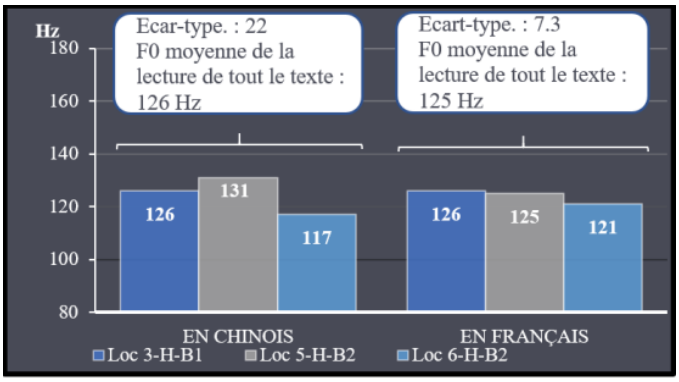


FIGURE 1 : *Fo* moyenne de la lecture du texte en chinois et en français par 3 apprenants masculins.

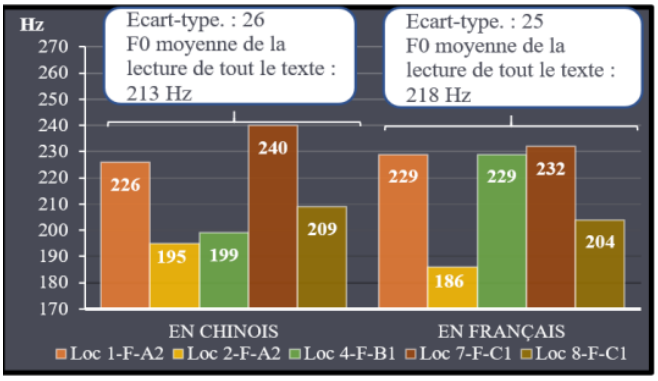


FIGURE 2 : *Fo* moyenne de la lecture du texte en chinois et en français par 5 apprenants féminins.

6 Discussion

L'étude de Pillot-Loiseau (2013 : 19) a conclu que « 63% sinophones ressentent une voix plus haute en français », cependant nos résultats ne permettent pas de le confirmer. Le quart de nos apprenants chinois ressent effectivement une hauteur plus élevée dans la lecture en français, tandis que 5 sur 8 ressentent le contraire. En comparaison avec notre étude, nos résultats divergent de ceux de Pillot-Loiseau (2013 : 20). Par conséquent, nous pensons qu'il existe un grand décalage entre le jugement subjectif de la hauteur et la mesure objective de la *Fo*. Les deux ne peuvent se substituer l'un à l'autre. L'autre problème est que dans nos expériences subjectives, le jugement des apprenants chinois sur leur propre hauteur ne correspond pas au jugement des Français sur la hauteur des apprenants chinois et ne coïncide pas également avec notre jugement de configuration de leur hauteur. Mais nos expériences subjectives ont montré que la plupart des apprenants chinois sont capables de ressentir le changement de hauteur quand ils parlent en français et en chinois, mais ils ne peuvent pas décrire avec précision ce changement. Parallèlement, les sujets français peuvent également reconnaître ce changement, bien que leurs descriptions soient plus précises que les descriptions des apprenants chinois, ils sont également incapables de fournir des descriptions détaillées. En outre, nous ne pouvons trouver de corrélation entre le niveau de français et la configuration de la hauteur chez les apprenants chinois, mais ce point reste à approfondir avec davantage de sujets et avec des niveaux de compétence différents en langue.

Concernant la voyelle [a], nos mesures permettent de dire que tous les locuteurs chinois voient leur Fo moyenne baisser en passant de L1 à L2. Ce constat contredit en apparence les résultats de l'étude menée par Oh (2011) sur la voyelle [a] en L1 coréen et L2 chinois. Certains pourraient en conclure hâtivement que les deux études donnent des résultats opposés, la *Fo* moyenne augmente en passant de L1 à L2 dans l'étude de Oh (2011) tandis qu'elle baisse dans notre étude en passant de L1 à L2. Ce paradoxe pourrait être expliqué en grande partie par le système tonal chinois. Dans la présente étude, les locuteurs chinois présentent une *Fo* élevée sur la voyelle en chinois car celle-ci est le domaine de l'actualisation tonale. En revanche, les mêmes sujets présentent une *Fo* moyenne plus basse en français car ils ont conscience que cette langue ne distingue pas le sens par des modulations mélodiques sur le domaine strict de la voyelle. Les apprenants coréens de chinois ont vraisemblablement conscience eux aussi de cette particularité, d'où l'augmentation de leur *Fo* moyenne en L2 chinois. La baisse vs l'augmentation de la *Fo* en passant de L1 à L2, contradictoires en apparence, obéissent à nos yeux aux mêmes contraintes. Les sujets ont conscience de la valeur phonologique d'une *Fo* moyenne élevée en chinois, qu'ils soient de langue maternelle chinoise ou de simples apprenants de cette langue.

Notre étude, qui a un caractère purement exploratoire sur la qualité vocale en L1 et L2, permet d'esquisser quelques pistes de recherche intéressantes pour la phonétique, la phonologie, l'acquisition des langues et la didactique. Elle pose entre autres la question des ajustements de la qualité vocale en L1 et les changements opérés lors de l'apprentissage d'une L2. Les pistes esquissées ici pourraient susciter des réflexions en didactique sur la prise en compte de la voix dans la production des apprenants de FLE.

7 Conclusion

Les résultats de l'expérience montrent que, comparativement aux configurations de la qualité vocale utilisées en chinois, les apprenants chinois montrent des différences importantes dans certaines configurations lors de leur production en français. Les données concernant l'activité labiale (arrondissement des lèvres ou leur étirement), le débit de parole et l'estimation de la hauteur diffèrent entre L1 et L2. Cependant, les mesures objective et subjective de la configuration de la hauteur vocale sont divergentes. Quelques tendances fragiles concernent la corrélation entre le niveau de compétence en français et la qualité vocale, car certains sujets avancés manifestent en français certaines configurations ressemblant aux locuteurs natifs de français. Nous avons montré qu'il existe bel et bien des différences de la qualité vocale chez les apprenants chinois lors de leur lecture en français et en chinois, même si ces différences ne sont pas conformes aux résultats de la littérature.

Références

ABERCROMBIE D. (1964). *English phonetic texts*. London, Faber & Faber

ABERCROMBIE D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

BRADLOW A.R. & PISONI D.B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America*, 106: 2074–2085.Doi: 10.1121/1.427952

BROWN B. L., STRONG W. J. & RENCHER A.C. (1973). Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *Journal of the Acoustical Society of America*, 54, 29-35. DOI : 10.1121/1.1913571

BRUCKERT L. & SANGUIN-BRUCKERT C. (2013). Le rôle de la voix de l'enseignant dans les situations d'apprentissage. 6ème Colloque international du RIPSYDEVE ; *Actualités de la Psychologie du Développement et de l'Éducation*, May 2013, France. pp.67-72. HAL : hal-01018635.

ESLING J.H. & WONG R.F. (1983). Voice Quality Settings and the Teaching of Pronunciation. *TESOL Quarterly*, 17: 89-95. DOI : 10.2307/3586426

GOLDSTEIN U. G. (1980). An articulatory model for the vocal tract of the growing children, *Thesis of Doctor of Science*, MIT, Cambridge, Massachusett.

HILLEBRAND J., GETTY L.A., CLARK M.J. & WHEELER K. (1995). Acoustic characteristics of American English vowels., *Journal of the Acoustical Society of America*, 97, 3099-3111. DOI : 10.1121/1.411872

HOLMBERG E.B., HILLMAN R.E. & PERKELL J.S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the Acoustical Society of America,* 84, 511. DOI : 10.1121/1.396829

HONIKMAN B. (1964). Articulatory settings. In D. ABERCROMBIE et al., Éds., *In Honour of Daniel Jones*, London: Longman , 73-84.

FITCHE W.T. & GIEDD J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106, 1511. DOI: 10.1121/1.427148.
LAVER J. (1968). Voice quality and indexical information. *International Journal of Language & Communication Disorders, 3*, 43-54. DOI : 10.3109/13682826809011440
LAVER J. (1974). Labels for voices, *Journal of the International Phonetic Association, 4*, 62-75 (repris dans J. Laver 1991, p. 171-183). DOI : 10.1017/S0025100300001031
LAVER J. (1975). *Individual features in voice quality.* Ph.D. dissertation, University of Edinburgh.
LAVER J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.
LAVER J. (1991). *The Gift of Speech: Papers in the Analysis of Speech and Voice*. Edinburgh University Press.
LAVER J. (1994). *Principles of Phonetics.* Cambridge University Press.
DOI : 10.1017/CBO9781139166621
LAVER J.& TRUDGILL P. (1979). *Phonetic and linguistic markers in speech*. In Scherer, KLAUS R. & GILES H. EDS. *Social Markers in Speech*. Cambridge University Press: 1–32.
LAVER J., WIRZ S., MACKENZIE B. & HILLER S. (1981). A perceptual protocol for the analysis of vocal profiles. *Work in Progress, Department of Linguistics, University of Edinburgh,* 14, 139– 155.
LHOTE E. (1982). *La parole et la voix*, Hamburg, Buske : 228-353
NOLAN F. (1982). John Layer, The phonetic description of voice quality. Cambridge: Cambridge University Press. Pp. ix 186. *Journal of Linguistics,* 18(2), 442-454.
DOI : 10.1017/S0022226700013724
OH H. (2011). Voice Quality Processing Strategy of Korean Learners of Chinese. *Proceed. of XVIIth ICPhS*, Hong Kong, August 17-21, 2011, 1526-1529: electronic edition @ internationalphoneticassociation.org
PETERSON G. E. & BARNEY H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184. DOI : 10.1121/1.1906875
PILLOT-LOISEAU C., BENOIST-LUCY A. & VAISSIÈRE J. (2012). Fréquence fondamentale moyenne, qualité vocale et bilinguisme : quelles implications pour la rééducation vocale ? *12èmes Rencontres Internationales d'Orthophonie*, Dec 2012, Montrouge, France. p.37-78. HAL : hal-00748693.
PILLOT-LOISEAU C. (2013). Travail de la voix dans la langue : le cas de la prononciation du Français Langue Etrangère. *La langue, la voix, la parole*, Paris, France, p.17-23. HAL : hal-00862340.
TITZE I.R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85, 1699-1707. DOI : 10.1121/1.397959
TITZE I. R. (1994). *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall.
ZUCKERMAN M. & MIYAKE K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal Behavior*, 17, 119-135. DOI : 10.1007/BF01001960

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Réduction temporelle en français spontané : où se cache-t-elle ?

Une étude des segments, des mots et séquences de mots fréquemment réduits

Yaru Wu[1, 2] Martine Adda-Decker[1, 3]
(1) Laboratoire de Phonétique et Phonologie (LPP), UMR7018, CNRS, France
(2) Modèles, Dynamiques, Corpus (MoDyCo), UMR 7114, CNRS, France
(3) LIMSI-CNRS bât. 508, BP 133, 91403 Orsay cedex, France
yaru.wu@sorbonne-nouvelle.fr, madda@limsi.fr

RÉSUMÉ

Cette étude vise à proposer une méthode adaptée à l'étude de divers phénomènes de variation dans les grands corpus utilisant l'alignement automatique de la parole. Cette méthode est appliquée pour étudier la réduction temporelle en français spontané. Nous proposons de qualifier la réduction temporelle comme la réalisation de suites de segments courts consécutifs. Environ 14% du corpus est considéré comme réduit. Les résultats de l'alignement montrent que ces zones impliquent le plus souvent plus d'un mot (81%), et que sinon, la position interne du mot est la plus concernée. Parmi les exemples de suites de mots les plus réduits, on trouve des locutions utilisées comme des marqueurs discursifs.

ABSTRACT

Temporal reduction in spontaneous French : where is it hidden ? A study of frequently reduced segments, words and word sequences

This study aims to propose a method to explore large corpora using automatic speech alignment and to apply this method to the special case of reduction in spontaneous French. By locating sequences of short segments of at least three 30 or 40 ms segments, we were able to identify potential reduction zones. 14% of the corpus is considered as temporally reduced. Short segment sequences are most often observed in cross-word position (81%) rather than in single words. In the latter, the corresponding phone segments are frequently located in word-internal position. The identified reduced sequences often concern phrases used as discourse markers, which carry very little semantic content in real-life communication.

MOTS-CLÉS : français spontané, réduction temporelle, durée segmentale, suite de segments courts, grand corpus.

KEYWORDS: spontaneous French, temporal reduction, segmental duration, sequence of short segments, large corpus.

1 Introduction

La variation de la parole est très présente en parole continue (Duez, 1997; Ernestus, 2000; Duez, 2003; Johnson, 2004; Meunier & Espesser, 2011). Dans la communication quotidienne, les mots sont souvent articulés avec moins de précision et les segments sont affaiblis par rapport à une forme standard. Grâce au traitement automatique de la parole et à la quantité

Volume 1 : Journées d'Études sur la Parole, pages 627–635.
HAL : hal-02798596.

croissante de données accessibles, nous sommes aujourd'hui en mesure d'explorer la parole spontanée et d'étudier des phénomènes variés globalement, sans forcément formuler une hypothèse précise et recueillir des données contrôlées afin d'examiner cette hypothèse en question. Les outils de traitement automatique permettent ainsi des approches différentes de l'analyse traditionnelle basée sur des hypothèses sur un ensemble de données contrôlées.

Un des deux objectifs de cette étude est de suggérer une méthode utilisant l'alignement automatique de la parole afin d'aider à localiser des zones de réduction temporelle. Nous visons également à mieux comprendre où se situent ces zones. Sont-elles principalement situées au sein d'un mot ou vont-t-elle au-delà de la frontière des mots ? Y a-t-il une position dans le mot qui favorise la réduction ? Notre hypothèse de travail ou notre question d'intérêt consiste donc dans le fait que la parole spontanée contienne de nombreux phénomènes de réduction temporelle. Par réduction temporelle nous entendons des zones de parole où les prononciations des mots seraient seulement partiellement réalisées et qu'elles contiendraient potentiellement moins de segments que le nombre prévu par une prononciation canonique. Cette question est intéressante à notre avis non seulement pour les technologies vocales, comme la reconnaissance ou la synthèse de la parole, mais également pour l'apprentissage des langues et pour mieux comprendre le traitement cognitif de la parole.

2 Corpus et alignement

Le corpus NCCFr (Nijmegen Corpus of Casual French, Torreira *et al.*, 2010) a été utilisé dans cette étude. Ce corpus est composé de 36 heures d'enregistrements de conversations entre amis, dont 24 femmes et 22 hommes.

Les données du corpus ont été automatiquement segmentées et étiquetées à l'aide du système de reconnaissance automatique de la parole au LIMSI (Gauvain *et al.*, 2002) en mode d'alignement forcé. Des modèles acoustiques de phones indépendants du contexte (HMM de phones estimés à partir de segments d'au moins 50 ms) ont été utilisés afin d'établir au mieux la correspondance entre les segments de parole et les modèles HMM selon la ou les transcriptions phonémiques proposées par le dictionnaire de prononciation du système. Des variantes de prononciation peuvent être ajoutées à ce dictionnaire pour qu'elles soient mieux adaptées à la production réelle des locuteurs. Le système produit, à la sortie de l'alignement, les étiquettes des mots et des phones, ainsi que les frontières respectives. Le système fournit également des étiquettes et les frontières pour les tronçons de signal hors parole, tels que les pauses, la respiration et le bruit. La durée minimale d'un segment est de 30 ms, ce qui correspond à 3 trames acoustiques (Adda-Decker & Lamel, 2000). Le dictionnaire utilisé contient comme variantes systématiques la liaison et le schwa.

3 Méthodologie

Dans la suite, nous décrivons comment nous proposons de qualifier les zones de réduction et la méthode adoptée afin de les localiser automatiquement dans le signal de parole.

3.1 Méthode ascendante

Nous avons décidé de localiser les phénomènes de réduction temporelle en exploitant les caractéristiques du système d'alignement forcé : ce dernier cherche à associer à chaque phone prévu dans la prononciation d'un mot une portion de signal, dont la durée minimale est au moins de 30 ms. Lorsque, dans le signal observé, certains phones prévus par la prononciation du dictionnaire ne sont pas ou presque pas présents, le système force quand-même l'alignement de ces phones, ce qui engendre, comme résultat de l'étape d'alignement, une séquence de plusieurs segments courts (de 30 ou 40 ms) à la suite. Ce défaut du système d'un point de vue de la précision de l'étiquetage et de la segmentation en phones, peut être exploité comme qualité afin de localiser ces phénomènes de réduction temporelle. Ainsi, nous qualifions comme zone de réduction temporelle une séquence de plusieurs segments courts (30 ou 40 ms) consécutifs produits lors de l'alignement forcé. Ainsi, nous sommes en mesure de séparer les segments dans nos données en deux parties : les segments « normaux » (« Nrm ») et les segments « d'alerte potentiellement réduits » (« Alrt »). Nous ajoutons une contrainte : les segments courts sont étiquetés comme « Alrt », uniquement s'ils sont à l'intérieur d'une séquence d'au moins 3 segments courts consécutifs. Cette méthode, que nous appelons « méthode ascendante », nous permet d'identifier les mots et séquences de mots qui sont réduits d'après ces critères. Par exemple, la séquence de mots « je ne sais pas » (/ʒənəsɛpa/, 4 syllabes en français standard) peut être prononcée [ʃpa] (séquence monosyllabique). Une telle prononciation n'est pas modélisée de manière satisfaisante par le système. L'alignement le mettra en évidence en forçant la présence de tous les segments mais en leur attribuant des durées très courtes.

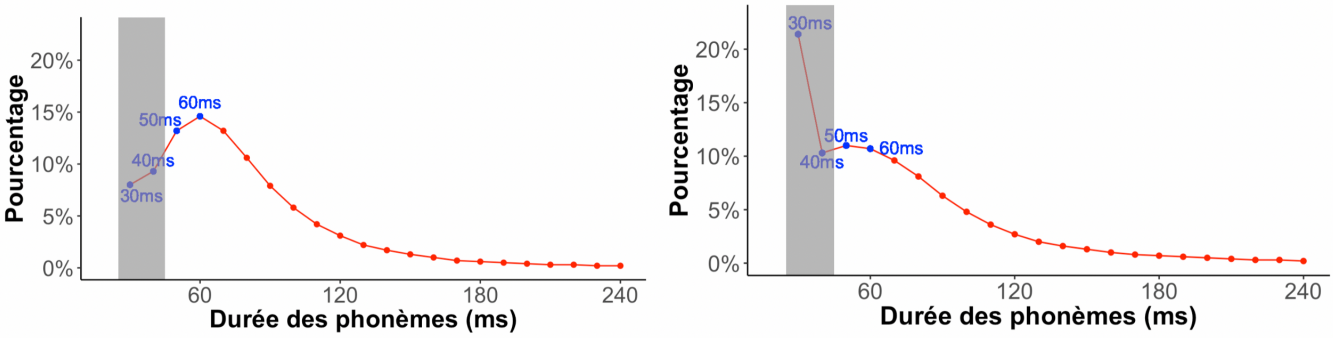


FIGURE 1 – Distribution de la durée des segments (a) dans le corpus journalistique formel ESTER (Galliano *et al.*, 2009) et (b) dans le corpus spontané familier NCCFr (Torreira *et al.*, 2010). L'abscisse concernant la durée segmentale est donnée en milliseconde. L'ordonnée indique le pourcentage de cette durée dans chaque corpus.

La figure 1 illustre la répartition des durées de phones entre (a) la parole formelle journalistique et (b) la parole spontanée familière. La durée des phones sur l'abscisses est indiquée en millisecondes (ms). Les segments potentiellement réduits (à gauche) sont encadrés en gris. La distribution des durées de phones sur la parole formelle journalistique (Figure 1a) est présentée ici pour nous aider à mieux comprendre les caractéristiques de la distribution des durées de phones provenant de parole spontanée familière (Figure 1b). Comme nous pouvons le voir sur la figure 1a, la distribution de la durée des phones

correspond à une courbe globalement en forme de cloche et le sommet de la courbe se situe à 60 ms (voir l'abscisse). Cela suggère que la durée la plus fréquemment observée dans la parole formelle journalistique est de 60 ms. En ce qui concerne la parole spontanée (Figure 1b), le sommet de la courbe correspond à la durée minimale de 30 ms, ce qui suggère que la durée la plus fréquemment observée ici est de 30 ms (>20 %).

Ces observations sont conformes à celles d'Adda-Decker & Lamel (2017) sur la durée des phones en parole préparée et en parole spontanée (en français et en anglais). La parole spontanée familière traitée par l'alignement forcé, de la même manière que la parole formelle, contient alors une proportion extrêmement élevée de segments de durée minimale de 30 ms et de 40 ms. Dans cette étude, nous nous intéressons à la zone en gris de la parole spontanée familière de la Figure 1b.

3.2 Traitement des données

Nous pouvons constater sur la figure 1b que les segments courts (30 ou 40 ms) sont très fréquents dans notre corpus NCCFr de parole conversationnelle familière. Dans la suite, nous expliquerons comment nous avons localisé les suites de segments d'au moins trois segments consécutifs de 30 ou 40 ms.

L'alignement automatique fournit dans la forme de surface résultante (1) les segments d'une durée supérieure à 40 ms, (2) les segments avec moins de trois segments courts (30 ou 40 ms) consécutifs, ou (3) avec au moins trois segments courts (30 ou 40 ms) de suite. (1) et (2) sont considérés comme des segments « normaux » sans rien à signaler (nommés « Nrm ») ; (3) est considéré comme une zone qui est potentiellement impacté par la réduction (nommée « Alrt »). Le tableau 1 donne des exemples sur la façon dont les segments sont classés en fonction de leur durée.

Ex. /stʁ/ du mot « ministre » /ministʁ/	
– Si les segments [s], [t] et [ʁ] (qui se suivent) sont alignés chacun avec une durée courte (30 ou 40ms)	→ [s] segment en alerte : « Alrt » → [t] segment en alerte : « Alrt » → [ʁ] segment en alerte : « Alrt »
– Si les segments [s] et [t] sont alignés chacun avec une durée courte (30 ou 40ms) et le [ʁ] est aligné avec une durée de 50ms	→ [s] segment sans alerte : « Nrm » → [t] segment sans alerte : « Nrm » → [ʁ] segment sans alerte : « Nrm »

TABLE 1 – Exemple illustrant la catégorisation des segments comme « Nrm » ou « Alrt » à partir du mot « ministre ».

4 Résultats

Ci-après, nous étudions les suites de segments courts (zones de réduction potentielle) par rapport aux mots. Est-ce que la zone « Alrt » se trouve à l'intérieur d'un seul mot ou impacte-elle au moins deux mots ? Ensuite, nous regardons dans quelle position se trouvent les segments qui composent les zones « Alrt » à l'intérieur du mot. Ces résultats seront suivis d'une analyse de quelques mots et séquences de mots fréquents.

4.1 Position des segments réduits dans les mots

Nous obtenons un total de 23,002 zones « Alrt » impliquant un total de 86,265 segments. Pour information, le corpus contient un total d'environ 623,844 segments. Nous voulons savoir où se trouvent les séquences de segments étiquetés « Alrt » en opposant la position interne au mot à une localisation chevauchant deux ou plusieurs mots (figure 2). Ensuite, nous regardons pour les segments composant les séquences « Alrt », leur position à l'intérieur d'un mot (figure 3).

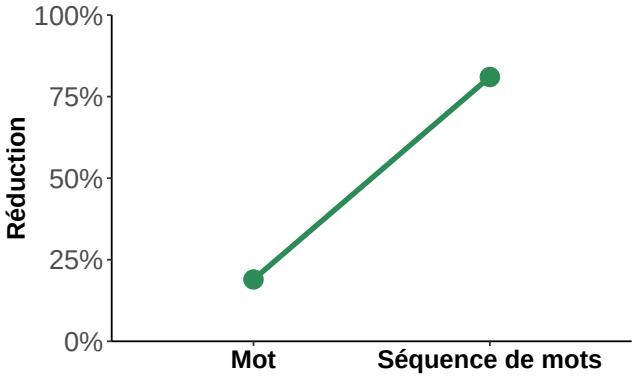


FIGURE 2 – Taux des suites de segments courts dans les mots et les séquences de mots.

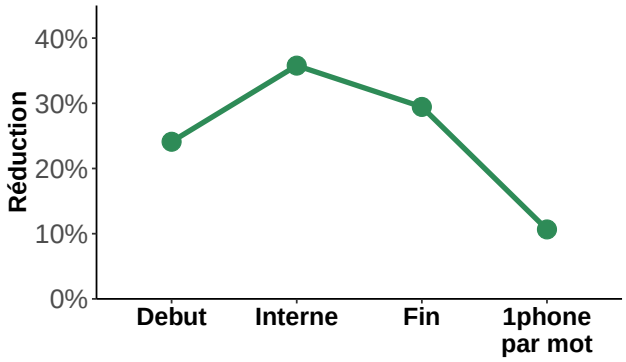


FIGURE 3 – Taux de segments courts en fonction de sa position dans le mot.

La figure 2 montre la proportion de suites de segments courts se réalisant complètement à l'intérieur d'un mot (Mot) [1] et celles impactant au moins deux mots (Séquence de mots) [2] observée dans les zones de réduction. En effet, parmi les « suites de segments courts », la plupart des segments courts font partie des suites de phones qui couvrent plus d'un mot (81%) et les suites de segments courts ne se limitent que rarement à un seul mot (19%). Ces résultats suggèrent que la zone de réduction concerne plus d'un mot en général.

La figure 3, montre la proportion de segments courts en fonction de leur position dans le mot. Nous observons plus de segments en position interne de mot qu'en position finale de mot (et davantage en position finale qu'en initiale) dans la zone de réduction. Cela suggère que la position interne du mot est la position préférée pour la réduction. La proportion la plus faible se trouve sur « 1phone par mot » (c'est-à-dire un segment qui est aussi un mot).

La figure 4 détaille les résultats de la figure 3 en fonction de la nature des suites de segments présentées dans la figure 2 (mot vs. séquence de mots). En ce qui concerne les suites de segments courts qui sont au sein d'un mot, la position interne de mot est la position privilégiée pour la réduction (80%). Il est intéressant de noter qu'aucune tendance particulière n'est observée en ce qui concerne la position dans un mot lorsqu'il s'agit de suites de segments courts au-delà de la frontière des mots (28% vs. 26% vs. 33% pour « Début », « Interne » et « Fin » respectivement) et que la position « 1 phone par mot » a une proportion légèrement inférieure par rapport aux trois autres positions.

1. Ex. suite de segments courts [ʁkw] provenant de la prononciation du mot « pourquoi » /puʁkwa/.
2. Ex. suite de segments courts [aʁʁa] provenant de la séquence de mots « par rapport » /paʁ#ʁapɔʁ/.

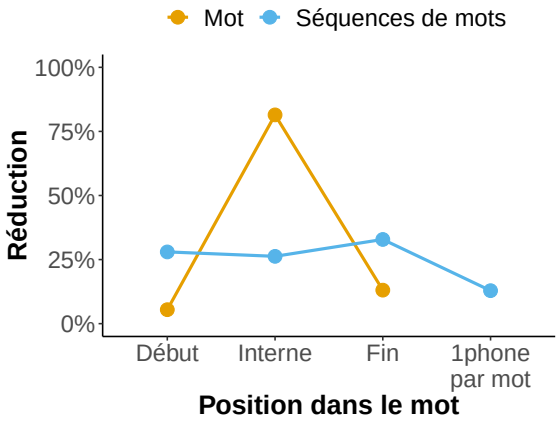


FIGURE 4 – Taux de segments courts suivant la position du segment dans le mot (Début vs. Interne vs. Fin de mot) et la nature des séquences (Mot vs. Séquence de mots).

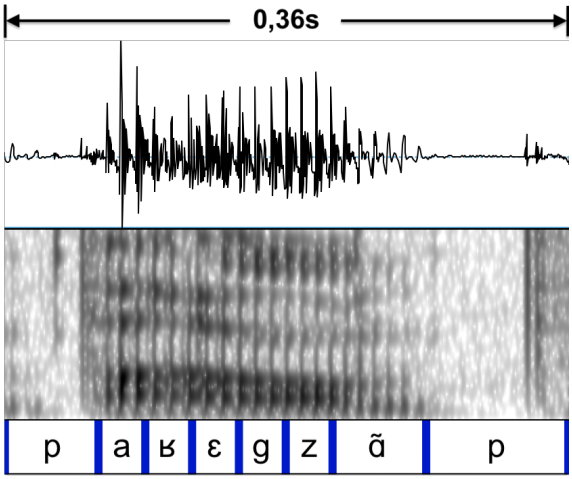


FIGURE 5 – Spectrogramme (0˜5000Hz) et signal de la locution *par exemple* /paʁ#ɛgzɑ̃pl/ (NCCFr, 26_11_07_nb1_2_16.wav, 1726,91s ˜1727,27s).

4.2 Mots et suites de mots fréquemment réduits

Dans la suite, nous présentons quelques mots et séquences de mots qui ont été fréquemment alignés à l'aide de suites de segments courts. Nous tentons d'identifier des phénomènes de réduction au-delà des segments. En effet, certains phénomènes de réduction peuvent se limiter à la simple chute d'un segment, comme la chute de la coda /l/ dans les mots « il », « ils », voire « elle », « elles ». Cependant, nous pensons que la réduction peut englober plusieurs segments (sous-jacents) consécutifs et résulter dans des formes de surface simplifiées, modifiées avec différents processus phonologiques à l'œuvre qu'il s'agit de mieux décrire dans le futur. Notre ambition ici se limite à essayer à mettre en évidence de tels phénomènes sur des exemples de mots et séquences de mots les plus représentatifs.

Nous présenterons tout d'abord les **mots** (c'est-à-dire sans considérer leur voisinage) qui ont été les plus fréquemment alignés avec des suites de segments courts (plus de 100 occurrences), avant de présenter des séquences de mots fréquemment alignées avec des suites de segments courts. Il faut noter qu'avec le critère appliqué (plus de 100 occurrences d'un même mot étiquetées « Alrt »), on ne peut tirer de conclusions que sur les mots fréquents.

Mot	**Alrt (occ.)**	**Total (occ.)**	$Taux\ (\%) = \frac{Alrt\ (occ.)}{Total\ (occ.)}$	**Détails** Seq. Occ.	Seq. Occ.	Seq. Occ.
voilà /vwala/	481	1723	28%	vwal 182 vwala 56	vwa 106 wala 24	wal 92 ala 21
crois /kʁwa/	129	486	27%	kʁwa 56	ʁwa 37	kʁw 36
trouve /tʁuv/	134	495	27%	tʁuv 84 uvə 2	ʁuv 32 tʁuvə 2	tʁu 18 ʁuvə 1

TABLE 2 – Mots alignés avec des suites de segments courts. « Alrt » et « Total » donnent respectivement le nombre d'occurrences du mot aligné avec des séquences de segments courts (Seq.) et le nombre total d'occurrences (Occ.). La colonne « Détails » précise les phones impliqués dans les « Alrt » alignés avec leur nombre d'occurrences.

Séquence de mots	Alrt (occ.)	Total (occ.)	$Taux\ (\%) = \frac{Alrt\ (occ.)}{Total\ (occ.)}$	Détails					
				Seq.	Occ.	Seq.	Occ.	Seq.	Occ.
par rapport /paʁ#ʁapɔʁ/	119	172	69%	aʁʁa	60	aʁʁ	46	paʁʁa	6
				ʁʁa	2	paʁʁ	2	ʁʁapɔʁ	1
par exemple /paʁ#ɛgzãpl/	117	239	49%	aʁɛgz	40	aʁɛg	24	ʁɛg	20
				ʁɛgz	8	aʁɛgzãp	6	paʁɛg	5
				aʁɛgzã	4	ʁɛgzãp	3	paʁɛgz	3
				aʁɛ	3	paʁɛ	1		
quand même /kã#mɛm/	177	765	23%	ãm	85	ãmɛm	63	kãm	18
				kãmɛm	4	kãmɛ	4	ãmɛmə	3

TABLE 3 – Séquences de mots alignées avec des suites de segments courts. « Alrt » et « Total » donnent respectivement le nombre d'occurrences de la séquence de mots alignée avec des suites de segments courts et le nombre total de ses occurrences. La colonne « Détails » donne les phones impliqués dans les « Alrt »avec leur nombre d'occurrences.

Le tableau 2 illustre des mots ayant plus de 100 occurrences alignés avec une suite d'au moins 3 segments courts (> 100 occurrences, au total 11 mots). Après vérification manuelle, nous avons pu constater que ces mots remplissent souvent un rôle de marqueur du discours. Ils ne sont pas très informatifs en soi, ils se comportent davantage comme une ponctuation (ex. voilà, je crois, je trouve, alors, tu vois, enfin, quoi) – que les locuteurs utilisent facilement dans un style de parole familier.

Le tableau 3 donne quelques séquences de mots fréquemment alignées avec une suite d'au moins 3 segments courts (> 100 occurrences, 6 séquences de mots). La même tendance a été observée que précédemment : il s'agit souvent de marqueurs discursifs et de « tics de langage ». Cependant, on y trouve également des expressions comme « par rapport » et « par exemple » qui ont les « taux d'alerte » les plus élevés parmi les séquences de mots trouvées. Il s'agit ici de deux expressions adverbiales qui restent parfaitement intelligibles même si elles sont prononcées d'une manière extrêmement réduite. On peut se demander comment sont réalisées de telles séquences de mots réduites. Quelles formes de surface sont produites ? Quels segments disparaissent ?

La figure 5 présente le spectrogramme avec la segmentation automatique de la locution « par exemple ». Une transcription phonétique manuelle donnerait à peu près [paʁãp]. Nous observons que le meilleur appariement entre signal et transcription a été obtenu grâce à la variante la plus courte sans le /l/ du cluster final /pl/, qui est la prononciation la plus adaptée selon le signal acoustique et le dictionnaire de prononciation. Ensuite, nous observons que la suite de phonèmes /ɛgzã/ a été alignée avec une suite de segments de durée minimale : il s'agit d'une réduction massive où la suite de segments [ɛgzã] a été prononcée comme un [ã].

La méthode ascendante mise en œuvre dans cette étude nous a permis de localiser des zones de parole à réduction massive. Les séquences de mots réduits ainsi mises en évidence révèlent qu'il s'agit avant tout de marqueurs discursifs. Mais on peut également y trouver des locutions polysyllabiques comme « par rapport » et « par exemple ». Ces locutions polysyllabiques se trouvent raccourcies en des réalisations de surface avec un nombre de syllabes plus faible. Dans des études futures, il serait intéressant d'analyser ces réductions en y intégrant une grille d'analyse prosodique dans la mesure où on peut avancer l'hypothèse que les syllabes non-accentuées seraient davantage sujettes au raccourcissement, voire à la disparition que les syllabes sous accent final.

5 Discussion

Cette étude sur la réduction temporelle montre que la combinaison de grands corpus avec l'alignement automatique présente des possibilités innovantes qui permettent de nouveaux points d'entrée d'analyses. Dans l'approche ascendante proposée, nous avons su profiter du « point faible » de l'alignement qui est la production d'une rafale de segments courts en zones temporellement réduites et nous avons pu apporter de nouvelles connaissances sur ce phénomène peu étudié auparavant. Dans la méthode ascendante appliquée dans cette étude, nous avons considéré comme « séquence réduite » les suites de segments d'au moins trois segments consécutifs de 30 ou 40 ms. Cette procédure très sélective nous a permis de localiser automatiquement les zones qui ont une forte probabilité d'être réduites (environ 14% du corpus). Nous avons constaté que les suites des segments courts impliquent souvent deux ou plusieurs mots (au lieu d'un seul mot). De plus, si la suite des segments courts fait partie d'un seul mot, elle est en général en position interne.

Nos résultats sur les mots et séquences des mots fréquemment réduits concernent surtout des locutions utilisées comme marqueurs discursifs qui ne portent que peu d'information sémantique et qui se comportent davantage comme une ponctuation ou qui peuvent éventuellement jouer un rôle dans la gestion des tours de parole lors de l'interaction.

En examinant les résultats sur la propension à la réduction des segments obtenus au niveau des mots, on peut remarquer que les séquences de segments réduits les plus fréquents partagent souvent des traits phonologiques (cf. « quand même » dans le tableau 3 pour lequel les séquences réduites les plus fréquentes sont toutes voisées, presque toutes nasales et les consonnes partagent le lieu d'articulation bilabiale) . Cette observation nous invite à étudier plus en détails le rôle des traits partagés dans une séquence de segments réduits sur la réduction en parole spontanée.

Il est intéressant cependant de noter que des locutions adverbiales polysyllabiques comme « par exemple » et « par rapport » apparaissent comme les plus réduites dans le corpus NCCFr. Ce résultat nous invite à rester vigilants sur la présence de réductions potentielles sur des mots ou locutions polysyllabiques moins fréquents. En effet, il faut garder à l'esprit qu'avec notre critère de sélection pour l'étude des mots et des séquences de mots (nous n'avons inclus que des séquences ayant plus de 100 occurrences étiquetées « Alrt »), on ne peut tirer de conclusions que sur les mots ou suites de mots les plus fréquents du corpus. Nous pensons que la réduction est certes favorisée par la fréquence des mots, mais qu'elle peut également être favorisée par d'autres régularités (ex. types de phonème impliqués, position syllabique, accentuation...) et des mécanismes (fusion, chute ou recombinaison de segments, de syllabes). Ce type de mécanismes peuvent potentiellement être à l'œuvre de manière similaire dans des phénomènes de réduction concernant des mots moins fréquents. Des études futures sur des plus grands corpus permettront de continuer ces pistes de recherche.

Remerciements

Ce travail est financé par Investissements d'Avenir – Projet Labex EFL (ANR-10-LABX-0083).

Références

Adda-Decker M. & Lamel L. (2000). The use of lexica in automatic speech recognition. In *Lexicon Development for Speech and Language Processing*, p. 235–266. Springer.

Adda-Decker M. & Lamel L. (2017). Discovering speech reductions across speaking styles and languages. In *Rethinking reduction : Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Walter de Gruyter GmbH & Co KG.

Duez D. (1997). Acoustic markers of political power. *Journal of Psycholinguistic Research*, **26**(6), 641–654.

Duez D. (2003). Modelling aspects of reduction and assimilation of consonant sequences in spontaneous french speech. In *Proceedings of Spontaneous Speech Processing and Recognition, IEEE-ISCA*, p. 120–124 : University of Tokyo.

Ernestus M. T. C. (2000). *Voice assimilation and segment reduction in casual Dutch : A corpus-based study of the phonology-phonetics interface*. Thèse de doctorat, LOT, Utrecht.

Galliano S., Gravier G. & Chaubard L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Gauvain J.-L., Lamel L. & Adda G. (2002). The limsi broadcast news transcription system. *Speech communication*, **37**(1), 89–108.

Johnson K. (2004). Massive reduction in conversational american english. In *Spontaneous speech : Data and analysis. Proceedings of the 1st session of the 10th international symposium*, p. 29–54 : Tokyo, Japan : The National International Institute for Japanese Language.

Meunier C. & Espesser R. (2011). Vowel reduction in conversational speech in french : The role of lexical factors. *Journal of Phonetics*, **39**(3), 271–278.

Torreira F., Adda-Decker M. & Ernestus M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**(3), 201–212.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Les variations du schwa transitionnel en tachlhit : une analyse acoustique

Minmin Yang[1] Rachid Ridouane[1]
(1) Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins 75005 Paris, France
minmin.yang@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

RESUME

Les caractéristiques temporelles et spectrales du schwa transitionnel en tachlhit sont analysées dans cette étude. Nous avons examiné 18 items du type C_1C_2VC afin d'explorer comment la durée et la qualité de ce vocoïde sont affectées par le contexte consonantique et vocalique avoisinant. Les résultats obtenus à partir de la réalisation de 7 locuteurs natifs montrent que la durée du schwa est beaucoup plus court comparées aux voyelles pleines. Alors que cette durée varie peu selon le contexte, la qualité du schwa peut être affectée par une combinaison de facteurs incluant la nature de la voyelle qui suit, ainsi que le lieu et le mode d'articulation des consonnes adjacentes. Ces variations sont observées pour F1, F2 et F3, et la plupart d'entre elles peuvent être prédites selon que la consonne qui suit est une occlusive emphatique ou une sonante battue.

ABSTRACT

Variations of transitional schwa in Tashlhiyt: an acoustic analysis

The temporal and spectral characteristics of the Tashlhiyt transitional schwa are analyzed in this study. We examined 18 C_1C_2VC type items aiming to explore how the duration and quality of this vowel are affected by the consonantal context and surrounding vowels. The results obtained from the realization of 7 native speakers show that the duration of the schwa is much shorter compared to full vowels. While this duration varies little according to context, the quality of the schwa can be affected by a combination of factors including the nature of the vowel that follows, and the place and mode of articulation of adjacent consonants. These variations are observed for F1, F2 and F3, and most of them can be predicted by whether the following consonant is an emphatic stop or a rhotic sonorant.

MOTS-CLES : schwa transitionnel, durée, qualité, séquences consonantiques, berbère tachlhit
KEYWORDS: transitional schwa, duration, quality, consonant sequences, Tashlhiyt Berber

1 Introduction

Plusieurs types de schwa sont attestés dans les langues du monde. Le schwa peut être (i) sous-jacent (i.e. donné par le lexique), (ii) dérivé de la réduction d'une voyelle pleine, (iii) épenthétique (i.e. inséré par la composante phonologique), ou (iv) transitionnel (i.e. résultant de la transition phonétique entre deux consonnes adjacentes). Dans cette étude, nous allons nous intéresser à ce dernier type, encore peu étudié en phonétique et en phonologie. Le tachlhit, la langue examinée,

Volume 1 : Journées d'Études sur la Parole, pages 636–644.
HAL : hal-02798597.

est connu pour l'extrême souplesse qu'il offre pour former des séquences consonantiques au niveau sous-jacent, et se présente donc comme un terrain fertile pour traiter de ce vocoïde. Nous allons plus spécifiquement nous intéresser à sa variabilité sur le plan acoustique, en examinant comment sa qualité et sa durée sont affectées par le contexte consonantique et vocalique avoisinant. L'inventaire vocalique du tachlhit contient 3 voyelles au niveau sous-jacent /i u a/. Au niveau de la surface, en plus des différentes réalisations de ces 3 voyelles pleines, un élément schwa, transcrit ici comme [ə], est fréquemment observé dans le signal acoustique. La question du statut de ce vocoïde a fait l'objet d'après discussions dans la littérature. Deux tendances ont longtemps animé ce débat : (i) la première, représentée par les travaux de Dell & Elmedlaoui (1985, 1988, 1996, 2002), Ridouane (2003, 2008), Ridouane et Fougeron (2011), considère schwa comme un élément transitionnel irrépressible, gouverné uniquement par la nature phonétique des consonnes adjacentes, (ii) la deuxième, représentée notamment par les travaux de Coleman (1996, 1999, 2001 ; mais aussi Louali et Puech, 2000), considère au contraire qu'il s'agit là d'une voyelle épenthétique insérée par la composante phonologique pour occuper le noyau de toute syllabe n'ayant pas de voyelle pleine. Plus récemment, des travaux sur la structuration prosodique du tachlhit et son interaction avec la distribution du schwa ont permis d'y voir plus clair (Gordon & Nafi, 2012 ; Grice et al. 2011, 2015 ; Roettger, 2016 ; Ridouane 2008 ; Ridouane & Cooper-Leavitt 2019). Les résultats de l'étude menée par Ridouane & Cooper-Leavitt (2019) montrent ainsi la coexistence de deux types de schwa dans la langue, qui font surface dans des conditions différentes et ont des relations différentes avec la structure phonologique : un vocoïde transitionnel (T-vocoïde) et un vocoïde déclenché prosodiquement (P-vocoïde). Le P-vocoïde ne fait surface que lorsqu'un marquage prosodique saillant ne s'attache pas à une voyelle lexicale ou à une sonante déjà présente dans le mot. Le T-vocoïde (schwa transitionnel), beaucoup plus fréquent, est ignoré par le système phonologique de la langue, et son apparition dépend des caractéristiques phonétiques des consonnes adjacentes.

Au-delà de la variabilité liée aux locuteurs et au débit, deux conditions sont nécessaires pour l'émergence d'un schwa transitionnel entre deux consonnes en tachlhit : (i) au moins une des deux consonnes doit être phonologiquement voisée, et (ii) le conduit vocal doit être suffisamment ouvert au moment de la transition de la première à la deuxième consonne. Ainsi, dans une forme de type C_1C_2VC (où V = voyelle pleine), seul un schwa transitionnel peut être acoustiquement présent entre C_1 et C_2, si les conditions (i) et (ii) sont satisfaites (e.g. dans [bədan] 'ils ont commencé'). Ce sont des items de ce type que nous allons examiner dans cette étude. L'objectif est de déterminer comment le lieu et le mode d'articulation de C_2 ainsi que la nature de la voyelle pleine affectent la durée et la qualité des schwas transitionnels. Les résultats obtenus seront notamment comparés aux résultats obtenus par Coleman (2001). Contrairement au travail de Coleman, dont les données étaient déséquilibrées et limitées à la réalisation d'un seul locuteur, cette étude porte sur plusieurs locuteurs produisant une liste de mots dont les contextes d'occurrence du schwa ont été soigneusement contrôlés.

2 Méthodologie

Sept locuteurs natifs du tachlhit (L1-7) ont été enregistrés dans le cadre de cette étude. L'enregistrement s'est déroulé dans la chambre sourde du Laboratoire de Phonétique et Phonologie (CNRS/Sorbonne Nouvelle, Paris). Les participants, enregistrés individuellement à l'aide du logiciel Protools, sont tous de sexe masculin et âgés entre 24 et 47 ans (âge moyen : 28 ans). Tous les locuteurs ont grandi au Maroc et parlent, en plus du tachlhit, l'arabe marocain, l'arabe standard et le français (comme c'est le cas pour la majorité des Amazighs ayant grandi au Maroc avant de venir en France). Le corpus utilisé est constitué de six verbes dans trois formes

grammaticales différentes, permettant d'obtenir des items ayant une forme C_1C_2VC (où V = /i/ pour le perfectif négatif, /a/ pour le perfectif affirmatif, et /u/ pour l'aoriste). La table 1 présente les 18 items utilisés dans cette étude (toutes ces formes sont à la 3e personne du masculin pluriel, d'où le suffixe /n/). Ces items nous ont permis d'effectuer trois types de comparaison selon la nature de V et de C_2 :

- Type 1 : Nature de V. L'objectif ici est de déterminer si la durée et la qualité de schwa varient selon que la voyelle qui suit est antérieure /i/, postérieure /u/ ou basse /a/ (e.g. [bədin] vs. [bədun] vs. [bədan]).
- Type 2 : Lieu d'articulation de C_2. L'objectif est de déterminer si la durée et la qualité de schwa varient selon que C_2 est une coronale simple /d/, une coronale emphatique /dˤ/ ou une dorsale /g/ (e.g. [bədan] vs. [bədˤan] vs. [bəgan]).
- Type 3 : Mode d'articulation de C_2. L'objectif est de déterminer si la durée et la qualité de schwa varient selon que la sonante C_2 est une nasale /n/, une latérale /l/ ou une battue /ɾ/ (e.g. [gənun] vs. [gəlun] vs. [gəɾun]).

		/i/	/a/	/u/	Glossaire
C_2 = Occlusive	a. /bdu/ b. /bgu/ c. /bdˤu/	[bədin] [bəgin] [bədˤin]	[bədan] [bəgan] [bədˤan]	[bədun] [bəgun] [bədˤun]	commencer percer partager
C_2 = Sonante	d. /gnu/ e. /glu/ f. /gɾu/	[gənin] [gəlin] [gəɾin]	[gənan] [gəlan] [gəɾan]	[gənun] [gəlun] [gəɾun]	coudre guide ramasser

TABLE 1 : Liste des items utilisés dans cette étude.

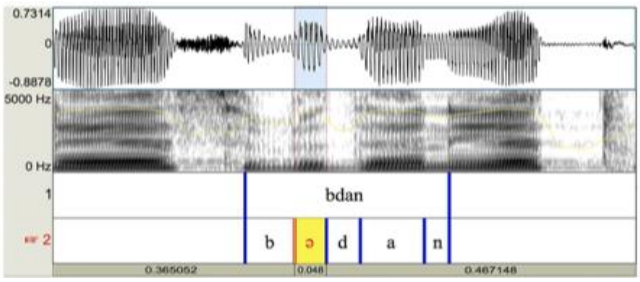


FIGURE 1 : Le signal acoustique et le spectrogramme d'une réalisation de la forme /bdan/ 'ils ont commencé' avec un schwa entre les consonnes [b] et [d] (i.e. [bədan]).

Une phrase cadre a été utilisée – *innajs … jatt twaltt* « Il lui a dit … une fois », et chaque phrase a été répétée par chaque locuteur 5 fois, donnant lieu à un corpus composé de 630 items (18 verbes * 7 locuteurs * 5 répétitions). Nous avons segmenté et annoté les données sur Praat (Boersma & Weenick, 2020), et nous avons extrait les durées et les valeurs F1, F2 et F3 pour chaque schwa transitionnel et chaque voyelle pleine (valeurs prises au début, au milieu et à la fin des voyelles, mais seules les valeurs au milieu sont présentées dans cette étude). La présence d'un schwa transitionnel dans une séquence de consonnes a été déterminée en se basant sur une combinaison d'indices. Ainsi, schwa correspond dans nos données à un intervalle suivant le relâchement de l'occlusive C_1, présentant des vibrations périodiques accompagnées d'une augmentation d'énergie et une structure formantinque dans la région F2/F3 caractéristique des voyelles. Un exemple est donné dans la figure 1. Comme attendu, et conformément à ce qui a déjà été observé pour des données semblables (Ridouane & Fougeron, 2011), le schwa transtionnel est très fréquent dans nos

données, et ce pour tous les locuteurs (619 occurrences sur 630) : L1 (98%), L2 (100%), L3 (98%), L4 (100%), L5 (98%), L6 (97%), L7 (98%).

3 Résultats

3.1 Effets sur la durée du schwa

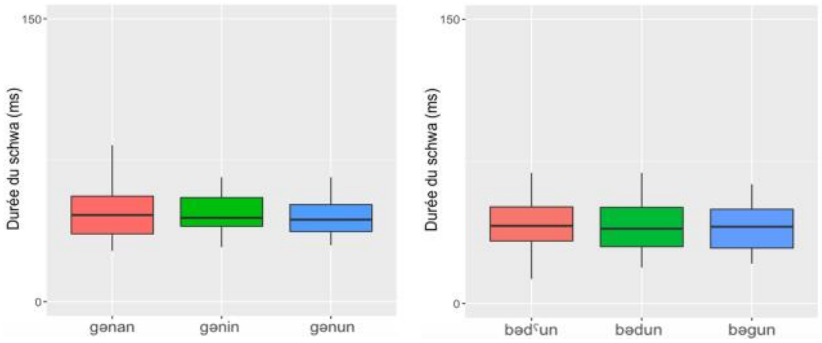


FIGURE 2ab : Boxplots illustrant l'absence d'effet de la nature de la voyelle (gauche) et du lieu d'articulation de la consonne C2 (droite) sur la durée du schwa.

Nous avons mesuré la durée du schwa et l'avons, dans un premier temps, comparé avec les durées des voyelles pleines /i a u/. Les résultats montrent que [ə] est beaucoup plus court, avec une durée de 50 ms en moyenne, soit au moins 2 fois plus court que les voyelles pleines (110 ms). La durée plus courte du schwa renvoie à une caractéristique de ce type de voyelles, partagée dans d'autres langues du monde. Pour autant, il est important de signaler que les voyelles pleines dans nos données sont particulièrement longues, très probablement en raison de la position finale de mot qu'elles occupent. Quand on compare la durée du schwa selon la nature de la voyelle pleine qui suit, les résultats montrent une absence d'effet (voir figure 2a pour le triplet [gənan], [gənin] et [gənun]). Cette absence d'effet est aussi observée quand on compare les durées du schwa selon le lieu d'articulation de C2 (voir figure 2b pour le triplet [bədun], [bədˤun] et [bəgun]).

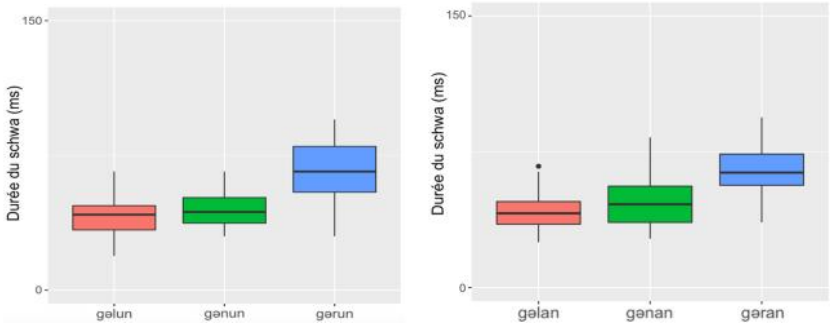


FIGURE 3 : Boxplots illustrant l'effet du mode d'articulation de la consonne C2 sur la durée du schwa.

En revanche, le mode d'articulation de C2 affecte significativement la durée du schwa, comme l'illustre la figure 3. Le test ANOVA indique que les durées du schwa dans le contexte de /l ɾ n/ sont significativement différentes ($F(2,101) = 15,5$; $p<,0001$) pour le triplet [gəlin], [gəɾin] et [gənin] ; ($F(2, 100)= 21, 7$; $p<,0001$) pour [gəlan], [gəɾan] et [gənan] ; et ($F(2,100)=36,2$; $p<,0001$) pour le triplet [gəlun], [gəɾun] et [gənun]). Le test post-hoc TukeyHSD montre que cet effet du mode d'articulation est dû à la différence entre le contexte /ɾ/ d'un côté et les contextes /l/

et /n/ de l'autre ([gəɾin] vs. [gəlin] (p<,0001) ; [gəɾin] vs. [gənin] (p<,0001) ; [gəɾun] vs. [gəlun] (p<,0001), [gəɾun] vs. [gənun] (p<,0001) ; [gəɾan] vs. [gəlan] (p<,0001) ; [gəɾan] vs. [gənan] (p<,0001)). Dans le contexte de /ɾ/, schwa affiche une durée moyenne de 65 ms, soit 22 ms plus longue que dans les autres contextes.

3.2 Effets sur la qualité du schwa

Avant d'examiner l'effet des trois paramètres sur les formants F1, F2 et F3, nous avons comparé les valeurs F1/F2 du schwa avec celles des voyelles pleines. Les résultats, présentés dans la figure 4, montrent que [ə] est une entité fortement instable, affichant une structure formantique pouvant englober une grande partie de l'espace vocalique, avec une tendance claire vers la centralisation (avec un F1 moyen de 339 Hz, et un F2 moyen de 1665 Hz).

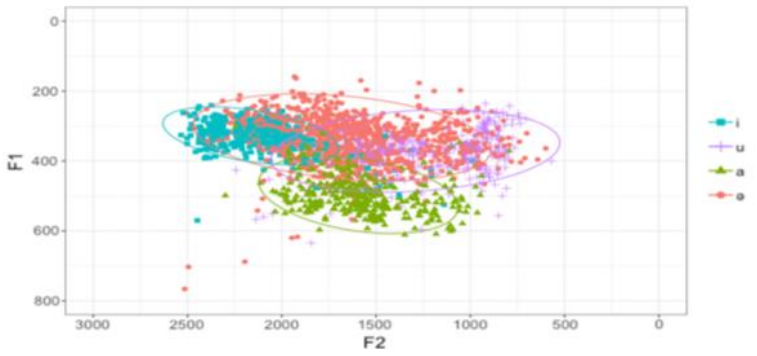


FIGURE 4 : Les valeurs F1 et F2 du schwa comparées à celles des voyelles pleines /i a u/.

3.2.1 Variations de F1

		F1(HZ)		F1(HZ)		F1(HZ)	
a	bədin	283	bədun	279	bədan	291	*NS*
b	bəgin	269	bəgun	310	bəgan	310	F(2,97)=5,6 ; p<,01
c	bədˁin	384	bədˁun	386	bədˁan	353	*NS*
d	gənin	300	gənun	301	gənan	323	*NS*
e	gəɾin	345	gəɾun	339	gəɾan	378	F(2,100)=11 ; p<,001
f	gəlin	334	gəlun	339	gəlan	306	*NS*

TABLE 2 : Valeurs moyennes de F1 selon la nature de la voyelle qui suit (NS = statistiquement non significatif). Ce tableau indique aussi les valeurs F1 selon le lieu et le mode d'articulation de C_2.

La comparaison des valeurs F1 du schwa selon la nature de la voyelle pleine qui suit montre des résultats différents selon les triplets. Alors qu'aucun effet n'a été observé pour la majorité des triplets (e.g. [bədin] vs. [bədun] vs. [bədan] avec [ə] ayant un F1 de 283 Hz, 279 Hz et 291 Hz, respectivement), des différences significatives ont été relevées pour le triplet [gəɾin] vs. [gəɾan] vs. [gəɾun]. Le test post-hoc TukeyHSD montre que c'est dû à la différence entre [gəɾan] d'un côté et [gəɾin] et [gəɾun] de l'autre ([gəɾan] vs. [gəɾin] (p<,001) et [gəɾan] vs. [gəɾun] (p<,0001)). Des différences significatives ont aussi été observées pour le triplet [bəgin], [bəgun] et [bəgan], mais cette fois-ci entre le contexte /i/ d'un côté, et /a u/ de l'autre. Même si ces résultats ne sont pas concordants, ils n'en demeurent pas moins dépendants du degré d'aperture de la voyelle pleine qui suit : schwa affiche un F1 plus élevé dans le contexte de la voyelle basse /a/ (pour les séquences [gəɾ]), et un F1 plus bas dans le contexte de la voyelle fermée /i/ (pour les séquences [bəg]). Le lieu d'articulation de la consonne C_2 affecte le F1 du schwa, et notamment en présence de

l'emphatique /dˤ/. La présence de cette emphatique constitue en effet l'effet le plus marquant et le plus systématique observé dans nos données. L'emphatique /dˤ/ induit systématiquement une élévation importante du F1. Ainsi la séquence [bədˤ] (toutes voyelles pleines confondues) affiche un F1 de 374 Hz, comparées aux séquences [bəd] et [bəg] qui affichent un F1 de 284 Hz et 296 Hz, respectivement (voir la table 2 qui présente les valeurs pour chaque item). Du point de vue statistique, toutes les comparaisons deux à deux entre les séquences [bədˤ] d'un côté et les autres séquences de l'autre sont significatives (p<,001). Concernant le mode d'articulation, le seul effet significatif observé a été entre [gəɾan] d'un côté et [gəlan] et [gənan] de l'autre (p<,0001), un autre effet visiblement dû à la présence de la battue /ɾ/.

3.2.2 Variations de F2

		F2(HZ)		F2(HZ)		F2(HZ)	
a	bədin	1761	bədun	1625	bədan	1654	F(2,102)=10,1 ; p<,001
b	bəgin	1892	bəgun	1503	bəgan	1721	F(2,98)=79,5 ; p<,001
c	bədˤin	1365	bədˤun	1350	bədˤan	1178	F(2,102)=3,8 ; p<,001
d	gənin	2014	gənun	1935	gənan	1800	F(2,104)=3,2 ; p<,05
e	gəɾin	1950	gəɾun	1677	gəɾan	1781	F(2,100)=11 ; p<,001
f	gəlin	2125	gəlun	1900	gəlan	2037	F(2,101)=3,8 ; p<,05

TABLE 3 : Valeurs moyennes de F2 selon la nature de la voyelle, ainsi que le lieu et le mode d'articulation de C2.

L'effet de la voyelle pleine sur le F2 est significatif pour tous les triplets, comme le montre la table 3. La différence la plus claire est celle observée entre le contexte /i/, qui a tendance à élever le F2 du schwa (1851 Hz, tous contextes consonantiques confondus), et le contexte /u/ qui a tendance à le baisser (1665 Hz) ; les valeurs de F2 dans le contexte /a/ sont intermédiaires (1695 Hz). Le lieu d'articulation de la consonne C2 affecte aussi la qualité du schwa. Là aussi, comme pour le F1, ce sont les séquences contenant l'emphatique /dˤ/ qui affichent les différences les plus notables et les plus systématiques. La présence de la consonne emphatique induit ainsi un abaissement très important du F2 : le schwa dans le contexte [bədˤ] présente un F2 qui est 382 Hz plus bas comparé au contexte [bəd] et 407 Hz plus bas comparé au contexte [bəg]. Ce sont là aussi, et de loin, les différences les plus importantes observées dans nos données. Concernant le mode d'articulation, les différences significatives observées sont entre [gəɾin] et [gəlin] (p<,05) ; [gəɾun] et [gəlun] (p<,001) ; [gəɾun] et [gənun] (p<,001) ; [gənan] et [gəlan] (p<,01) ;[gəɾan] et [gəlan] (p<,01). Comme pour le F1, ces différences sont principalement dues à la présence de la consonne battue /ɾ/.

3.2.3 Variations de F3

La nature de la voyelle pleine n'a pas d'effet significatif sur le F3 du schwa. Le F3 varie en effet très peu, que la voyelle qui suit soit un /i/ (2694 Hz), un /u/ (2704 Hz) ou un /a/ (2619 Hz). En revanche, le F3 du schwa peut varier selon le lieu d'articulation et le mode d'articulation de la consonne C2. Un examen plus détaillé de ces résultats montre là aussi un effet de la consonne emphatique (pour les items qui se distinguent par le lieu d'articulation), et un effet de la consonne battue (pour les items qui se distinguent par le mode d'articulation). Le test post-hoc TukeyHSD montre ainsi une différence significative entre [bədˤan] et [bəgan] (p<,05) ; entre [bədˤun] et [bədun] (p<,01). De même toutes les comparaisons entre les séquences [gəɾ] d'un côté et [gən] et [gəl] de l'autre, sont significatives ([gəɾin] vs. [gəlin] et [gənin] (p<,001) ; [gəɾun] vs. [gəlun] et [gənun] (p<,001) ; [gəɾan] vs. [gəlan] et [gənan] (p<,01)).

	F3(HZ)		F3(HZ)		F3(HZ)	
bədin	2698	bəgin	2679	bədˁin	2698	*NS*
bədan	2645	bəgan	2556	bədˁan	2714	$F(2,102)=2,9$; $p<,05$
bədun	2606	bəgun	2722	bədˁun	2849	$F(2,94)=5,9$; $p<,01$
gəlin	2773	gəɾin	2519	gənin	2799	$F(2,101)=10,7$; $p<,001$
gəlun	2756	gəɾun	2538	gənun	2750	$F(2,100)=7,4$; $p<,001$
gəlan	2653	gəɾan	2465	gənan	2683	$F(2,104)=6,9$; $p<,01$

Table 4 : Valeurs moyennes de F3 selon le lieu et le mode d'articulation de C_2, ainsi que la nature de la voyelle qui suit.

4 Discussion et conclusion

Dans cette étude, nous avons réalisé différentes analyses afin de répondre à la question suivante : comment les caractéristiques temporelles et spectrales du schwa transitionnel en tachlhit sont-elles affectées par la nature des voyelles et des consonnes avoisinantes ? L'étude est basée sur les productions de 7 locuteurs produisant des formes ayant la structure $C_1əC_2VC$, où C_2 varie selon le lieu et le mode d'articulation et où la voyelle pleine est antérieure /i/, postérieure /u/, ou basse /a/. Les résultats de l'analyse de la durée du schwa montrent que ce vocoïde est plus court que les voyelles pleines ; ces dernières étant environ deux fois plus longues. Ce résultat rejoint celui obtenu par Coleman (1999), et reflète, au-delà du cas spécifique du tachlhit, un aspect caractéristique de ces voyelles, dans différentes langues du monde, et ce quelle que soit la nature de ce schwa. Schwa est en effet intrinsèquement court qu'il soit lexical, issu d'une réduction vocalique ou épenthétique (voir Hall, 2006 ; Kalaldeh, 2008 ; Silverman, 2011 pour une revue). Pour autant, il est important de rappeler, comme nous l'avons signalé plus haut, que le schwa et les voyelles pleines ne sont pas dans le même contexte (les voyelles pleines sont en position finale de mot, une position susceptible d'allonger la durée des voyelles). Il sera utile pour la suite de ce travail de comparer la durée de schwa dans le même environnent que celui des voyelles pleines. Pour autant, la durée du schwa ne varie pas selon la nature de la voyelle pleine qui suit. Nous n'avons en effet trouvé aucune différence significative de durée du schwa, que la voyelle qui suit soit ouverte ou fermée. Des résultats similaires ont été observés pour le schwa lexical, notamment en néerlandais (Koopmans-van Beinum, 1994). L'absence d'effet a aussi observée quand on compare les durées du schwa selon le lieu d'articulation de C_2. En revanche, la comparaison selon le mode d'articulation présente des résultats différents. Nos données ont en effet monté que la durée du schwa était systématiquement plus longue lorsque celui-ci est suivi de la coronale battue /ɾ/. Dans ce contexte, schwa est environ 20 ms plus long que dans les autres contextes. Là aussi, cette caractéristique n'est pas propre au tachlhit (voir Koopmans-van Beinum (1994) pour le néerlandais).

En ce qui concerne la structure spectrale du schwa, les résultats de nos analyses montrent qu'elle peut être affectée par les trois paramètres examinés. Premièrement, concernant l'effet de la voyelle pleine, la majorité des différences significatives observées concernent le F2 du schwa, qui varie selon que la voyelle est antérieure ou postérieure. Fort logiquement, quand la voyelle pleine qui suit est antérieure (i.e. /i/), le schwa affiche un F2 plus élevé comparé au contexte où la voyelle pleine qui suit est postérieure (i.e. /u/). L'effet le plus systématique et le plus constant concerne celui que le lieu d'articulation exerce sur F1 et F2 du schwa, notamment en raison de la présence de l'emphatique /dˁ/. Cette consonne induit une élévation importante du F1 du schwa et un abaissement tout aussi important de son F2, ce qui rejoint les résultats obtenus par Coleman (2001). Cet effet, loin d'être limité au schwa en tachlhit, concerne toutes les voyelles dans le contexte

d'une consonne emphatique (Ridouane 2014). Appelé 'propagation de l'emphase', cet effet coarticulatoire a fait l'objet de plusieurs études notamment sur les variétés de l'arabe comme l'égyptien (Wahba, 1993), le libanais (Obrecht, 1968), le jordanien (Khattab et al., 2006), le tunisien (Ghazeli, 1981), et le marocain (Yeou, 2001 ; Zeroual et al., 2007 ; Lahrouchi et Ridouane 2016). Les voyelles qui jouxtent la consonne emphatique présentent un F1 plus élevé et un F2 plus bas que ceux des voyelles au contact des non-emphatiques, ce qui a comme conséquence, un rapprochement de ces deux formants. Ce rapprochement est principalement dû à l'articulation dorsopharyngale qu'implique ce type de consonnes, mais aussi à la "*simultaneous depression of the palatine dorsum*" (Ali & Daniloff, 1972 : 100). Les effets du mode d'articulation sur la structure spectrale du schwa se limitent presque tous aux effets de la battue /ɾ/. Cette consonne a pour effet d'augmenter F1 et d'abaisser F2. L'augmentation de F1 est probablement due au fait que cette sonante implique une ouverture du conduit vocal plus large comparée aux autres sonantes /l/ et /n/ (voir Coleman 2001 pour le même résultat). Des différences ont aussi été observées sur le plan F3. Là aussi, elles sont dues à un effet de la consonne emphatique qui a tendance à élever F3, ou un effet de la consonne battue qui a tendance à le baisser. Pour autant, ces différences semblent moins importantes que celles induites par ces mêmes consonnes sur les deux premiers formants.

Pour conclure, les données présentées ici montrent que la qualité du schwa transitionnel en tachlhit peut être affectée par une combinaison de facteurs incluant la nature de la voyelle qui suit, ainsi que le lieu et le mode d'articulation des consonnes adjacentes. La plupart de ces variations peuvent être prédites selon que la consonne qui suit est une emphatique ou une battue. D'autres données sont en cours d'analyse testant l'effet potentiel d'autres facteurs, en variant le mode et lieu d'articulation de la consonne initiale, et en incluant davantage de consonnes, comme les fricatives pas incluses dans cette étude.

Références

ALI L. & DANILOFF R. G. (1972). A contrastive cinefluorographic investigation of the articulation of emphatic-non emphatic cognate consonants. *Studia Linguistica*, *26*(2), 81-105. DOI : 10.1111/j.1467-9582.1972.tb00589.x.

BOERSMA P & WEENINK D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.09, http://www.praat.org.

COLEMAN J. (1996). Declarative syllabification in Tashlhit Berber. *Current Trends in Phonology: Models and Methods*, *1*, 175–216.

COLEMAN J. (1999). The nature of vocoids associated with syllabic consonants in Tashlhiyt Berber. In *Proceedings of the 14th International Congress of Phonetic Sciences* (Vol. 1, pp. 735-738).

COLEMAN J. (2001). The phonetics and phonology of Tashlhiyt Berber syllabic consonants. *Transactions of the Philological Society*, *99*(1), 29–64. DOI : 10.1111/1467-968X.00073.

DELL F. & ELMEDLAOUI M. (1985). Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics*, *7*(2), 105–130. DOI : 10.1515/jall.1985.7.2.105.

DELL F. & ELMEDLAOUI M. (1988). Syllabic consonants in Berber: Some new evidence. *Journal of African Languages and Linguistics*, *10*(1), 1–18. DOI : 10.1515/jall.1988.10.1.1.

DELL F. & ELMEDLAOUI M. (1996). On consonant releases in Imdlawn Tashlhiyt Berber. *Linguistics* 34. 357-395. DOI : 10.1515/ling.1996.34.2.357.

DELL F. & ELMEDLAOUI M. (2002). *Syllables in Tashlhiyt Berber And in Moroccan Arabic*. Dordrecht: Kluwer Academic Publications. DOI : 10.1017/S0025100304211860.

GHAZELI S. (1981). La coarticulation de l'emphase en arabe. *Arabica*, *28*(2), 251-277.

GRICE M., ROETTGER T. B., RIDOUANE R. & FOUGERON C. (2011). Tonal association in Tashlhiyt Berber. *Proceedings of the 17th International Congress of Phonetic Sciences*, *Hong Kong*. 775–778.

GRICE M., RIDOUANE R. & ROETTGER T. B. (2015). Tonal association in Tashlhiyt Berber: Evidence from polar questions and contrastive statements. *Phonology*, *32*(2), 241-266. DOI : 10.1017/S0952675715000147.

GORDON M. & NAFI L. (2012). Acoustic correlates of stress and pitch accent in Tashlhiyt Berber. *Journal of Phonetics*, *40*(5), 706-724. DOI : 10.1016/j.wocn.2012.04.003.

HALL N. (2006). Cross-linguistic patterns of vowel intrusion. *Phonology*, *23*(3), 387-429. DOI : 10.1017/S0952675706000996.

KALALDEH R. (2008). Hiberno-English Vowel System: Drogheda English.

KHATTAB G., AL-TAMIMI F. & HESELWOOD B. (2006). Acoustic and auditory differences in the/t/-/T/opposition in male and female speakers of Jordanian Arabic. In *Perspectives on Arabic Linguistics XVI: Papers from the sixteenth annual symposium on Arabic linguistics* (pp. 131-160). John Benjamins Cambridge, UK.

KOOPMANS-VAN BEINUM F. J. (1994). What's in a Schwa? *Phonetica*, *51*(1–3), 68–79. DOI : 10.1159/000261959.

LAHROUCHI M., & RIDOUANE R. (2016). On diminutives and plurals in Moroccan Arabic. *Morphology* 26 (2), 1-23. DOI : 10.1007/s11525-016-9290-7, HAL : hal-01324192.

LOUALI N. & PUECH G. (2000). Etude sur l'implémentation du schwa pour quatre locuteurs berbères de tachelhit. *Actes des 23e Journées d'Etudes sur la Parole. Aussois*, 25-28.

OBRECHT D. (1968). *Effects of the second formant on the perception of velarization consonants in Arabic*. The Hague: Mouton. DOI : 10.1515/9783111357393.

RIDOUANE R. (2003). Geminates vs. singleton stops in Berber: An acoustic, fiberscopic and photoglottographic study. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 1743-6.

RIDOUANE R. (2008). Syllables without vowels: Phonetic and phonological evidence from Tashlhiyt Berber. *Phonology*, *25*(2), 321–359. DOI : 10.1017/S0952675708001498.

RIDOUANE R. & FOUGERON C. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology*, *2*(2), 275-300. DOI : 10.1515/labphon.2011.010.

RIDOUANE R. (2014). Tashlhiyt Berber. *Journal of the International Phonetic Association*, 44(2), 207-221. DOI : 10.1017/S0025100313000388.

RIDOUANE R. & COOPER-LEAYITT J. (2019). A story of two schwas: a production study from Tashlhiyt. *Phonology*, *36*(3), 433-456. DOI : 10.1017/S0952675719000216.

ROETTGER T. B. (2016). *Stress and intonation in Tashlhiyt Berber*. PhD dissertation, University of Cologne.

SILVERMAN D. (2011). Schwa. *The Blackwell companion to phonology*, 1-15. DOI : 10.1002/9781444335262.wbctp0026.

WAHBA K. *(*1993*). A sociolinguistic treatment of the feature of emphasis in Egypt*. Unpublished doctoral dissertation, University of Texas at Austin.

YEOU M. (2001). Pharyngealization in Arabic: Modelling, acoustic analysis, airflow and perception. *Revue de La Faculté des Lettres El Jadida, 6, 51-70.*

ZEROUAL C., HOOLE P., FUCHS S. & ESLING H. (2007). EMA study of the coronal emphatic and non-emphatic plosive consonants of Moroccan Arabic. In *Proc. 16th ICPhS.*

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Effets du sexe et de la langue parlée sur la production de la parole chez les locuteurs coréens et français

Dayeon Yoon, Nicolas Audibert, Cécile Fougeron

Laboratoire de Phonétique et Phonologie (UMR7018, CNRS–Sorbonne Nouvelle, France)
{dayeon.yoon;nicolas.audibert;cecile.fougeron}@sorbonne-nouvelle.fr

RÉSUMÉ

Cette étude a pour but d'examiner l'effet du sexe et de la langue sur la production de la parole lue des locuteurs coréens et français. Dix paramètres acoustiques sont utilisés pour caractériser trois grandes dimensions : la voix (moyenne et écart-type de la F0, pente de LTAS et CPPs) ; les résonances du conduit vocal (F1 et F2 de /a/ et /i/) ; la gestion temporelle (débit de parole et articulatoire). Comme attendu, on observe une interaction entre sexe et langue sur la plupart des paramètres acoustiques supposés différencier les voix de femmes de celles d'hommes. Seuls le F1 de /i/ et la gestion temporelle ne montrent pas d'interaction entre sexe et langue. Ces résultats suggèrent que la différenciation de la voix entre sexes dépend de la langue parlée.

ABSTRACT

Effects of sex and language spoken on speech production among Korean and French speakers

The purpose of this study is to examine the effect of sex and language on read speech produced by Korean and French speakers. Ten acoustic parameters are used to characterize three dimensions: voice (mean and standard deviation of F0, spectral slope of the LTAS and CPPs); resonances of the vocal tract (F1 and F2 of /a/ and /i/); time management (speech and articulation rate). As expected, we observe an interaction between sex and language on most of the acoustic parameters supposed to differentiate the female from male voices. Only F1 of /i/ and the time management do not show any interaction between sex and language. These results suggest that the differentiation of voices between men and women depends on the language spoken.

MOTS-CLÉS : voix et sexe, voix et langue, interaction entre sexe et langue, Coréen, Français

KEYWORDS: voice and sex, voice and language, interaction between sex and language, Korean, French

Volume 1 : Journées d'Études sur la Parole, pages 645–653.
HAL : hal-02798598.

1 Introduction

Les études sur la voix sexuée font partie depuis longtemps de la recherche en phonétique. Les différences vocales entre hommes et femmes se basent tout d'abord sur des différences morphologiques. Une des caractéristiques les plus discriminantes entre les sexes est la fréquence fondamentale (ci-après F0) qu'on sait dépendre de la longueur et de l'épaisseur des plis vocaux, deux caractéristiques qui peuvent varier entre hommes et femmes. En général, la F0 des femmes est plus élevée que celle des hommes car les plis vocaux moins longs et moins épais des femmes permettent un nombre plus élevé de vibrations par unité de temps (Titze, 1989). Les plis vocaux jouent aussi un rôle principal dans la qualité de voix. Les voix de femmes sont souvent décrites comme relativement plus soufflées du fait d'une fermeture incomplète de la glotte par leurs plis vocaux moins longs et épais (Simpson, 2009).

Les résonances du conduit vocal sont un autre exemple de caractéristiques discriminantes entre homme et femme liées à des différences morphologiques. Les femmes présentent des fréquences de résonance plus élevées du fait d'un conduit vocal généralement moins long (en moyenne 14-14,5 cm) que celui des hommes (en moyenne 17 cm) (Fant, 1960).

Enfin, les différences dimorphiques influent également sur le débit de parole et articulatoire (Weirich & Simpson, 2013). Même si plusieurs recherches démontrent que les hommes parlent plus vite que les femmes (Byrd, 1994; Schwab & Avanzi, 2015), Weirich et Simpson (2013) suggèrent que les femmes seraient perçues comme plus rapides que les hommes. Par contre, pour les hommes, leur grande cavité buccale entraîne une vitesse de parole et articulatoire moins rapide par rapport à celle des femmes s'ils veulent atteindre les différentes cibles articulatoires.

La voix sexuée dépend non seulement du dimorphisme sexuel mais est aussi construite de façon culturelle. Ainsi, plusieurs études ont montré que les différences acoustiques entre voix d'hommes et de femmes pouvaient dépendre de la langue étudiée, et donc de l'origine des locuteurs. Par exemple, Van Bezooijen (1995) a observé que la F0 était plus élevée chez les locutrices japonaises que chez les locutrices néerlandaises et moins élevée chez les locuteurs japonais que chez les locuteurs néerlandais, ce provoque une différence de F0 entre sexes plus grande chez les Japonais que chez les Néerlandais. Van Bezooijen (1995) interprète ces résultats par une différence culturelle entre Japon et Pays-Bas, avec une voix idéale des femmes et des hommes qui n'est pas équivalente dans les deux pays. L'étude de locuteurs bilingues fournit d'autres exemples. Selon l'étude de Pépiot et Arnold (2018), la F0 de femmes bilingues français-anglais est plus élevée quand elles parlent français tandis que la modulation de leur F0 est plus forte quand elles parlent anglais. Au contraire, les hommes bilingues français-anglais, qui ne changent pas de F0 en fonction de la langue parlée, montrent une variation plus grande dans la modulation de F0 en français qu'en anglais. Benoist-Lucy et Pillot-Loiseau (2013) montrent que l'utilisation du « creaky voice », qui est un marqueur saillant chez les jeunes américaines, est moins fréquente lorsque celles-ci parlent une autre langue (ici le français). Les études susmentionnées suggèrent que les caractéristiques vocales qui dépendent du genre varient entre les langues.

Notre étude s'inscrit dans cette lignée de recherches, avec comme objectif de comparer la distinction entre hommes et femmes français et hommes et femmes coréens sur plusieurs paramètres acoustiques.

2 Méthode

Deux corpus de parole lue par des locuteurs et locutrices coréens et français ont été analysés comme indiqué dans la Table 1. Nous avons sélectionné une phrase par langue (pour le coréen, une phrase parmi d'autres dans une liste ; pour le français, une phrase isolée dans un protocole incluant différentes tâches) en comparant le nombre des syllabes, la présence des voyelles /a/ et /i/, et la structure prosodique (deux syntagmes accentuels (AP) avec un contour descendant à la fin de la phrase). Puis, en prenant en considération un possible effet de l'âge des locuteurs, nous avons divisé la population en trois groupes d'âge avec un équilibre entre hommes et femmes dans chaque groupe.

<table>
<tr><td></td><td colspan="3">CORÉEN</td><td colspan="3">FRANÇAIS</td></tr>
<tr><td>Corpus</td><td colspan="3">« Seoul Korean Speech Corpus »
(NIKL, 2005)</td><td colspan="3">« MonPaGe_HA »
(Fougeron et al., 2018)</td></tr>
<tr><td>Phrase</td><td colspan="3">«엄마가 해주신 밥이 왜 그렇게 맛있는지. »
(Comme la cuisine de maman est délicieuse.)
[ʌmmaga hɛdzuɕin pabi ǀ wɛ kɯrʌk^{h}e madinnɯndzi ǁ]</td><td colspan="3">« Anne-Marie et moi allons à la mer. »
[anmaʁi e mwa ǀ alɔ̃ (z) a la mɛʁ ǁ]</td></tr>
<tr><td>Nombre des locuteurs</td><td colspan="3">72</td><td colspan="3">70</td></tr>
<tr><td rowspan="5">Distribution de l'âge des locuteurs</td><td>Sexe / Âge</td><td>Femme</td><td>Homme</td><td>Sexe / Âge</td><td>Femme</td><td>Homme</td></tr>
<tr><td>19~29</td><td>9</td><td>9</td><td>19~29</td><td>7</td><td>7</td></tr>
<tr><td>30~49</td><td>9</td><td>9</td><td>30~49</td><td>13</td><td>13</td></tr>
<tr><td>50~68</td><td>18</td><td>18</td><td>50~68</td><td>15</td><td>15</td></tr>
<tr><td>Somme</td><td>36</td><td>36</td><td>Somme</td><td>35</td><td>35</td></tr>
</table>

TABLE 1 Récapitulatif des données analysées et de la population observée dans les deux langues

Nous avons choisi d'analyser avec ces données trois dimensions de parole : voix, résonances et gestion temporelle. Pour la dimension « voix », une F0 moyenne a été mesurée en hertz sur les segments voisés dans la phrase. La segmentation a été faite manuellement. Puis, l'écart-type de la F0 a été mesuré en demi-ton de façon à normaliser la modulation de F0. Puis, pour analyser la qualité de voix, la pente de LTAS et le CPPs ont été mesurés. Une pente de LTAS a été mesurée en décibel en calculant le rapport d'énergie entre 0-1 kHz et 1-5 kHz sur toute la phrase, puis les valeurs ont été exprimées en valeurs absolues. La pente de LTAS dépend de la qualité de voix : plus la pente est grande, plus la voix est perçue comme craquée ; moins la pente est grande, plus la voix est perçue comme soufflée (Mendoza et al., 1996). Pour le CPPs, il a été mesuré en décibel sur toute la phrase pour analyser la qualité générale de la voix comme « breathiness »,

« hoarseness » et « roughness », etc. (Hillenbrand & Houde, 1996 ; Maryn & Weenink, 2015). Plus grande est la valeur de l'amplitude (dB) du CPPs, moins dysphonique est la parole.

Pour la dimension « résonance », les formants F1 et F2 des voyelles /a/ et /i/ ont été mesurés au milieu des voyelles puis moyennés pour avoir une valeur par voyelle par phrase.

Pour la dimension temporelle, un débit de parole et un débit articulatoire (sans pause) ont été calculés par phrase en syl/sec.

Afin d'évaluer l'effet du sexe et de la langue sur les variables acoustiques susmentionnées, un modèle linéaire à effets mixtes a été appliqué avec le SEXE et la LANGUE des locuteurs comme facteurs fixes et avec le GROUPE D'ÂGE comme intercept aléatoire, de façon à prendre en compte des différences potentielles en fonction de l'âge des locuteurs (VD ~ sexe + langue + sexe:langue + (1 | groupe d'âge)). En raison d'erreurs de convergence du modèle complet dues à la variance trop faible, un modèle linéaire simple a été appliqué pour certaines variables : la moyenne et l'écart-type de la F0, le CPPs, les F1 et F2 de /a/ et le F2 de /i/ (VD ~ sexe + langue + sexe:langue). Enfin, les contrastes entre groupes ont été évalués en appliquant une correction de Tukey.

3 Résultats

VARIABLE	SEXE	LANGUE	SEXE:LANGUE
F0 moyenne (Hz)	$F(1,14) = 355.96$; $p < .001$ ***	$F(1,14) = .9$; $p < .345$	$F(3,14) = 120.8$; $p < .001$ ***
Écart-type de la F0 (dt)	$F(1,14) = 43.01$; $p < .001$ ***	$F(1,14) = 85.55$; $p < .001$ ***	$F(3,14) = 174.99$; $p < .001$ ***
Pente de LTAS (dB)	$\chi^2(5) = 1.51$; $p = .22$	$\chi^2(5) = .07$; $p = .795$	$\chi^2(6) = 10.44$; $p = 0.001$ **
CPPs (dB)	$F(1,14) = .18$; $p = .67$	$F(1,14) = 109.06$; $p < .001$ ***	$F(3,14) = 36.79$; $p < .001$ ***
F1 de /a/ (Hz)	$F(1,14) = 196.61$; $p < .001$ ***	$F(1,14) = 4.49$; $p < .036$ *	$F(3,14) = 75.96$; $p < .001$ ***
F2 de /a/ (Hz)	$F(1,14) = 176.11$; $p < .001$ ***	$F(1,14) = .24$; $p = .627$	$F(3,14) = 71.79$; $p < .001$ ***
F1 de /i/ (Hz)	$\chi^2(5) = 20.31$; $p < .001$ ***	$\chi^2(5) = 7.29$; $p = .007$ **	$\chi^2(6) = 1.5$; $p = .22$
F2 de /i/ (Hz)	$F(1,14) = 7.43$; $p = .007$ **	$F(1,14) = 2.35$; $p = .127$	$F(3,14) = 3.78$; $p = .012$ *
Débit de parole (syl/sec)	$\chi^2(5) = .14$; $p = .709$	$\chi^2(5) = 21.05$; $p < .001$ ***	$\chi^2(6) = .28$; $p = .594$
Débit articulatoire (syl/sec)	$\chi^2(5) = 2.04$; $p = .153$	$\chi^2(5) = 10.32$; $p = .001$ **	$\chi^2(6) = 2.29$; $p = .13$

TABLE 2 Récapitulatif des résultats statistiques

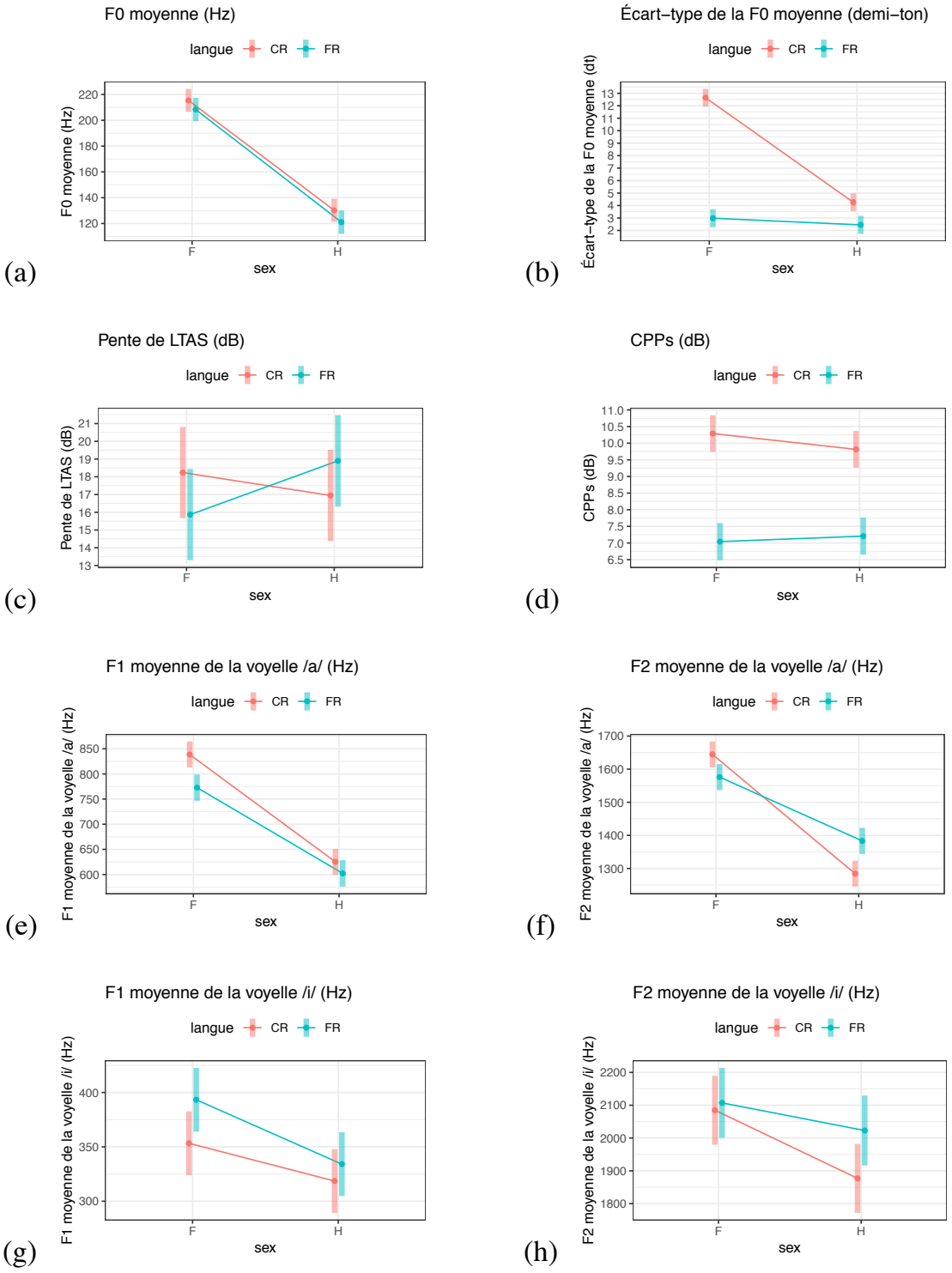


FIGURE 1 (a) F0 moyenne (Hz), (b) Écart-type de la F0 moyenne (demi-ton), (c) Pente de LTAS (dB), (d) CPPs (dB) (e) F1 de /a/ (Hz), (f) F2 de /a/ (Hz), (g) F1 de /i/ (Hz) et (h) F2 de /i/ (Hz) en fonction du sexe (F : femme, H : homme) et de la langue (CR : coréen, FR : français)

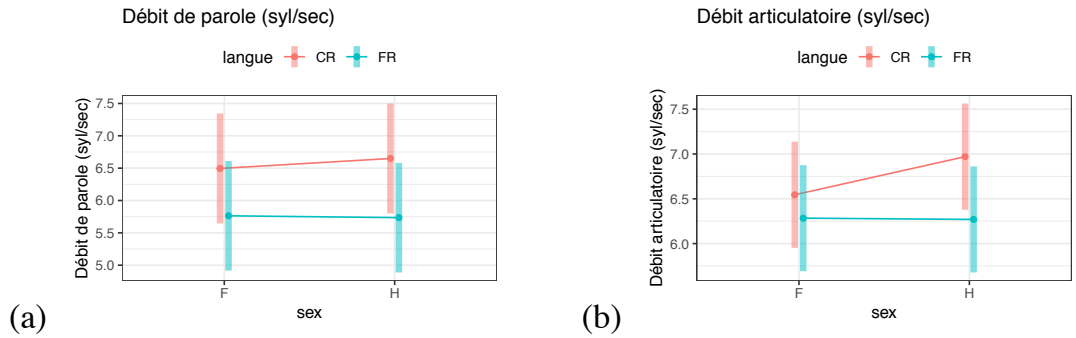


FIGURE 2 (a) Débit de parole et (b) Débit articulatoire (nombre des syllabes par seconde) en fonction du sexe (F : femme, H : homme) et de la langue (CR : coréen, FR : français)

Les résultats de l'évaluation statistique des paramètres sont présentés dans la Table 2. Comme le montre la Figure 1(a), une interaction entre sexe et langue s'observe avec la F0 moyenne plus élevée chez les femmes par rapport aux hommes, mais sans différence significative entre langues. Les femmes coréennes (215,4 Hz) et françaises (208,3 Hz) montrent respectivement la F0 moyenne plus élevée que les hommes coréens (130,2 Hz) et français (121 Hz). Quant à la modulation de F0, comme le montre la Figure 1(b), il y a une interaction entre les facteurs : un effet du sexe s'observe uniquement chez les Coréens, avec une modulation de F0 plus grande chez les femmes que chez les hommes alors qu'il n'y a pas de différence entre sexes chez les Français. D'autre part, la F0 apparait globalement plus modulée dans la phrase coréenne que dans la phrase française aussi en comparant les femmes que les hommes, mais l'écart entre langues est plus important en comparant les femmes (9,7 demi-tons) que les hommes (1,8 demi-tons). En outre, il y a une forte corrélation positive entre la F0 et sa modulation uniquement chez les femmes et les hommes coréens ($r = 0,85$ et $r = 0,9$, respectivement) et pas chez les femmes et les hommes français ($r = 0,15$ et $r = 0,03$, respectivement).

Pour la pente de LTAS, comme on peut le voir sur la Figure 1(c), il y a une interaction entre les facteurs de sexe et de langue, avec une différence entre sexes uniquement chez les Français et pas de différence entre langues. Donc, seuls les Français montrent une pente plus grande chez les hommes (val. abs., 19 dB en moyenne) par rapport aux femmes (val. abs., 16 dB en moyenne). En ce qui concerne le CPPs, on observe un effet de la langue et une interaction entre sexe et langue. Comme représenté sur la Figure 1(d), le CPPs varie en fonction de la langue avec un CPPs plus important sur les productions des femmes coréennes (10,3 dB en moyenne) que celles des femmes françaises (7 dB en moyenne) et sur les productions des hommes coréens (9,8 dB en moyenne) que celles des hommes français (7,2 dB en moyenne).

Quant aux formants, il y a une interaction entre sexe et langue pour les F1 et F2 de /a/. Comme on peut le voir sur la Figure 1(e), la différence globale des F1 entre sexes est évidente, mais cette différence varie en fonction de la langue : la différence entre sexes est moins grande entre homme et femmes français (différence de 170,5 Hz en moyenne) que celle entre homme et femmes coréens (différence de 213,5 Hz en moyenne). Les femmes ont également un F2 de /a/ plus haut que celui des hommes dans les deux langues comme dans la Figure 1(f), avec une interaction entre

sexe et langue. Comme c'était le cas pour le F1, la différence des F2 entre sexes est moins grande chez les Français (différence de 192,7 Hz en moyenne) que celle entre homme et femmes coréens (360,2 Hz en moyenne). Pour les formants de /i/, un effet global du sexe et celui de la langue s'observent pour le F1, mais cette fois-ci sans interaction entre les facteurs. Comme le montre la Figure 1(g), les F1 de /i/ sont plus élevés chez les femmes (370,9 Hz en moyenne) que chez les hommes (324,1 Hz en moyenne) indépendamment de la langue. Et, un effet de la langue se trouve aussi avec des F1 de /i/ globalement plus hauts en français (361,3 Hz en moyenne) qu'en coréen (334,1 Hz en moyenne). Quant au F2 illustré dans la Figure 1(h), une interaction entre sexe et langue se trouve avec une différence entre sexes uniquement chez les femmes et hommes coréens, avec les femmes dont le F2 est plus élevé par rapport aux hommes.

En ce qui concerne le débit de parole et articulatoire, illustré sur la Figure 2, aucune différence significative s'observe entre sexes, ni interaction entre sexe et langue. Par contre, le débit de parole et articulatoire varie en fonction de la langue, avec une vitesse de parole et articulatoire plus rapide en coréen (en moyenne 6,47 et 6,69 syl/sec, respectivement) qu'en français (en moyenne 5,67 et 6,22 syl/sec).

4 Discussion et conclusion

Dans cette étude nous avons cherché à mieux comprendre l'effet du sexe et de la langue sur la production de la parole lue par des locuteurs coréens et français. Nous avons examiné dix paramètres acoustiques utilisés pour caractériser trois grandes dimensions connues pour porter des différences entre femmes et hommes : la voix, les résonances du conduit vocal et la gestion temporelle de la parole. Nos résultats confirment les différences attendues entre hommes et femmes, mais surtout, montrent que la distinction entre les deux sexes dépend de la langue parlée : certaines différences apparaissent dans une langue et pas dans une autre, d'autres sont plus marquées dans l'une des deux langues.

Premièrement, la F0 moyenne et les F1 et F2 de /a/ varient en fonction du sexe dans les deux langues mais avec une différence plus saillante dans l'une que dans l'autre. Pour la F0 moyenne, nous retrouvons les différences attendues entre hommes et femmes (Vaissière, 2006) : les femmes ont en général une F0 plus élevée que les hommes dans les deux langues. Les deux premiers formants de la voyelle /a/ diffèrent également entre femmes et hommes et entre langues. Par contre, les F1 et F2 de la voyelle /i/ dépendent respectivement uniquement du sexe des locuteurs sans différences entre langues, ou uniquement de la langue mais pas du sexe du locuteur. Ces différences entre /a/ et /i/ suggèrent que les résonances de la voyelle /a/ seraient plus sensibles aux différences entre sexes. Néanmoins, les résultats que nous obtenons pourraient être influencés par les différences entre les systèmes vocaliques des deux langues ainsi que par la prosodie des phrases sélectionnées.

Deuxièmement, l'effet du sexe sur la modulation de F0 et la pente de LTAS n'apparaissent que dans une des deux langues. La modulation de F0 montre une différence entre sexes uniquement chez les Coréens, avec les femmes dont la modulation est plus forte que les hommes. Ces résultats

des Coréens sont contraires aux études précédentes qui affirment que, dans d'autres langues, il y a peu de différences entre sexes dans le même style de parole lorsqu'on mesure l'écart-type de la F0 moyenne en demi-tons (Traunmüller & Eriksson, 1993). De même, la pente de LTAS ne se distingue que chez les Français, en suggérant que la voix des hommes français est moins modale par rapport aux celle des femmes françaises. Les différences de contenu segmental entre langues pourraient toutefois avoir un impact sur la sensibilité du LTAS aux différences entre sexes.

Troisièmement, pour le CPPs et la gestion temporelle, l'effet du sexe est faible contrairement à l'effet de la langue. La valeur du CPPs plus élevée des Coréens que celle des Français suggère une voix plus périodique chez les Coréens que chez les Français, avec un effet modéré de l'interaction mais pas d'effet du sexe. Ce résultat est toutefois à nuancer en raison des différences de contenu segmental entre les phrases choisies dans chacune des deux langues, mais aussi des potentielles différences de conditions d'enregistrement au sujet desquelles le corpus coréen est peu documenté. Pour le débit de parole et le débit articulatoire, on observe uniquement un effet de la langue avec des débits plus rapides en coréen qu'en français, qui pourrait être en partie dû à un nombre différent de syllabes dans les phrases lues (16 syllabes en coréen contre 10 en français) : la longueur de l'énoncé peut en effet influer sur la durée des syllabes (Lehiste, 1974). L'absence d'effet du sexe dans nos données est en revanche en désaccord avec les études précédentes (Byrd, 1994 ; Schwab & Avanzi, 2015 ; Weirich & Simpson, 2013).

Enfin, cette étude a constaté que l'effet du sexe peut varier selon la langue parlée. Différentes hypothèses peuvent être avancées : tout d'abord, les différences acoustiques entre femmes et hommes coréens et français pourraient s'expliquer, en partie, par le caractère universel du dimorphisme entre sexes. Les différences morphologiques pourraient toutefois être plus ou moins marquées en fonction de différences ethniques. Il faudra donc vérifier s'il existe des données scientifiques indiquant de façon consistante que, par exemple, la différence de taille entre hommes et femmes coréens est plus importante que chez les Français. La taille seule ne permettant de prédire qu'une faible part des différences de voix (Pisanski et al., 2014), il sera nécessaire dans une étude ultérieure de collecter des informations morphologiques sur les locuteurs pour répondre plus directement à cette question (Weirich et al., 2016). Un autre aspect qui devra être approfondi est celui des caractéristiques attendues d'une voix d'homme ou de femme par des sociétés et cultures différentes (Johnson, 2006 ; Yuasa, 2010).

Remerciements

Ce travail est soutenu en partie par le Labex EFL (ANR-10-LABX-0083) et l'ANR VoxCrim (ANR-17-CE39-0016).

Références

BENOIST-LUCY A. & PILLOT-LOISEAU C. (2013). The influence of language and speech task upon creaky voice use among six young American women learning French. In F. BIMBOT, C. CERISARA, C. FOUGERON, G. GRAVIER, L. LAMEL, F. PELLEGRINO & P. PERRIER, Éds., *Speech in Life Sciences and Human Societies, 14th Annual Conference of the International Speech Communication*

Association, INTERSPEECH 2013, Lyon, France, August 25-29, 2013 Proceedings, volume 4, p. 2395-2399 : Curran. HAL : hal-00862349.

BYRD D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15, 39-54.

FANT G. (1960). *Acoustic theory of speech production*. Mouton.

FOUGERON C., DELVAUX V., MÉNARD L. & LAGANARO M. (2018). The MonPaGe_HA Database for the Documentation of Spoken French Throughout Adulthood. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *LREC 2018 Miyazaki, 11th International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018 Proceedings*, p. 4301-4306 : European Language Resources Association.

HILLENBRAND J. & HOUDE R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research*, 39(2), 311-321.

JOHNSON K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4), 485–499. DOI : 10.1016/j.wocn.2005.08.004.

LEHISTE I. (1974). Interaction between test word duration and length of utterance. In S. GARNES, I. LEHISTE, P. MILLER, L. SHOCKEY, & A. ZWICKY, Éds., *Ohio State Working Papers in Linguistics*, 17, p. 160–169. DOI : 10.1121/1.3437206.

MARYN Y. & WEENINK D. (2015). Objective dysphonia measures in the program praat: Smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice*, 29(1), 35-43.

MENDOZA E., VALENCIA N., MUÑOZ J. & TRUJILLO H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *J Voice*, 10(1), 59-66.

NATIONAL INSTITUTE OF KOREAN LANGUAGE. (2005). Seoul Korean Speech Corpus. https://ithub.korean.go.kr/user/total/referenceManager.do.

PÉPIOT E. & ARNOLD A. (2018). Étude des variations de fréquence fondamentale relatives au genre chez des bilingues Anglais/Français. In M. COOKE, B. BIGI & J. LAVAUD, Éds., *Actes des 32e journées d'études sur la parole Journées d'études sur la parole*, Aix-en-Provence. AFCP, LPL.

PISANSKI K., FRACCARO P., TIGUE C., O'CONNOR J., RÖDER S., ANDREWS P., FINK B., DEBRUINE L., JONES B. & FEINBERG D. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, 95, 89-99. DOI : 10.1016/j.anbehav.2014.06.011.

SCHWAB S. & AVANZI M. (2015). Regional variation and articulation rate in French. *J Phonetics*, 48, 96-105. DOI : 10.1016/j.wocn.2014.10.009.

SIMPSON A. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2), 621–640. DOI : 10.1111/j.1749-818x.2009.00125.x.

TITZE I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85(4), 1699-1707. DOI : 10.1121/1.397959.

TRAUNMÜLLER H. & ERIKSSON A. (1993). The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript. Stockholm.

VAISSIÈRE J. (2006). *La phonétique*. Paris : Presses Universitaires de France.

VAN BEZOOIJEN R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, 38(3), 253–265. DOI : 10.1177/002383099503800303.

WEIRICH M. & SIMPSON A. (2013). Acoustic vowel space size and perceived speech tempo. The Journal of the Acoustical Society of America, 133, 3571. DOI : 10.1121/1.4806540.

WEIRICH M., FUCHS S., SIMPSON A., WINKLER R. & PERRIER P. (2016). Mumbling: Macho or Morphology? *Journal of Speech Language and Hearing Research*, 59, S1587-S1595.

YUASA I. P. (2010). Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women? *American Speech*, 85(3), 315-337.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude des caractéristiques spatio-temporelles de la production de la parole chez des patients glossectomisés

Hasna Zaouali[1], Béatrice Vaxelaire[1], Christian Debry[2], Rudolph Sock[1&3]
[1]U.R. 1339-Linguistique, Langues et Parole (LiLPa) –ER Parole et Cognition
Institut de Phonétique de Strasbourg (IPS) – Université de Strasbourg
22 rue Descartes – 67084 Strasbourg– Cedex, France.
[2]Service O.R.L. - Hôpitaux Universitaires de Strasbourg
1 av. Molière – 67098 Strasbourg – Cedex, France.
[3]LICOLAB - Université Pavla Jozefa Safarika, Faculté des Lettres Košice – Slovaquie.
hasnazaouali@live.fr

RESUME

Cette étude porte sur les caractéristiques spatio-temporelles de la production de séquences VCV par des patients glossectomisés suite à un cancer endo-buccal. Plus précisément, il s'agit d'analyser les différents paramètres acoustiques (V1, VOT, VTT, silence, occlusion, V2), cette analyse nous permettra de rendre compte des conséquences d'une glossectomie sur le timing des gestes articulatoires (Sock, 1998). Dix patients ont été enregistrés sur plusieurs phases pré- et post-chirurgicales ; il s'agit donc d'une étude longitudinale. Le corpus étudié est composé de huit séquences de type VCV. L'objectif est principalement d'observer à partir du signal acoustique, différents événements acoustiques interprétables directement en termes articulatoires, en tentant ainsi de remonter aux configurations articulatoires. L'analyse statistique a montré des modifications significatives au niveau de la durée pour toutes les variables mesurées lors des phases d'enregistrements post-chirurgicales. En effet, une amélioration dans la production des séquences apparaît progressivement avec le temps et la réhabilitation orthophonique qui se manifeste pour certains patients par un retour aux valeurs initialement relevées lors de la phase pré-chirurgicale.

ABSTRACT

An acoustic study of spatio-temporal characteristics of speech production in glossectomised patients

This study examines the spatio-temporal characteristics of production of VCV sequences uttered by glossectomised patients following endo-oral cancer. More specifically, it involves analysing different acoustic parameters such as: (V1, VOT, VTT, the acoustic silent phase, occlusion, V2). This investigation should allow us to report the consequences that glossectomy may have on the the timing of articulatory gestures (Sock, 1998). Ten patients were recorded in several pre- and post-surgical phases; it is thus a longitudinal study. The corpus analysed consists of eight VCV sequences. The objective is mainly to observe in the acoustic signal various acoustic events directly interpretable in articulatory terms, and thereby attempting to infer articulatory configurations from the timing between specific acoustic events. Statistical analyses showed significant changes in the duration of all variables measured during post-surgical recording phases. An improvement in the

Volume 1 : Journées d'Études sur la Parole, pages 654–663.
HAL : hal-02798599.

production of the sequences appears gradually over time and speech therapy, as can be seen for some patients whose productions resemble initial values recorded during the pre-surgical phase.

MOTS-CLES : parole pathologique, glossectomie, analyse acoustique, perturbations, timing, VOT, VTT, réajustements.
KEYWORDS: speech pathology, glossectomy, acoustic analysis, perturbations, timing, VOT, VTT, readjustments.

1 Introduction

Nous proposons dans ce travail d'étudier les caractéristiques spatiotemporel de la production de la parole de patients glossectomisés suite à un cancer endo-buccal, et ce en fonction du site, de la taille de la résection et des traitements subis par chaque patient. Il est question d'estimer les effets d'une glossectomie partielle ou d'une pelvi-glossectomie[1] sur le *timing* des articulateurs, à partir d'un corpus composé de séquences de type VCV.

Notre démarche se veut articulatori-acoustique ; pour cela nous analysons, les différents indices articulatoires, à partir d'un signal acoustique continu, essayant ainsi de remonter aux gestes et aux configurations articulatoires (Abry *et al.,* 1985). Nous nous focalisons, à travers cette expérience, sur les perturbations impactant le niveau temporel suite à une exérèse carcinologique localisée dans le système de production de la parole.

Notre étude se veut longitudinale du moment où la parole des patients est enregistrée lors de différentes phases pré et post-chirurgicales (Préop, Post-Op1, 2 et 3), ce qui devrait nous permettre d'étudier les perturbations et les réajustements que les patients pourraient déployer après la glossectomie. Notons que selon la taille de l'exérèse, le site de la tumeur et le type de reconstruction, les patients glossectomisés trouvent des difficultés à réaliser certains sons, et en particulier les occlusives linguales [t, d, k, g] (Savariaux *et al*., 2000 ; Acher *et al*., 2014 ; Zaouali *et al*., 2018). En effet, après une résection carcinologique de la langue, l'occlusion est partielle, et certains patients ont tendance à réaliser une occlusive ayant des propriétés de constrictive, engendrant une extension de la durée des paramètres inter et intra-segmentaux. Pour certains patients, l'atteinte de la cible articulatoire n'est pas toujours évidente. Les patients glossectomisés mettre en œuvre des stratégies de compensation ou de réajustement, suite à la nouvelle configuration de leur cavité buccale provoquée par la résection et les prises en charges consécutives à la chirurgie. Ces stratégies compensatoires restent individuelles et se manifestent au niveau des perturbations des valeurs de la durée des segments vocaliques et consonantiques (Vaxelaire, 2007 ; Zaouali *et al*., 2018). Une amélioration au niveau de la production de la parole est donc discernable au cours du temps et de la réhabilitation orthophonique.

Par rapport aux travaux déjà effectués dans ce domaine, l'originalité de notre étude réside dans le fait que, outre l'examen du VOT, nous analysons de près un autre timing intra-segmental, le VTT (Voice Termination Time ou délai d'arrêt du voisement), et le timing inter-segmental de V1, de C et de V2 des séquences VCV.

[1] Pelvi-glossectomie : ablation d'une partie de la langue et du plancher buccal

2. Procédure expérimentale

2.1. Participants

Cette étude est réalisée à partir d'une cohorte de dix patients : trois femmes (ZIM, PETR et HACH), et sept hommes (SIB, SOM, GLAD, JCT, BIRL, ANT et ROJ) dont la production de la parole a été enregistrée dans différents établissements hospitaliers localisés en Alsace. Les enregistrements ont été réalisés sur quatre phases pré- et post-chirurgicales : Préop (la veille de l'intervention), Post-Op1 (entre 1 et 1,5 mois après l'intervention), Post-Op2 (3 mois après l'intervention), Post-Op3 (6 mois après l'intervention). Ces patients ont subi différentes ablations linguales, partielles ou sub-totales, suivies ou non de reconstructions (*cf*. Tableau1). Signalons que le patient JCT a été exclu de notre analyse. En effet, suite à la lourdeur de la résection subie par ce patient, certains paramètres n'étaient pas mesurables ; une étude spectacle a été réalisée pour ce cas clinique (*cf*. Zaouali *et al*., 2018). Nous nous sommes concentrés dans cette étude sur les données de six patients ayant accompli les 4 sessions d'enregistrements pré- et post-chirurgicales. Notons que la radiothérapie a eu lieu dans les deux, voire trois mois qui suivaient l'exérèse. Selon le site et le type de résections, le début de la thérapie était variable d'un patient à un autre, et les patients n'avaient pas le même nombre de séances de traitement, ni le même dosage.

Tableau 1 *:* Répartition des exérèses et informations complémentaires concernant les cas cliniques étudiés (TNM) : classification de la taille de la tumeur (T), de la présence d'adénopathies (N) et de la présence de métastases (M). M : homme, F : femme, (hémiglossect : (hémi-) glossectomie, mandibulect : mandibulectomie, G : gauche, D : droite).

Identification patients	Âge	Sexe	Profession	TNM	Type d'exérèse	Reconstruction	Traitements complémentaires	Rééducation ortho
SIB	42	M	Manager	T2N0M0	Glossect-Partielle G	Suture	Radiotherapie	Oui en libéral
SOM	30	M	Ingénieur	T1N0M0	Glossect-Partielle D	Suture	Curiethérapie	Non
ZIM	69	F	Infirmière	pT4aN2bM0	Pelvi-Glossect D	Suture	Radiotherapie	Non
GLAD	53	M	Conducteur	T2N0M0	Glossect-Partielle D	Suture	Radiotherapie	Oui en libéral
PETR	65	F	Retraité	T4N0M0	Pelvi-Glossect D	Suture	Radiotherapie	Oui au service ORL
JCT	53	M	Manager	T4N0M0	Pelvi-Glossect -Totale	Lambeau libre Antérolatéral	Radiotherapie	Oui en libéral
HACH	24	F	Etudiante	T4N2bM0	Hémi-Glossectomie D	Lambeau libre Anté-brachial	Radiotherapie	Oui en libéral
BIRL	47	M	Fonctionnaire	T3N0M0	Pelvi-Glosso-mandib G	Lambeau libre du péroné	Radiotherapie	Oui en libéral
ANT	68	M	Sans	T4N2bM0	Oro-pharyngectomie D	Lambau peaucien	Radiotherapie	Non
ROJ	60	M	Fonctionnaire	T1N0M0	Glossect-Partielle D	Suture	Radiotherapie	Non

2. 2 Corpus

Le corpus utilisé dans cette étude est composé de huit logatomes. Les logatomes ont été construits comme suit : si V1 = [i] alors V2 = [a] et *vice versa*. La consonne est l'une des 4 occlusives [t, k, d, g]. Ces consonnes ont été choisies car elles offrent la possibilité d'observer l'effet de la chirurgie et des traitements postopératoires sur le recul du lieu d'articulation et particulièrement de la masse de la langue, de l'avant vers l'arrière de la cavité buccale.
Les huit logatomes sont donc les suivants : [ati], [aki], [adi], [agi], [iga], [ida], [ika] et [ita].
Le corpus a été prononcé aléatoirement entre 5 et 10 fois, selon les possibilités de chaque patient, tout en respectant son degré de fatigabilité. Nous nous sommes limités, au bout du compte, à 5 répétitions qui était le nombre de répétions minimum atteint par l'ensemble des patients.

2.3 Mesures

Pour chaque logatome, nous avons mesuré les durées : 1) de la première voyelle (V1) ; 2) de l'occlusion pour les consonnes voisées et du silence acoustique pour les consonnes non-voisées ; 3) de la seconde voyelle (V2). En outre, nous avons mesuré : 4) le VTT (*Voice Termination Time*), ou délai d'arrêt du voisement qui correspond, pour les occlusives non voisées seulement, à l'intervalle entre la disparition de la structure formantique clairement définie de V1 à la fin d'oscillations périodiques dans la phase silencieuse de la consonne C, soit la période de transition menant à la fermeture complète du conduit vocal. Il est à noter qu'Agnello (1975) a été le premier à utiliser cette mesure ; 5) le Voice Onset Time (VOT), ou délai d'établissement du voisement, qui correspond à l'intervalle allant de l'explosion-friction (*burst*), due au relâchement consonantique, à l'apparition d'une structure formantique clairement définie de V2, reflétant un conduit vocal dégagé (Klatt, 1975). Notons que la durée absolue de la voyelle a été mesurée entre le VVO et le VVT, c'est-à-dire entre le début et la fin de la structure formantique clairement définie (*cf.* Figure 1 *infra*). Nous avons utilisé le model linéaire (*lm*) sous R pour effectuer nos analyses statistiques. Une analyse de variances ANOVA à mesures répétées a été réalisée.

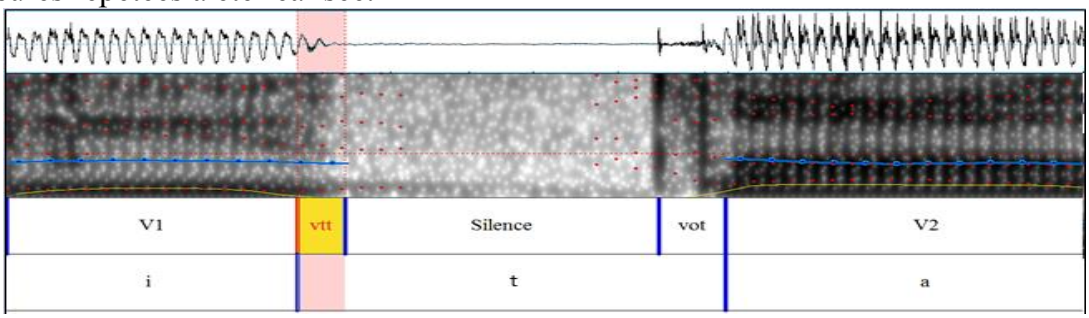


Figure 1 : Mesures temporelles pour une séquence VCV, où C correspond à une consonne non-voisée. Exemple de [ita]

3. Hypothèses

Suite à la glossectomie, de possibles modifications pourraient s'engendrer au niveau de la configuration de la cavité buccale du patient et du timing des articulateurs. En conséquence, nous avons émis les hypothèses suivantes :

Nous pensons pouvoir observer des altérations qui pourraient être perçues au niveau des occlusives [t, d, k, g] et qui se révéleraient dans les perturbations du timing au niveau inter- et intra-segmental : les durées vocaliques et consonantiques et notamment les timings du VTT (Voice Termination Time, ou délai d'arrêt du voisement) et du VOT (Voice Onset Time, ou délai d'établissement du voisement), indices importants dans la réalisation de l'opposition de sonorité chez les occlusives du français. Ainsi, l'opposition, au niveau des caractéristiques acoustiques temporelles, entre consonne voisée et consonne non-voisée devrait être également perturbée.

Relativement aux difficultés du contrôle de l'occlusion des occlusives dentales et vélaires, et au relâchement subséquent, le bruit de l'explosion provoqué par le relâchement de l'occlusive pourrait affecter la voyelle suivante, retardant l'apparition d'une structure formantique clairement définie et augmentant *de facto* la durée du VOT.

En raison des difficultés que présentent certains patients glossectomisés à obtenir une occlusion suffisamment étanche, les occlusives linguales étudiées [t, d, k, g] pourraient présenter des propriétés de fricatives. Savariaux *et al*. (2008) ont mis en évidence un bruit important lors de l'articulation, en français, des occlusives [t, d, k, g], qui pourrait être considéré comme un phénomène de compensation.

Suite à la diversité des stratégies compensatoire déployées selon les patients, une variabilité inter et intra individuel serait notable lors des phases post-chirurgicales. Le temps et la rééducation devraient permettre une amélioration dans la production de la parole chez nos sujets patients ; qui devrait progressivement apparaître en (Post-Op2 et 3).

4. Résultats

Les analyses statistiques de variances (ANOVA à mesures répétées) multifactorielles avec correction Tukey ont été réalisées pour toutes les variables mesurées, à savoir : (V1, le VTT, le silence acoustique, l'occlusion, le VOT, V2). Nous avons essayé de savoir s'il existerait des effets significatifs des facteurs principaux suivants : *temps*, *chirurgie*, *contexte vocalique* et *sonorité*.

Dans cette investigation, nous avons relevé des effets significatifs pour la plupart des variables mesurées, certaines interactions indiquaient également des effets significatifs ($p<0,05$). Pour illustrer nos résultats nous avons choisi de présenter les paramètres VTT et VOT des occlusives [t] et [k] dans deux contexte différents [i-a] et [a-i].

En ce qui concerne la voyelle précédente V1 dans les séquences VCV étudiées, les résultats de l'analyse statistique montrent une différence significative pour les 3 effets principaux « temps » (phase d'enregistrement), « chirurgie » et « sonorité » : pour le facteur « temps » les différences significatives résident entre Préop et Post-Op1 : [$F(3)=-12,64$; $p<0,0004$] ; Préop et Post-Op2 : [$F(3)=-9,458$; $p<0,0158$] ; pour le facteur « chirurgie » [$F(1)=12,675$; $p<0,000004$] ; [$F(1)=-15,446$; $p<0,0012$] ; et pour le facteur « sonorité » : [$F(1)=-15,446$; $p<0,0012$].

Pour la variable VTT, les 2 effets principaux qui se sont révélés statistiquement significatifs sont : « temps » et « chirurgie » : pour le facteur « temps », nous avons relevé des différences significatives entre Préop et Post-Op1 [$F(3)=-4,855$; $p<0,03761$] ; Préop et Post-Op2 : [$F(3)=-6,10556$; $p<0,00490$] ; pour le facteur « chirurgie » les différences sont de : [$F(1) = 5,981$, $p<0,00001$]. Le « contexte vocalique » s'est également révélé significatif pour la variable VTT. En ce qui concerne l'interaction du « contexte vocalique » [i-a] et du facteur « temps », elle montre que le VTT est significativement plus important entre (Préop et Post-Op1) : [$F(5)=-8,422$; $p<0,02507$] et (Préop et Post-Op2) : [$F(5)=-9,25556$; $p<0,00903$] en contexte [i-a]. Une autre interaction s'est révélée significative pour la variable VTT entre le facteur « temps » et « chirurgie », c'est à dire les deux types d'exérèses (A[2] et B[3]), lors de la phase Post-Op1 : [$F(3)=4,90694$; $p<0,03487$]. L'interaction des facteurs « chirurgie » et « contexte vocalique » indique à son tour une différence significative pour la variable VTT dans les deux contextes vocaliques [i-a] : [$F(3)=7,05556$; $p<0,00082$] et [a-i] : [$F(3) = 4,90694$, $p<0,03487$]. Ces interactions démontrent que le VTT est plus notable en Post-Op1, et cela dans les deux contextes vocaliques étudiés pour les patients pelvi-glossectomisés, par

[2] Patients partiellement glossectomisés

[3] Patients pelvi-glossectomisés

rapport aux patients ayant subi une glossectomie partielle. Cela nous indique que les patients pelvi-glossectomisés ont plus de difficultés à réaliser une occlusion linguo-palatale. Cette perte dans le contrôle de l'activité supraglottique est remarquée dans les deux contextes [i-a] et [a-i]. Cette difficulté de contrôle est retrouvée non seulement dans le contexte de la voyelle de petite aperture [i], mais elle est encore plus notable dans celui de la voyelle de grande aperture [a] (*cf*. Figure 2). L'interaction entre les facteurs « chirurgie » (type d'exérèse) et « sonorité » ne montre pas de différences significatives pour la variable VTT entre les patients pelvi-glossectomisés et leurs homologues partiellement glossectomisés, en contexte consonantique non-voisée.

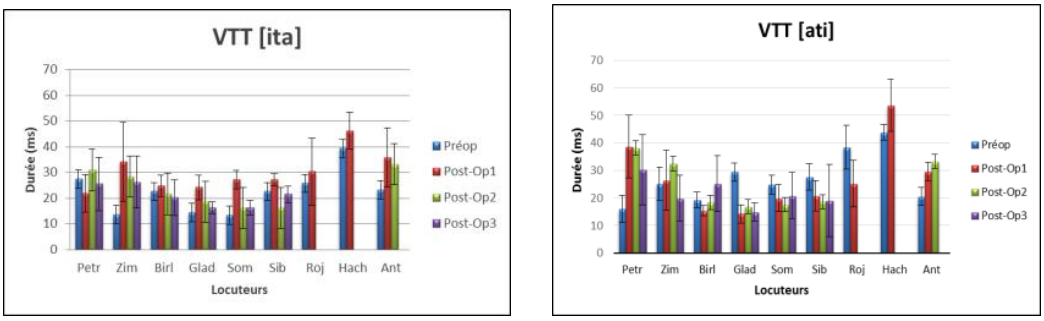


Figure 2:Valeurs de la durée du VTT en contexte [t] pour les logatomes [ita] (à gauche) et [ati] (à droite) en fonction du patient

En ce qui concerne la variable silence acoustique (SIL), les facteurs « temps » et « chirurgie » se sont révélés significatifs : « temps » entre Pérop et Post-Op1 [F (3) =-13,344 ; p< 0,0086] ; chirurgie [F (1) =-18,131 ; p< 0,0001]. Le silence acoustique pour les occlusives non-voisées est visiblement plus important en Post-Op1 après une pelvi-glossectomie qu'après une glossectomie partielle. Notons que pour les productions de certains patients, le silence acoustique n'était pas mesurable (ex patient JCT). Ce dernier était caractérisé par des vibrations presque régulières en remplaçant ainsi le silence acoustique qui est considéré, en théorie, comme l'une des caractéristiques d'une occlusive non-voisée. Cette remarque est à prendre avec précaution puisqu'elle ne concerne que certains patients et, parfois, que quelques items dans les productions du même patient.

Nous avons également observé une altération de la variable durée de l'occlusion, les effets principaux « temps », « contexte vocalique » et « chirurgie » se sont révélés significatifs : pour le facteur « temps » une différence significative est indiquée entre (Préop-Post-Op1) : [F (3) =-13,962 ; p< 0,00002)] ; (Préop-Post-Op2) : [F (3) =-10,719 ; p< 0,0021)] ; (Post-Op1-Post-Op3) : [F (3) =-8,0513 ; p< 0,0356)] ; (Post-Op1-Post-Op3) ; « contexte vocalique » : [F (1)= -4.150, p<0.0488)] ; « chirurgie » : [F (3) =14,717 ; p< 0,00003)].

Par rapport à la variable VOT, les 4 effets principaux, « temps », « chirurgie », « contexte vocalique » et « sonorité » se sont montrés statistiquement significatifs : pour le facteur « temps » une différence significative est notée entre (Préop-Post-Op1) : [F (3) = -8,730 ; p< 0,000005)] ; (Préop-Post-Op2) : [F (3) =-6,972 ; p< 0,000004)] ; (Préop-Post-Op3) : [F (3) =-4,009 ; p< 0,018)] ; (Post-Op1-Post-Op3) : [F (3) =-4,721 ; p< 0,003)]; « chirurgie » : [F (3) = 8,189 ; p< 0,00000)]; « contexte vocalique » [F (1)= -4,560, p<0.00000)] ; « sonorité » [F (1) = 20,448 ; p<0.0000)]. L'interaction entre le facteur « temps » et « contextes vocaliques » indique des effets significatifs pour la variable VOT entre (Préop-Post-Op2) : [F (5) = -7,2 ; p<0,005)] dans le contexte [a-i] ; (Préop-Post-Op1) [F (5) = -7,722 ; p<0,0020)] ; (Préop-Post-Op2) : [F (5)= -7,2 ; p<0,005)] dans le contexte [i-a]. Une autre interaction entre les facteurs « temps » et

« chirurgie » pour la variable VOT pourrait être intéressante ; celle-ci indique des différences significatives entre les patients ayant subi une pelvi-glossectomie et ceux partiellement glossectomisés. En effet, l'interaction dévoile un VOT plus important pour les patients pelvi-glossectomisés par rapport au VOT relevé chez les patients glossectomisés en Post-Op2 : [F (5) = -13,005 ; p<0,000004)] ; en Post-Op3 [F (5) = 9,845 ; p<0,000004)]. Cela peut s'expliquer non seulement par la différence des schémas opératoires (la variabilité des traitements chirurgicaux en termes d'étendue de l'exérèse), mais aussi l'importance des traitements complémentaires nécessaires pour chaque exérèse comme la radiothérapie qui viennent généralement bousculer le déroulement de la récupération. Ces contraintes font de la population un groupe hétérogène ce qui aura une incidence non négligeable sur les délais de récupération. En Post-Op3, la différence significative entre les deux types de chirurgie pour la variable VOT témoigne d'une meilleure et réelle réadaptation (réajustement) pour les patients glossectomisés, par rapport à ceux pelvi-glossectomisés. Ces derniers ne réajustent pas forcement le geste articulatoire après 6 mois, mais le réajustement peut probablement avoir lieu plus tard (Zaouali, 2019). L'interaction entre les facteurs « temps » et « chirurgie » pour la variable VOT n'était pas significative, ni en Préop [F (5) = -5,672 ; p<0,067)] ni en Post-Op 1 [F (5) = -4,233 ; p<0,353)]. Cette non-significativité peut être intéressante du moment où cela nous révèle que lors de la phase Préop, les deux groupes de patients ne présentaient pas de modifications significatives pour le VOT et cela malgré les différences dans la stadification qui sont indiquées dans le diagnostic ORL préopératoire et la classification TNM. Cela peut aussi révéler que le Préop peut être considéré comme la parole de référence du patient, quels que soient la taille de la tumeur et son stade d'avancement. Concernant le Post-Op1, la non-significativité entre les deux types de chirurgie pour la variable VOT révèle que lors de cette phase, les perturbations peuvent être importantes malgré la différence entre les deux prises en charge chirurgicales (pelvi-glossectomie et glossectomie). Cela dépend fortement du patient et de sa capacité à adapter ses gestes linguaux dans une nouvelle configuration buccale, après une résection suturée ou une résection nécessitant un changement anatomo-physiologique, donc des possibles reconstructions.

Nous pouvons observer ci-dessous les résultats de l'analyse temporelle de la durée du VOT dans les séquences [ita] *vs*. [ati], pour l'ensemble des patients. Signalons que les patients ROJ, HACH et ANT n'ont pas pu être suivis jusqu'au Post-Op3 (nous nous sommes limités à comparer les résultats de leurs productions qui étaient à notre disposition). En effet, à partir de la figure *infra* (*cf*. Figure 3), nous avons constaté que dans les deux contextes vocaliques [i-a] et [a-i], la durée du paramètre VOT est remarquablement plus longue chez l'ensemble des patients, principalement lors des phases post-chirurgicales 1 et 2. Cela est encore plus prononcé dans le contexte vocalique [a-i] que dans le contexte [i-a]. En ce qui concerne la variabilité, nous en avons observé une plus importante dans le contexte [a-i] que dans le contexte [i-a], et cela chez tous les patients. Cette dernière correspond à des différences inter et intra-individuelles. Cette différence peut être liée à des questions de transition, lors du passage de la consonne à la voyelle (V2=i) [C=> i] (voir *infra* pour une explication possible). Notons que le VOT et le VTT n'étaient pas mesurables pour le patient JCT.

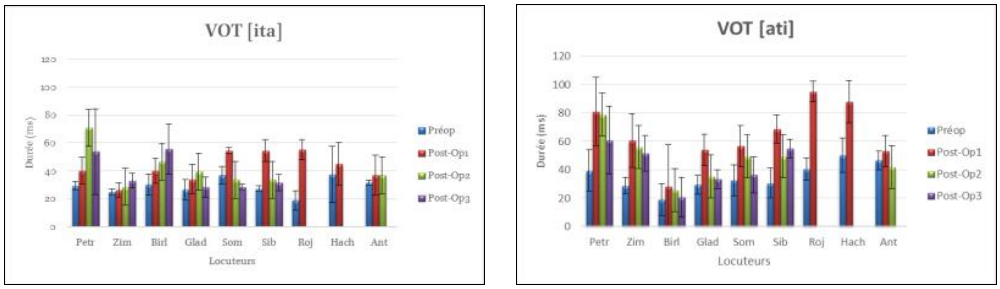


Figure 3 : Valeurs de la durée du VOT en contexte [t] pour les logatomes [ita] (à gauche) et [ati] (à droite), en fonction du patient

En observant de près les productions des patients, la durée du VOT en post-Op1 est généralement plus élevée que celle notée en Préop dans le contexte [i-a]. Cependant, cette observation n'est pas toujours valable pour tous les patients et dans les deux contextes vocaliques. Notons que ce phénomène est à examiner en fonction du patient (type de résection étudiée). En effet, nous avons constaté que la valeur du VOT en Post-Op 1 est élevée chez six patients sur les neuf présentés. C'est le cas chez les patients (ZIM, BIRL, SOM, SIB, ROJ et HACH) pour le contexte [i-a]. Nous avons remarqué une légère diminution des valeurs du VOT toujours en Post-Op1 chez les patients (PETR, GLAD et ANT) dans le contexte [i-a]. Cette baisse dans les valeurs de la durée du VOT est plus notable dans le contexte [a-i], et cela pour les patients (BIRL, SOM, SIB et GLAD). Les valeurs du VOT ont tendance à diminuer ou à se stabiliser, plus ou moins, dans les deux contextes vocaliques, et cela pour tous les patients, à l'exception des patients (PETR et BIRL) dans le contexte [i-a], et de ZIM dans le contexte [a-i] (*cf*. Figure 4).

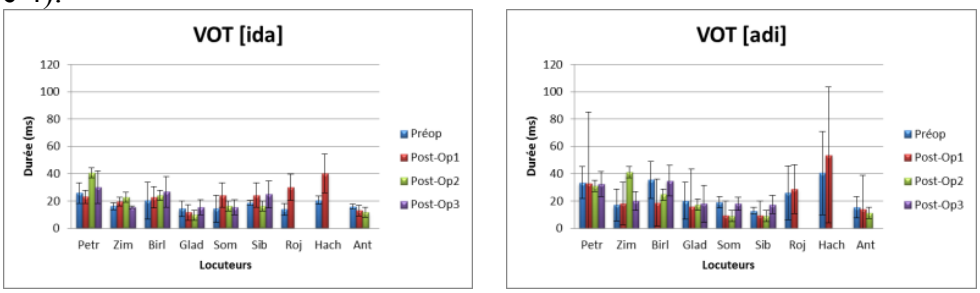


Figure 4 : Valeurs de la durée du VOT en contexte [d] pour les logatomes [ida] (à gauche) et [adi] (à droite), en fonction du patient

En ce qui concerne la variable V2, deux effets principaux se sont révélés significatifs « temps » et « sonorité ». Pour le facteur « temps », la différence se trouve entre le Préop et Post-Op1: [F(3)= -11,023 ; p< 0,003)]. Pour la sonorité, la différence était également entre le Préop et Post-Op1: [F(1)= -22,2625 ; p< 0,0001)]. Ces résultats n'indiquent pas de différences significatives pour les durées de V2 dans les deux contextes vocaliques. Aussi, il apparaît que la durée de V2 est significativement plus longue lorsqu'elle suit une consonne voisée.

5 Discussion et conclusion

À partir de l'analyse des résultats des différents paramètres temporels mesurés, nous avons conclu que les différentes exérèses carcinologiques subies par chaque patient modifient significativement la durée des paramètres retenus. Globalement, l'intervalle de l'ensemble des

paramètres mesurés ont été altérés. En effet, la glossectomie impact de façon significative la durée des segments vocaliques V1 et V2 dans les deux contextes consonantiques voisés ou non-voisés. Néanmoins, l'augmentation de la valeur de V1 est significativement plus importante dans le contexte des occlusives voisées par rapport aux occlusives non-voisées. La durée de V2 s'est montrée remarquablement plus longue que V1 également dans le contexte des occlusives voisées par rapport à leurs homologues non-voisées. Le voisement consonantique favoriserait ainsi une augmentation de la durée des voyelles adjacentes avec qui cette consonne partage le même trait de sonorité.

L'accroissement de la durée du silence acoustique est remarquable pour les deux occlusives non-voisées [t] et [k] et cela dans les deux contextes [i-a] et [a-i]. Concernant les consonnes voisées [d] et [g], la durée de l'occlusion s'est trouvée perturbée après glossectomie pour ces deux occlusives voisées, tant dans le contexte [i-a] que dans le contexte [a-i]. Cette perturbation de la durée de l'occlusion est plus marquée en Post-Op1 et 2. L'étude de la durée du paramètre VTT nous a permis d'observer des modifications significatives, plus ou moins importantes, en fonction du type d'exérèse et de la phase d'enregistrement. Les valeurs du VTT sont très variables entre les différents sujets glossectomisés, et ce dans les deux contextes vocaliques et consonantiques. Il est tout de même important de signaler que les valeurs sont considérablement plus prononcées après chirurgie en Post-Op1 et après radiothérapie en Post-Op2.

Il est fort probable que l'allongement remarquable du VTT soit lié à une difficulté pour les patients de réaliser une occlusion suffisamment étanche, lors de la production d'une occlusive non-voisée. Cette difficulté à obtenir la pression intraorale nécessaire pour la production de cette catégorie de consonne favoriserait un lapse d'amortissement de l'activité laryngée plus long qui déborderait sur la phase silencieuse de ces consonnes non-voisées.

Pour le VOT, nous avons observé que la durée du VOT est notable en Post-Op 1 par rapport au Préop, et cela quasiment pour l'ensemble des patients, tous contextes vocaliques confondus. Nous avons relevé deux tendances principales qui consistent en un allongement plus saillant de la durée du VOT pour les occlusive non- voisées par rapport au VOT de leurs homologues voisées. Le VOT des occlusives voisées tend à augmenter pour certains patients, alors que pour d'autres patients le VOT tend plutôt à diminuer en Post-Op1. Nous avons relevé plus de modifications pour la durée du VOT dans le contexte [a-i] que [i-a]. Cette observation nous révèle qu'il est possible que le problème réside au niveau de la transition [C=> i], puisque cette dernière requiert un contrôle plus précis et qui pose visiblement plus de problème à nos patients glossectomisés que la transition [C=> a]. Certains patients semblent trouver plus d'aisance à réaliser une constriction radico-pharyngale indispensable pour la réalisation de la qualité acoustique de la voyelle [a] que la constriction alvéo-palatale restreinte nécessaire à l'émergence de la voyelle [i]. Les difficultés de transition [C=> V] sont plus importantes pour les patients pelvi-glossectomisés que pour les patients partiellement glossectomisés. Notons qu'une variabilité inter- et intra-individuelle est marquante au niveau de l'ensemble des paramètres mesurés. Les écarts types sont généralement plus élevés en Post-Op1 et 2. Une diminution et une stabilisation des moyennes et des écarts types sont relevées en Post-Op3, suite au réajustement effectué par les patients.

Cette étude nous a permis de voir comment les patients glossectomisés, malgré la perturbation induite par la chirurgie et les différents traitements post-opératoires, peuvent réajuster leurs gestes articulatoires, grâce à l'examen du réaménagement spatiotemporel des différents paramètres acoustiques inter- et intra-segmentaux mesurés. Nous avons ainsi tenté d'aller plus loin que d'autres études traitant des effets de la glossectomie sur la production acoustique de patients (cités *supra*), en examinant le timing du VTT, de V1, de C et de V2, dans les séquences VCV.

Il serait bien de procéder à une orientation du travail de réhabilitation orthophonique vers des voyelles facilitatrices, telles que les voyelles mi-ouvertes et ouvertes du français, et se focaliser sur le temps de récupération à partir du 3ème ou du 6ème mois.

Il serait intéressant aussi de réaliser des tests de perception, et de comparer les deux niveaux acoustique et perceptif, afin de d'évaluer les efficiences de ces réorganisations articulatori-acoustiques.

Références

Abry C., Benoit C., Boë L.J., & Sock R. (1985). Un choix d'événements pour l'organisation temporelle du signal de parole. 14èmes Journées d'Etudes sur la Parole, Société Française d'Acoustique, 133-137.

Acher A., Perrier P., Savariaux C & Fougeron C., Speech production after glossectomy: methodological aspects, Clinical Linguistics and Phonetics. (2014), 28(4), pp. 241-256. DOI: 10.3109/02699206.2013.802015.

Agnello J., (1975). Voice Onset and Voice Termination features of stutterers. In L. M. Webster & L.C. Furst (Eds.), Vocal tract dynamics and dysfluency. 940-954, New York: Speech and Hearing Institute.

Savariaux C., Perrier P., Lebeau J., Magaña G., Dorange-Pattoret C. (2000). Production de parole après traitements de cancers de la cavité endobuccale. In Proceedings of the XXIIIrd Journées d'Étude de la Parole, (pp 433-436), Aussois, France.

Savariaux C., Vilain C., Baciu M., Abry C., Perrier P., Lebeau J., Segebarth C. (2008). Réorganisation du conduit vocal et réorganisation corticale de la parole: de la perturbation aux lèvres à la glossectomie. Études acoustiques et IRMf. Editions de la Maison des sciences de l'Homme (pp. 5–21). DOI: 10.4000/books.editionsmsh.13750

Klatt D.H., (1975). Voice onset time, frication and aspiration in word-initial consonant clusters. Journal of Speech and Hearing Research, 18, 686-706. DOI:10.1044/jshr.1804.686

Sock R (1998). Organisation temporelle en production de la parole. Émergence de catégories sensori-motrices phonétiques. (Doctorat d'Etat). Institut de la Communication Parlée de Grenoble/ INPG & Université Stendhal, Grenoble.

Sock R., & Vaxelaire B. (2001). Réflexions sur le timing de la quantité. Travaux de l'Institut de Phonétique de Strasbourg, TIPS, 31, 89–126.

Vaxelaire, B. (2007). La Résistivité spatio-temporelle des gestes linguistiques. Ou perturber le linguistique en augmentant la vitesse d'élocution. In B. Vaxelaire, R. Sock, G. Kleiber, & F. Marsac (Eds.), *Perturbations et réajustements : langue et langage* (pp. 179–199). Publications de l'Université Marc Bloch – Strasbourg Ville

Zaouali H., Vaxelaire B., Debry C., Schultz P., Bronner G., Sock R (2018). An acoustic study of plosive consonants produced by patients with and without reconstruction after partial or total glossectomy. 2nd International Conference on Natural Language and Speech Processing (ICNLSP). 25-26 Avril 2018 Algiers. pp 60-66. In IEEE *Xplore* Digital Library. DOI: 10.1109/ICNLSP.2018.8374377

Zaouali H. (2019). Étude acoustique de la production de la parole chez des patients glossectomisés (Thèse de Doctorat). UR 1339 Linguistique, Langues et Parole – LiLPa & Institut de Phonétique de Strasbourg – IPS, Université de Strasbourg.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Perception des tons du mandarin par les apprenants français : effets des contextes segmental et syllabique

Qing Zhou Didier Demolin

LPP, UMR 7018 CNRS – Université Sorbonne Nouvelle / Paris 3,
19 rue des Bernardins, 75005 Paris, France

qing.zhou@sorbonne-nouvelle.fr, didier.demolin@sorbonne-nouvelle.fr

RÉSUMÉ

Dans la présente étude, nous rapportons deux expériences visant à explorer les contributions des contextes segmental et syllabique à la perception des tons du mandarin par les apprenants français. Dans la première, des stimuli monosyllabiques produits naturellement, composés de 9 attaques ([ø(zéro), p, t, t^h, tɕ, ɕ, tʂ, $tʂ^h$, m]) et 2 rimes ([i, ɑu]), ont été identifiés par 19 apprenants français de mandarin de niveau débutant et 18 auditeurs de langue maternelle mandarin. Dans la deuxième, les stimuli composés de 6 types de syllabes (V, VV, VN, CV, CVV, CVN) ont été catégorisés par deux autres groupes d'auditeurs. Nos résultats montrent que contrairement aux auditeurs natifs, la perception tonale des apprenants français est influencée de manière significative non seulement par les caractéristiques tonales, mais aussi par les attaque-, rime- and syllabe-types. Cela suggère que les études d'acquisition des tons L2 devraient prendre en compte non seulement le système tonal de la L2, mais aussi le système phonologique segmental de la L2.

ABSTRACT

In the present study, we report two experiments aimed at exploring the contributions of segmental and syllabic contexts to French learners' perception of Mandarin tones. In the first experiment, naturally produced monosyllabic stimuli composed of 9 onsets ([ø(zero), p, t, t^h, tɕ, ɕ, tʂ, $tʂ^h$, m]) and 2 rimes ([i, ɑu]) were identified by 19 beginning-level French learners of Mandarin and 18 native Mandarin listeners. In the second one, stimuli consisting of 6 syllable-types (V, VV, VN, CV, CVV, CVN) were categorized by another two groups of listeners. Our results show that unlike native listeners, French learners' tone perception are significantly influenced not only by tonal features, but also by onset-, rime- and syllable-types. This suggests that L2-tone acquisition studies should take into account not only the L2 tonal system but also the L2 segmental phonological system.

MOTS-CLÉS : Tons du mandarin, perception des tons L2, contexte segmental, apprenants français

KEYWORDS: Mandarin tones, L2 tone perception, segmental context, French learners

Volume 1 : Journées d'Études sur la Parole, pages 664–672.
HAL : hal-02798600.

1 Introduction

1.1 Variations tonales et facteurs segmentaux

Il est largement admis que les contrastes tonals dans les langues asiatiques proviennent de contrastes segmentaux (Haudricourt, 1954 ; Sagart, 1999). De nombreuses études ont indiqué que les tons ont une relation étroite avec les attaques et les rimes en chinois mandarin (Hu, 1987 ; Wu & Lin, 1989 ; Yip, 2002). Les tons sont essentiellement portés par les rimes. Le contour F0 de certains tons varie en fonction de la voyelle (Howie, 1976). Les caractéristiques du F0 intrinsèque se trouvent dans tous les tons du chinois mandarin : plus la voyelle est fermée, plus la F0 intrinsèque est élevée ; plus les valeurs relatives de F0 sont élevées, plus la différence de F0 entre les voyelles est importante (Shi & Zhang, 1987 ; Whalen & Levitt, 1995).

Les effets des consonnes initiales sur la F0 d'une syllabe ont également été bien établis, non seulement pour les langues non-tonales mais aussi pour les langues tonales (Lehiste, 1970 ; Howie, 1974 ; Hombert, 1978). La F0 de départ, le timing de F0 et la durée de F0 sont différents après différentes consonnes pour des raisons biomécaniques et/ou aérodynamiques. Une consonne non-sonante interrompt le mouvement autrement continu de F0 et élève ou rabaisse F0 des voyelles adjacentes (Xu, 2006), alors que les consonnes sonantes sont connues pour présenter la moindre perturbation et interruption de la continuité des contours de F0 (Xu, 1999). Le début de F0 d'un ton est plus élevé précédé de consonnes non-aspirées que précédé de consonnes aspirées (Ballard, 1975 ; Xu & Xu, 2003).

1.2 La perception des tons du mandarin

Des études dans le domaine de la perception ont montré qu'un certain nombre d'indices acoustiques sont fonctionnellement intégrés dans l'identification de tons du mandarin par des auditeurs natifs. Le contour F0 et la hauteur F0 sont considérés comme les principaux indices de l'identité du ton (*e.g.*, Ohala, 1973). La durée syllabique est en corrélation avec la catégorie tonale, du moins lorsqu'elle est prononcée sous la forme de citation (Howie, 1976), et elle affecte également l'identification du ton (Yang, 1989 ; Blicher et al., 1990). En outre, le contour d'amplitude est spécifique à chaque ton et il peut être utilisé comme indice pour l'identification des tons (Coster & Kratochvil, 1984 ; Whalen & Xu, 1992). En ce qui concerne la perception des tons en L2, à part les caractéristiques tonales, ont été trouvé des effets importants de certains facteurs tels que le contexte tonal (Xu, 1994 ; Bent, 2005 ; Chang & Bowles, 2015), le contexte prosodique (Chen, 1997 ; Bent, 2005 ; Hao, 2018), l'expérience linguistique/prosodique de l'apprenant (White, 1980 ; So & Best, 2010, 2014), l'information lexicale (Lee et al., 1996), etc.

D'habitude, des stimuli relativement simples au niveau segmental sont employés dans les études sur la perception des tons, afin de limiter des effets potentiels du contexte segmental et de la structure syllabique sur la F0. Aucune étude antérieure n'a examiné les effets des contextes segmental et syllabique sur la perception tonale des apprenants français du mandarin. Pour les apprenants américains, les tons des syllabes avec une attaque alvéolaire et ceux avec une rime diphtonguée sont avérés plus difficiles à identifier que les tons associés aux autres segments, en revanche, ceux avec une attaque rétroflexe et ceux avec une rime nasale étaient les plus faciles (Lin, 2007 ; Yang, 2012). Le mandarin et le français diffèrent beaucoup en matière d'inventaires consonantique et vocalique ainsi que de structure syllabique. De plus, les auditeurs des langues non-tonales ne perçoivent pas les tons du mandarin de façon catégorielle comme le font les natifs du mandarin (*e.g.*, Hallé et al., 2004) et ils sont plus sensibles aux petites différences de fréquence au niveau phonétique (Stagray & Downs, 1993). On pourrait donc s'attendre à des interactions tons-segments/syllabes dans la perception des tons du mandarin chez les apprenants français.

2 Expérience 1

2.1 Méthode

Participants 19 apprenants français (11 femmes et 8 hommes) de niveau débutant en mandarin, du département chinois à l'Institut National des Langues et Civilisations Orientales (INALCO), âgés de 18 à 24 ans, ont été rémunérés pour leur participation. Tous étaient de langue maternelle française et avaient appris le mandarin durant une année universitaire au moment du test. Tous étaient droitiers. Aucun n'avait signalé de troubles de l'audition ou de la parole ; aucun n'avait reçu de formation musicale formelle avant le test. 18 auditeurs natifs de mandarin vivant en Chine continentale, âgés de 19 à 28 ans (10 femmes, 8 hommes), ont également participé à l'expérience en tant que groupe de contrôle.

Stimuli et design 72 mots monosyllabiques CV placés en position initiale d'une phrase cadre « __ 字我认识。» (Je connais le mot/caractère X.) ont été produits par une locutrice de 28 ans, de langue maternelle mandarin. La syllabe cible était composée de l'une des 9 attaques ([ø(zéro), p, t, t^h, tɕ, ɕ, tʂ, $tʂ^h$, m]) et l'une des 2 rimes ([i, ɑu]).

Le mot cible a été placé en position initiale de la phrase cadre afin d'éviter l'effet de report (*« carry-over » effect*) d'un ton précédent, qui, selon Xu (1994, 1997), est d'une plus grande ampleur que l'effet anticipatoire d'un ton suivant. Mettre le mot cible en position initiale permet aussi de le thématiser. La structure thème – rhème est une structure dominante dans les langues chinoises (Chao, 1968) et il se peut que les locuteurs la produisent avec un certain pattern prosodique, notamment accentuel. En outre, cela permet d'éviter l'interférence de l'intonation finale de la phrase. Les deux rimes (une monophtongue, une diphtongue) ont été choisies parce qu'elles peuvent se combiner avec toutes les 9 attaques et tous les 4 tons pour former 72 (9 attaques * 2 rimes * 4 tons) vrais mots monosyllabiques du mandarin.

Procédure Les fichiers audio des enregistrements ont fait l'objet de découpages dans Praat (Boersma & Weenink, 2018) pour faire correspondre un fichier son par syllabe cible, avec un silence homogène de 50 ms avant l'onset de la syllabe et après l'offset de la syllabe. Les participants ont été invités à s'asseoir dans la chambre sourde du LPP devant un Macbook et ont été équipés d'un écouteur professionnel. Le test a été réalisé avec le logiciel PsychoPy 3.0 (Peirce & MacAskill, 2018). Quatre touches centrales (*f, g, h* et *j*) du clavier (*qwerty*) du Macbook avaient été étiquetées comme des touches de réponse correspondant aux quatre tons. Les stimuli ont été présentés deux fois dans un ordre aléatoire différent pour chaque sujet, avec une pause à mi-chemin, après 72 essais. 16 stimuli différents de ceux utilisés dans le test, produits par un locuteur masculin de langue maternelle mandarin, ont été utilisés lors de la séance de familiarisation qui a précédé le test. La réponse correcte de chaque essai a été donnée uniquement lors de la session de familiarisation, après la réponse de l'auditeur, avec deux couleurs différentes (vert/rouge) indiquant correcte/incorrecte. L'intervalle inter-stimulus (*ISI*) a été fixée à 500 ms. Les participants ont été priés de faire leurs choix aussi rapidement et précisément que possible. Les temps de réponse (TR) ont été mesurés à partir de l'offset des stimuli.

2.2 Analyses acoustiques

2.2.1 Contours de F0

Chaque rime des 72 stimuli monosyllabiques a été segmentée et étiquetée dans Praat, à l'aide d'un script adapté de Xu (2013), le marquage des impulsions vocales de chaque rime a été vérifié et corrigé manuellement. Les valeurs brutes de F0 ont ensuite été lissées par un algorithme

d'ajustement de Xu (1999) et des fichiers de sortie avec F0 normalisée en durée ont été générés automatiquement. 21 mesures de F0 normalisées à durée équidistante ont été prises pour chacun d'entre eux. Les contours moyennés de F0 des quatre tons, conformes aux descriptions consensuelles des 4 tons du mandarin (Chao, 1930 ; Duanmu, 2007), sont démontrés dans la Figure 1 (gauche) : T1 est un ton haut et plat ; T2 est un ton montant ; T3 est un ton bas ; T4 est un ton descendant. T1 est considéré comme un ton « plat » alors que les trois autres comme des tons de « contour ». Notez que T3 en contexte de phrase et non-prépausal est produit comme un ton bas-descendant (demi-T3) (Chen, 2000 ; Zhu, 2012), qui perd la deuxième partie montante de sa forme en contexte isolé. Certains ont suggéré que la hausse finale observée dans sa forme de citation pourrait simplement refléter un mouvement mécanique de retour au niveau F0 de repos, plutôt qu'un marquage linguistique intentionnel (Hallé et al., 2004).

Si nous traçons les contours de F0 en fonction des deux rimes (Figure 1, droite), nous pouvons remarquer que les valeurs F0 de /ao/ sont toujours inférieures à celles de /i/, quel que soit le ton. Cela reflète les valeurs de F0 intrinsèques de ces deux rimes : /i/ est plus élevée que /ao/. À part la hauteur, les contours de F0 sont aussi plus ou moins différents : par exemple, pour T1, le contour F0 est plus courbé avec /ao/, qui s'approche de la forme montante du T2 ; pour T2, son contour F0 s'approche quant à lui de celui de T3 en contexte isolé avec /ao/ : degré plus important de la « chute initiale » et emplacement du « point tournant » plus tard pour T3 (Shen & Lin, 1991) ; etc. Ces variations tonales dues au contexte segmental se manifesteront-elles dans les jugements des auditeurs, surtout des apprenants français du mandarin ?

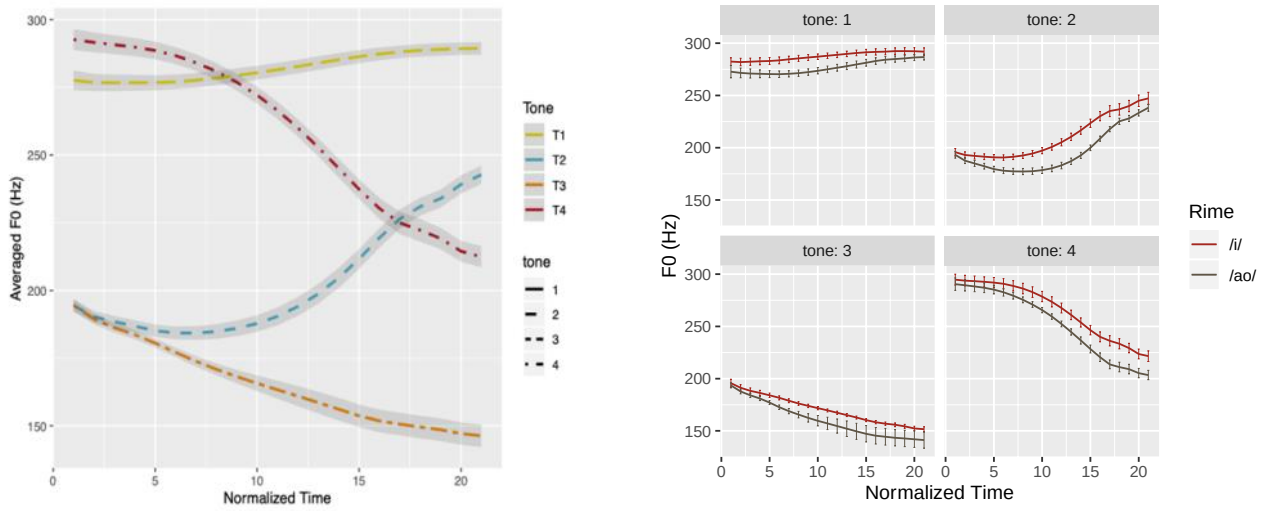


FIGURE 1 : À gauche, contours F0 des quatre tons du mandarin, normalisés en durée, obtenus à partir de 72 stimuli prononcés par une locutrice native. Pour chaque ton, la ligne colorée indique les valeurs moyennes de F0 et l'ombre grise représente les erreurs types. À droite, contours F0 des stimuli tracés en fonction du ton et de la rime (/i/ ou /ao/).

2.2.2 Durées des rimes

À propos des durées des rimes, T2 est le plus long (186ms, et=7.4), suivi de T4 (172ms, et=6.4) et T1 (166ms, et=6.8), tandis que T3 est le plus court (156ms, et=6.2). Une ANOVA à deux facteurs a révélé un effet significatif du ton sur la durée de la rime ($F(3, 64)=5.53$, $p<.01$). T2 est plus long que T1 ($p=.049$) et T3 ($p=.001$). Notez que lorsque T3 est prononcé en contexte de phrase en position non-prépausale (comme dans notre expérience), il a tendance à être plus court que les autres tons (Gottfried & Suiter, 1997), alors que lorsqu'il est prononcé isolément, il est en revanche intrinsèquement le plus long (Howie, 1976 ; Xu, 1997). *Rime* a un effet significatif sur la durée de la rime ($F(1, 64)=40.00$, $p<.001$) : /ao/ est plus longue que /i/. Des différences significatives ont été constatées pour tous les tons sauf T3, selon tests *post-hoc* par pair (Tukey HSD).

2.3 Résultats du premier test de perception

2.3.1 Taux de réponses correctes

Cible \ Réponse	T1	T2	T3	T4
T1	**70.81%**	15.84%	3.88%	9.47%
T2	8.36%	**60.52%**	21.98%	9.13%
T3	8.14%	12.21%	**50.08%**	29.58%
T4	19.97%	16.85%	7.96%	**55.23%**

TABLE 1 : Matrice de confusion des réponses des apprenants français dans le premier test d'identification

Comme le synthétise le Table 1, pour les apprenants français, ton 3 (ton bas, 50.08%) enregistre le taux de réponses correctes le plus bas parmi les quatre tons, suivi du ton 4 (ton descendant, 55.23%), ton 2 (ton montant, 60.52%) et enfin du ton 1 (ton haut, 70.81%) , tandis que les auditeurs mandarins ont obtenu un score proche du plafond dans l'ensemble, bien que T2 (ton montant, 93.6%) ait une précision légèrement inférieure aux trois autres tons. Les résultats de l'analyse statistique ont montré que *Group* était un prédicteur significatif ($\chi^2(1)=1148.7$, $p<.00001$) pour les données binomiales (correcte/incorrecte), reflétant une meilleure performance du groupe natif par rapport au groupe apprenant. Un apprenant français a obtenu le taux de réponses correctes en dessous du niveau de hasard (25%) et ses réponses ont été exclues de l'analyse statistique.

2.3.2 Temps de réponse (TR)

La mesure du TR a commencé à partir de l'offset des stimuli. Les TR (des réponses correctes) supérieurs à 4250 ms (2.5% du total des données) ou inférieurs à 100 ms ont été écartés avant l'analyse statistique. Les TR des apprenants français étaient presque deux fois plus longs que ceux des auditeurs mandarins (voir Table 2). Deux modèles linéaires à effets mixtes avec et sans *Group* comme facteur fixe (tous deux avec *Tone* comme facteur fixe et *Subject* comme facteur aléatoire) ont été effectuées sur les TR, et les résultats du test du rapport de vraisemblance avec la fonction *anova* ont confirmé que *Group* était hautement significatif, $\chi^2(1)=39.48$, $p<.0001$. Les TR des auditeurs français étaient significativement plus longs que ceux des auditeurs mandarins.

		T1	T2	T3	T4
Temps de réponse (ms)	*Français*	1540	1640	1620	1710
	Mandarin	840	812	801	813

TABLE 2 : Valeurs moyennes des temps de réponses correctes (ms) pour les apprenants français et pour les auditeurs mandarins, en fonction du ton.

Deux ANOVA séparées ont ensuite été réalisées sur les TR des deux groupes d'auditeurs, avec *Tone* comme facteur fixe et *Subject* comme facteur aléatoire. Pour les auditeurs mandarins, *Tone* n'était pas significatif, $F(3,23)=0.89$, $p=0.44$. Pour les apprenants français, *Tone* était en revanche significatif, $F(3,14)=4.63$, $p=.0032$. Des résultats similaires ont également été obtenus en appliquant deux modèles linéaires à effets mixtes. Les résultats des tests de comparaison *post-hoc* sur TR des apprenants français ont révélé que les TR de T1 étaient significativement plus courts que ceux de T4 ($p=.0017$). Une corrélation négative modérément forte mais hautement significative a été trouvée entre leurs taux de réponses correctes et les TR en fonction du ton, $r(4)= -0.409$, $p=.00036$, ce qui signifie qu'en général, plus d'erreurs un ton a causées chez les apprenants français, plus de temps il leur faut pour identifier ce ton correctement.

2.3.3 Effets de différents facteurs

Pour examiner les effets de différents facteurs, une série de modèles linéaires généralisés à effets mixtes (*glmer*) ont été appliqués aux données de réponse binomiale (correcte/incorrecte), en utilisant les packages *lme4* et *lmerTest* dans le logiciel R (R Core team, 2018). *Subject* a été défini comme facteur aléatoire, *Onset*, *Rime* et *Tone* comme facteurs fixes. Pour le groupe mandarin, ni *Onset* ni *Rime* ni *Tone* n'ont été significatifs, ni aucun effet d'interaction. Pour le groupe français, *Tone* était significatif, $\chi^2(3)=51.9$, $p<.001$. L'interaction *Rime* * *Tone* était significative, $\chi^2(3)=49.5$, $p<.0001$. L'interaction *Onset* * *Tone* était significative, $\chi^2(24)=76.9$, $p<.001$. L'*Onset* était également significatif, $\chi^2(8)=40.9$, $p<.001$: l'attaque de la fricative palatale (/ɕ/) a causé moins de réponses correctes par rapport aux autres types d'attaque. Quand on appliquait le même modèle au sous-ensemble de données privé de l'attaque /ɕ/, *Onset* n'était plus significatif, $\chi^2(7)=5.5$, $p=0.60$. Cependant, l'interaction *Onset* * *Tone* est restée significative, indiquant que l'identification des 4 tons était réalisée différemment selon les attaques.

Des tests *post-hoc* (*Bonferroni*) ont été effectués sur *Tone* et sur l'interaction *Rime* * *Tone*. T1 était mieux identifié que T3 ($p=.003$) et T4 ($p=.018$), tandis que la différence entre T2 et T3 approchait un niveau significatif ($p=.056$). Les taux d'indentification correcte pour T1 ($p<.001$) et pour T4 ($p=.014$) étaient significativement différentes en fonction de la rime : T1 était mieux identifié avec rime /i/ alors que T4 était mieux jugé avec rime /ao/, par les apprenants français.

2.3.4 Taux d'erreur des apprenants français en fonction de la rime

Le modèle linéaire à effets mixtes (*lmer*) a été adapté au taux d'erreur des réponses des apprenants français avec R, *Error type* et *Rime* étant les facteurs fixes et *Sujet* le facteur aléatoire. *Rime* n'était pas un prédicteur significatif du taux d'erreur ($\chi^2(1)=0.11$, $p=0.74$). Néanmoins, l'interaction *Rime* * *Error type* était significative, $\chi^2(11)=61.09$, $p<.001$. Les taux d'erreur moyens des types d'erreur T1→T2, T1→T3, T2→T1, T3→T2, T4→T1, T4→T3 étaient tous significativement différents en fonction de la rime au niveau $p<.05$, les types d'erreur T2→T3, T3→T1 et T4→T2 au niveau $p<.01$. D'après Figure 2, T1 a été moins souvent identifié comme les autres trois tons avec /i/-rime, alors que T2, T3, T4 ont été tous plus souvent identifiés comme T1 avec /i/-rime. Il semblerait que la monophtongue /i/ ait pu biaiser la perception d'un ton de contour vers un ton plat et haut ; et inversement pour /ao/. Il est possible que le mouvement formantique d'une diphtongue ait été mal interprété comme un mouvement dynamique au niveau tonal par les apprenants français. À cela s'ajoute probablement l'influence de la F0 intrinsèque : /i/ est plus élevée et plus stable que /ao/.

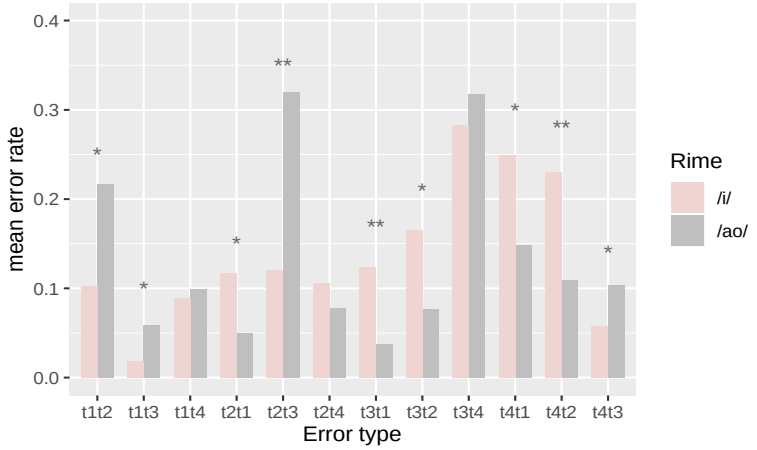


FIGURE 2 : Taux d'erreurs moyens des apprenants en fonction du la rime et du type d'erreur, le nom de chaque type d'erreur comprend le ton cible plus la réponse incorrecte, par exemple, t1t2 signifie le cas où T1 a été identifié comme T2. Niveau de significativité : * pour $p<.05$; ** pour $p<.01$.

3 Expérience 2

3.1 Méthode

Participants 15 apprenants français de niveau débutant en mandarin du département chinois de l'INALCO, âgés de 18 à 24 ans, ont été payés pour leur participation. 15 auditeurs natifs du mandarin vivant en Chine continentale, âgés de 21 à 27 ans, ont également participé à l'expérience en tant que groupe de contrôle.

Stimuli et design 96 mots monosyllabiques de 6 syllabes-types (V, VV, VN, CV, CVV, CVN) placés à la position initiale d'une phrase porteuse ont été produits par un homme de 26 ans, de langue maternelle mandarin. Selon Xu (1998), le rapport N-V (durée de la coda nasale divisée par la durée de la voyelle dans une syllabe avec coda nasale en mandarin) est généralement plus élevé pour les locuteurs masculins. Les syllabes cibles étaient composées de 2 attaques ([ø(zéro), tʂʰ]) et de 3 types de rimes (mono/V, multi/VV, nasal/VN). L'attaque *ch-* (/tʂʰ/) a été choisie parce qu'elle peut se combiner avec beaucoup de rimes et tous les 4 tons pour former de vrais mots monosyllabiques en mandarin, outre le fait qu'il s'agit d'une consonne typique du mandarin (affriquée rétroflexe aspirée) qui ne trouve pas d'homologue en français, tant par son lieu d'articulation que son mode d'articulation.

Procédure Similaire à l'expérience 1, sauf que les auditeurs ont utilisé le *touchpad* du Macbook au lieu du clavier pour faire leurs choix. 4 boutons représentant les 4 tons (avec le numéro de ton marqué en haut et son diacritique en bas) étaient affichés au centre de l'écran de l'ordinateur portable pendant le test, avec une fixation visuelle "+" au milieu, apparue avant chaque essai.

3.2 Résultats du deuxième test de perception

3.2.1 Taux de réponses correctes

Les auditeurs mandarins ont obtenu un score d'identification élevé globalement. Pour les apprenants français, T3 (44.3%) était le moins bien identifié, suivi de T2 (62.2%) et T4 (68.3%), alors que T1 (76.1%) était le mieux identifié. *Group* était un prédicteur significatif ($\chi^2(1)=33.47$, $p<.001$) pour les réponses de type binomiale. La performance des auditeurs mandarins était, sans surprise, significativement meilleure que celle des apprenants français.

3.2.2 Effets de différents facteurs

Pour le groupe français, *Tone* était significatif, $\chi^2(3)=187.8$, $p<.001$, T1 étant le plus facile et T3 le plus difficile à identifier. *Onset* était aussi significatif, $\chi^2(1)=12.5$, $p=.0004$. Les tons des syllabes à attaque vides étaient plus faciles à identifier que ceux avec l'attaque *ch-* (/tʂʰ/]). Selon comparaisons *post-hoc*, c'est le cas pour tous les tons ($p<.05$) sauf T1. *Syllable* était significative, $\chi^2(5)=187.8$, $p=.012$. Les tons des syllabes de type CVV étaient significativement plus difficiles à identifier que celles de type V ($p=.014$), de type VV ($p=.017$) et de type VN ($p=.002$) ; les tons des syllabes de type VN étaient significativement plus faciles à identifier que ceux de type CV ($p=.012$) et de type CVV ($p=.002$). Figure 3 permet de visualiser ces différences.

L'interaction *Onset* * *Tone* ($p=.003$), l'interaction *Rime* * *Tone* ($p=.0004$) et l'interaction *Syllabe* * *Tone* ($p<.001$) étaient toutes significatives, indiquant que la catégorisation des tons était réalisée différemment selon les types d'attaque, de rime ou de syllabe. *Rime* seule n'était pas significative ($p=0.38$). Les rimes avec une coda nasale étaient significativement plus faciles à identifier que les rimes à multiples voyelles ($p=.049$), qui étaient les moins bien identifiées parmi les trois types de rimes, bien que cette tendance n'ait pas atteint un niveau significatif.

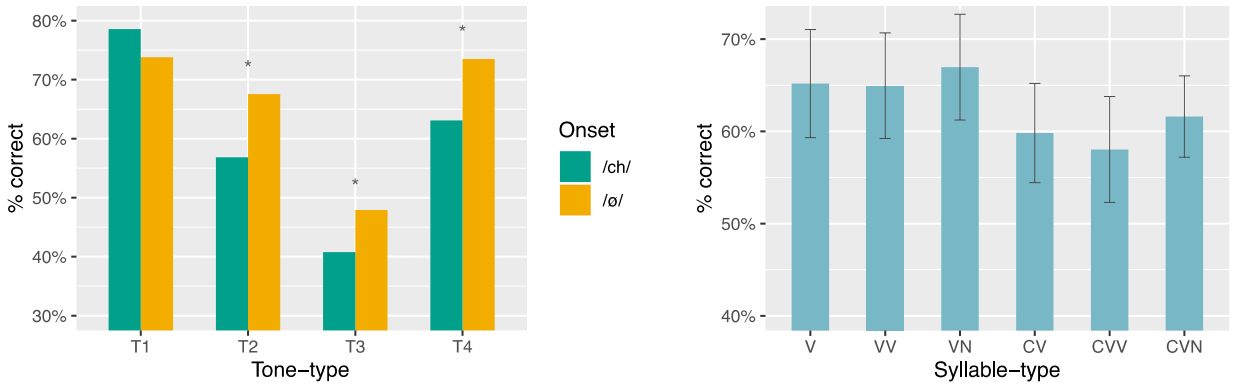


FIGURE 3 : À gauche, taux de réponses correctes pour l'identification tonale des apprenants français en fonction du ton et de l'attaque ; à droite, leur taux de réponses correctes en fonction du type de syllabe. (Niveau de significativité : * pour p<.05)

4 Conclusion

Dans deux expériences, la perception des tons du mandarin par les apprenants français a été examinés, en utilisant des stimuli avec des types de segments et de syllabes variés. Les résultats indiquent que le contexte segmental a souvent des effets significatifs sur l'identification tonale des apprenants français, que ce soit pour la rime, l'attaque, ou la structure syllabique. Par exemple, la rime monophtongue /i/ et la rime diphtongue /ao/ interagissaient différemment avec les caractéristiques tonales et ont rendu des patterns de catégorisation tonale différents ; les tons avec l'attaque /ɕ/ semblaient plus difficile à identifier que ceux avec d'autres attaques ; les tons associés à l'attaque rétroflexe affriquée aspirée /tʂʰ/ étaient plus difficiles à identifier que ceux avec zéro-attaque ; les tons des syllabes de type CVV étaient les plus difficiles à identifier et ceux des syllabes de type VN étaient les plus faciles. Pour les apprenants dont la langue maternelle est non-tonale, l'association fonctionnelle entre la structure segmentale et les caractéristiques suprasegmentales leur est peu familière et leur catégorisation tonale pourraient être perturbée par les variations au niveau segmental. Cela suggère aussi que l'enseignement des tons du mandarin ne devrait pas se limiter à donner des formes tonales abstraites et il semble nécessaire de leur exercer avec divers types de segments et de syllabes pour favoriser la formation des catégories tonales chez les apprenants.

Nos résultats montrent aussi que les caractéristiques tonales impactent significativement la catégorisation tonale des apprenants français : T3 était le plus difficile à identifier et T1 le plus facile. De plus, la confusion entre paires tonales était souvent asymétrique : T3 était le plus souvent classifié comme T4 alors que T4 était rarement identifié comme T3 ; T2 était le plus souvent jugé comme T3 et non pas vice-versa. Alors que la plupart des chercheurs s'accordent sur la difficulté du T2 dans l'acquisition, celle du T3 semble sous-estimée. Notez que T3 présente le plus de variations allophoniques parmi les quatre tons. Il est prononcé comme un ton bas-descendant sans partie remontante (demi-T3) dans la plupart des cas, en parole continue. Nous pensons que l'enseignement du T3 devrait mettre l'accent sur son côté « bas », au lieu de sa forme de citation « descendant-remontant », pour que les apprenant puissent mieux le connaître, le reconnaître et mieux le différencier tant du T4 que du T2.

Remerciements

Nous tenons à exprimer notre gratitude pour le précieux soutien des étudiants et des enseignants du département chinois de l'Institut national des langues et civilisations orientales (INALCO) à Paris. Ce projet en cours est financé par *China Scholarship Council* (CSC).

Références

Balard, W. L. (1975). Wu tone sandhi. Paper presented at *the 8th International Conference on Sino-Tibetan Languages and Linguistics*.

Bent, T. (2005). *Perception and production of non-native prosodic categories*. PhD Dissertation. Northwestern University.

Blicher, D., Diehl, R., & Cohen, L. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, **18**, 37–49.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

Chang, C., & Bowles, A. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America*, **138**, 3703–3716.

Chao, Y.-R. (1930). A system of tone letters. *La Maître phonétique*, **45**, 24–27.

Chao, Y.-R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, M. (2000). *Tone sandhi: Patterns across the Chinese dialects*. Cambridge: Cambridge University Press.

Chen, Q.-H. (1997). Toward a sequential approach for tonal error analysis. *Journal of the Chinese Language Teachers' Association*, **32**(1), 21–39.

Coster, D. C., & Kratochvil, P. (1984). Tone and stress discrimination in normal Peking dialect speech. In B. Hong (Ed.), *New papers in Chinese linguistics* (pp. 119–132). Canberra: Australian National University Press.

Duanmu, S. (2007). *The Phonology of Standard Chinese* (2nd edition). Oxford: Oxford University Press.

Gandour, J. (1984). Tone dissimilarity judgments by Chinese listeners. *Journal of Chinese Linguistics*, **12**, 235–261.

Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, **25**(2), 207–231.

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, **32**(3), 395–421.

Hao, Y.-C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication*, **97**, 32–42.

Haudricout, A-G. (1954). De l'origine des tons en vietnamien, *Journal Asiatique*, **242**, 69–82.

Hombert, J-M. (1978). Consonant types, vowel quality and tone. In: Victoria A. Fromkin (Ed.), *Tone: A Linguistic Survey* (pp. 77–111). New York: Academic Press.

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica*, **30**, 129–148.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press.

Hu, Y. (1987). *Xiandai Hanyu*. Shanghai: Shanghai Gaodeng Jiaoyu Chubanshe.

Lee, Y.-S., Vakock, D., & Wurm, L. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, **125**, 527–542.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge: MIT Press.

Lin, Y. (2007). *The sounds of Chinese*. Cambridge: Cambridge University Press.

Ohala, J. J. (1973). The physiology of tone. In L. Hyman (Ed.), *Consonant types and tone*, (pp. 1-14). University of Southern California.

Peirce, J. W., & MacAskill, M. R. (2018). *Building Experiments in PsychoPy*. London: Sage

R Core Development Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Sagart, L. (1999). The origin of Chinese tones. *Proceedings of the Symposium/Cross-Linguistic Studies of Tonal Phenomena/Tonogenesis, Typology and Related Topics*, Tokyo, Japan. Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, 91-104.

Shen, X.S., & Lin, M. (1991). A perceptual study of Mandarin Tone 2 and Tone 3, *Language and Speech*, **34**(2), 145–156.

Shi, B., & Zhang, J. (1987). Vowel intrinsic pitch in standard Chinese. *Proceedings of the 11th International Congress of Phonetic Sciences*, 142–145.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. *Language and speech*. **53**(2), 273–293.

So, C. K., & Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, **36**(2), 195–221.

Stagray, J. R., & Downs, D. (1993). Differential sensitivity for frequency among speakers of a tone and a non-tone language. *Journal of Chinese Linguistics*, **21**(1), 143–163.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, **49**, 25–47.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics*, **23**, 349–366.

White, C. (1980). *Mandarin tone and English intonation: a contrastive analysis*. MA Dissertation. University of Arizona.

Wu, Z. & Lin, M. (1989). *Shiyan Yuyinxue Gaiyao* (Introduction à la phonétique expérimentale), Shanghai: Gaodeng Jiaoyu Chubanshe.

Xu, C. & Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association*, **33**, 165–181.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, **95**(4), 2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, **25**, 61–83.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, **55**, 179–337.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, **27**, 55–105

Xu, Y. (2006). Principles of tone research. *Proceedings of International Symposium on Tonal Aspects of Languages*, La Rochelle, 3–13

Xu, Y. (2013). ProsodyPro – A Tool for Large-scale Systematic Prosody Analysis. In Proceedings of *Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France, 7–10.

Yang, B. (2012). The gap between the perception and production of tones by American learners of Mandarin: an intralingual perspective. *Chinese as a Second Language Research*, **1**(1), 31–52.

Yang, Y.-F. (1989). Yuanyin he shengdiao zhijue (Voyelles et la perception des tons du chinois*). Xinlixue Bao*, **34**, 29–34.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Zhu, X. (2012). Jiangdiao de Zhonglei (Types de tons descendants). *Yuyan Yanjiu*, **32**(2), 1–16.