Effets du locuteur sur la voix et la parole : des affects à la pathologie

Nicolas Audibert

Université Sorbonne Nouvelle

Habilitation à Diriger des Recherches

Volume 1 : Document de synthèse

Effets du locuteur sur la voix et la parole : des affects à la pathologie

Nicolas Audibert

Habilitation soutenue le 13 janvier 2025 à Paris

Composition du jury

- **Christine Meunier**, Directrice de recherche CNRS, Laboratoire Parole & Langage Aix-en-Provence, Présidente du jury
- **Rudolph Sock**, Professeur des universités, Université de Strasbourg, Rapporteur
- **Hans Van de Velde**, Professeur, Université d'Utrecht & Fryske Akademy, Rapporteur
- **Véronique Delvaux**, Chercheur qualifié FNRS, Université de Mons, Examinatrice
- **Corinne Fredouille**, Professeure des universités, Avignon Université, Examinatrice
- **Cécile Fougeron**, Directrice de recherche CNRS, Laboratoire de Phonétique et Phonologie, Garante

A Irene et Daria,

Remerciements

En tout premier lieu, je voudrais remercier Cécile Fougeron de m'avoir accompagné en tant que garante dans cette aventure, en me poussant tout d'abord à franchir le pas puis en m'aidant à structurer mes idées. Cécile, nous avons tellement collaboré au cours des quinze dernières années que personne ne connaît mieux que toi mon travail, tu étais donc la personne toute indiquée pour jouer ce rôle de garante. Au-delà du travail commun déjà accompli et des multiples codirections de travaux d'étudiants, j'espère bien ne pas m'arrêter en si bon chemin, et continuer à collaborer avec toi dans les prochaines années.

Je tiens aussi à remercier les collègues qui ont accepté de participer à mon jury, et tout d'abord mes deux rapporteurs Rudolph Sock et Hans Van de Velde, ainsi que mes examinatrices Véronique Delvaux, Corinne Fredouille et Christine Meunier. L'ambiance amicale dans laquelle s'inscrit le travail de recherche me tient particulièrement à cœur et a été l'un des aspects qui m'ont convaincu de poursuivre dans cette voie. Je suis donc absolument ravi d'avoir un aussi beau jury, composé de collègues pour qui j'ai non seulement une grande estime sur le plan scientifique mais aussi beaucoup de sympathie.

Je voudrais aussi remercier les collègues qui m'ont fait confiance après ma thèse et ont rendu possible tout le chemin parcouru depuis, en particulier Jean-François Bonastre qui m'a ouvert les portes du LIA pour mon premier séjour post-doctoral, puis Jacqueline Vaissière qui m'a permis de rejoindre le LPP, et enfin Christophe d'Alessandro qui m'a accueilli au LIMSI pour une nouvelle parenthèse post-doctorale avant mon retour au LPP.

Mes activités de recherche étant très largement collaboratives, je veux aussi remercier les nombreux collègues avec qui j'ai collaboré dans la bonne humeur pendant toutes ces années, à la fois au sein du LPP et dans d'autres unités de recherche dans le cadre des différents projets auxquels j'ai participé. La liste serait trop longue pour remercier chacune et chacun individuellement, et je ne voudrais pas vexer quiconque en oubliant de mentionner certains noms ou, pire, en désignant comme ma préférée l'une des doctorantes dont j'ai codirigé la thèse (pour les principales intéressées qui liraient ces lignes, il s'agit bien entendu d'un clin d'œil à une conversation avec trois d'entre vous lors de la dernière édition des JEP). Je remercie également mes camarades de l'AFCP que j'ai toujours autant de plaisir à retrouver à l'occasion des réunions du conseil d'administration et dans les éditions successives des JEP et des JPC.

Bien que ce dossier soit centré sur mes activités de recherche, la vie professionnelle d'un enseignant-chercheur s'articule aussi très largement autour des activités d'enseignement et de toutes les tâches annexes liées de près ou de loin à ces activités. Je veux donc remercier aussi chaleureusement mes collègues de l'ILPGA pour leur implication et tous les moments agréables passés ensemble. Aussi chronophages soient-elles, les tâches administratives semblent tellement plus légères lorsqu'elles sont effectuées au sein d'une équipe dynamique et solidaire et lorsqu'il est possible de rire ensemble de certaines situations.

Enfin je veux remercier tout particulièrement ma chère épouse Irene qui a fait d'immenses efforts pendant l'été précédant le dépôt de ce dossier d'habilitation pour me laisser le temps de me plonger dans la rédaction. Merci infiniment, sans toi tout cela n'aurait pas été possible ! Merci aussi à toi ma petite Daria d'avoir été si patiente et d'avoir accepté que je ne t'accompagne pas à la mer, que je rentre seul à Paris puis que je reste travailler pendant tes sorties au parc lors desquelles tu demandais à tes copines « Toi aussi ton papa travaille ? ». L'an prochain tu pourras compter sur moi pour les châteaux de sable, c'est promis !

Table des matières

Avant-propos : structure du dossier

La partie scientifique de ce dossier d'habilitation se subdivise en trois volumes :

- **Volume 1 : Document de synthèse** (251 pages, ce volume)
  Après une brève introduction qui résume mon parcours et expose les grandes orientations qui sous-tendent ma carrière scientifique, ce volume récapitule et met en perspective mes travaux de recherche postérieurs à ma thèse de doctorat à travers une structuration thématique. Dans ce volume, je propose également des analyses inédites centrées sur les spécificités individuelles des locuteurs à la suite de la présentation de mes travaux antérieurs s'appuyant sur les mêmes données.

- **Volume 2 : Volume de travaux publiés** (519 pages)
  Ce volume regroupe 62 articles publiés après ma thèse et non liés à celle-ci, parus dans des revues à comité de lecture, actes de conférences ou colloques ou chapitre d'ouvrage et ayant fait l'objet d'une évaluation par les pairs.

- **Volume 3 : Document analytique** (29 pages)
  Ce volume présente brièvement chacun des articles inclus dans le volume de travaux publiés, et expose les orientations que j'entends donner à mes recherches dans les prochaines années, ainsi que la façon dont j'envisage l'encadrement doctoral.

Outre cette partie scientifique et les pièces administratives requises, mon dossier inclut également en tant que document complémentaire mon Curriculum Vitæ analytique (40 pages), qui détaille mes activités d'animation de la recherche, d'enseignement et autres activités administratives que je ne mentionne par ailleurs que brièvement dans la partie introductive du Volume 1.

Conventions de notation

Afin d'éviter de surcharger ce document, j'ai pris le parti d'employer la forme classique du masculin générique plutôt que le point médian ou le recours à des paraphrases, et je serai amené dans le corps du texte à faire régulièrement référence aux locuteurs, aux auditeurs ou encore aux étudiants. Toutefois sauf mention contraire explicite, en particulier à l'occasion de la présentation des travaux dans lesquelles j'ai procédé à des comparaisons entre données produites par un groupe de femmes ou par un groupe d'hommes, les termes masculins désignant des groupes de personnes concernent autant les femmes que les hommes.

Dans la présentation que je fais dans ce document de synthèse de mes travaux scientifiques, bien que ceux-ci s'appuient sur des analyses statistiques je ne présente en général pas directement de résultats statistiques chiffrés. Néanmoins pour quelques-uns de mes travaux dans lesquels une partie de l'analyse repose sur des corrélations, la présentation de certaines valeurs est nécessaire à l'interprétation des résultats présentés. Dans ce cas, par souci de lisibilité je les présenterai en adoptant la convention anglo-saxonne de notation des nombres décimaux, avec le point comme séparateur décimal et en omettant le zéro dans les nombres dont la valeur est comprise entre 0 et 1 (par exemple pour une corrélation de Spearman : $\rho = .75$).

La bibliographie est formatée selon les normes de l'American Psychological Association (APA, 7ème édition), qui avec celles de l'IEEE dans les champs disciplinaires plus orientés vers les technologies de la parole sont les plus largement utilisées en sciences de la parole. En raison de quelques cas d'homonymie parmi les auteurs que je cite, conformément à ces normes certaines citations dans le texte sont précédées de l'initiale du prénom du premier auteur à des fins de désambiguïsation. Dans les cas de citations multiples du même premier auteur avec la même année de publication mais un second auteur différent, les citations sont désambiguïsées en mentionnant également dans le texte le nom du second auteur. Enfin dans les cas de citations multiples du même auteur avec la même année de publication, la désambiguïsation repose sur l'utilisation d'une lettre minuscule à la suite de l'année de publication.

1 Introduction : parcours et positionnement scientifique

1.1 Mon parcours : de l'informatique à la phonétique

Mon intérêt pour la recherche en phonétique, et plus généralement pour la linguistique, a été relativement tardif comparativement aux collègues et étudiants que je côtoie quotidiennement et qui pour la plupart sont issus d'une formation en sciences du langage ou qui intègre une composante linguistique. En effet, bien qu'intégrant en dernière année une spécialisation en génie logiciel, bases de données et intelligence artificielle, ma formation universitaire initiale en ingénierie informatique était éloignée de ces thématiques, le traitement automatique du langage n'y étant mentionné qu'à titre d'exemple d'application de l'intelligence artificielle en l'absence de spécialistes de cette discipline parmi les enseignants-chercheurs qui intervenaient dans cette formation. Ce n'est qu'au début de l'année 2002 que j'ai rejoint en tant que stagiaire l'Institut de la Communication Parlée de Grenoble (devenu au début de ma thèse département Parole et Cognition de Gipsa-lab) pour y effectuer mon stage de fin d'études de six mois sous la direction de Véronique Aubergé sur un projet de développement informatique d'une application destinée à la recherche en parole. Je suis ainsi arrivé sur la pointe des pieds, avec initialement une compréhension très partielle des objectifs scientifiques associés à l'outil que j'allais être amené à développer, mais déjà la conviction de rejoindre un environnement intellectuel stimulant et le souhait de poursuivre dans cette voie, conviction rapidement confortée par les échanges avec les doctorants et chercheurs du laboratoire et mon intérêt pour les séminaires que j'ai eu la possibilité de suivre. Afin de me consacrer plus avant à cette passion naissante pour la recherche sur le langage et la parole, j'ai donc fait le choix de rester dans le monde de la recherche à travers un DEA en sciences cognitives, puis un second en sciences du langage qui m'a permis de parfaire mes connaissances en la matière et d'évoluer progressivement de l'informatique vers la phonétique. Dans ce cadre, j'ai été associé à des premières publications et ai eu la chance de participer à plusieurs conférences, ce qui a achevé de me convaincre de tout faire pour consacrer la suite de ma carrière à la recherche en parole.

Grâce à l'attribution d'une allocation par le CNRS (Bourse de Doctorat pour Ingénieurs), j'ai pu débuter en 2004 une thèse dirigée par Véronique Aubergé et Jean-Luc Schwartz et portant sur l'étude des expressions prosodiques des émotions spontanées et simulées, reposant sur l'hypothèse d'une information prosodique affective portée par des contours gradients multiparamétriques incluant les variations de qualité de voix. Cette thèse m'a amenée à concevoir et mettre en place un protocole de recueil d'extraits de parole émotionnelle spontanée produite dans le contexte linguistique minimaliste de mots monosyllabiques produits isolément afin de séparer les variations de hauteur mélodique et de qualité de voix liées à l'émotion de celles conditionnées par la structure de l'énoncé, complétés par des productions simulées par les mêmes locuteurs ayant une expérience du jeu d'acteur dans un contexte directement comparable. Au-delà du français, ce protocole a été étendu au hongrois en collaboration avec le Speech Lab de Budapest. A partir de l'analyse acoustique des productions spontanées recueillies j'ai identifié des contours distincts associés au ressenti émotionnel étiqueté par les locuteurs. En collaboration avec l'équipe d'Orange-labs spécialisée en synthèse de la parole j'ai pu mettre en évidence un poids perceptif relatif des variations de fréquence fondamentale et de qualité de voix dépendant de la valence de l'émotion exprimée. D'autres expériences perceptives menées à partir de ces données ont montré la capacité d'auditeurs naïfs à discriminer expressions spontanées et simulées avec une grande variabilité

inter-individuelle, et ont suggéré un décodage de l'information affective à partir de la dynamique des contours et non simplement sur des caractéristiques globales telles que le registre de fréquence fondamentale. Une collaboration avec Petri Laukka (université d'Uppsala) a permis de montrer à partir du croisement des résultats de multiples analyses perceptives que les expressions des émotions dans la parole sont représentées comme des catégories liées au but au sens défini par Barsalou en psychologie cognitive plutôt que comme des catégories taxonomiques classiques. Enfin, en marge de ces travaux de thèse j'ai pu en collaboration avec Solange Rossato m'initier à l'approche métrologique de la phonétique à travers l'évaluation d'une mesure de qualité de voix obtenue par filtrage inverse.

Suite à ma soutenance en 2008, j'ai intégré en 2009 pour une année de post-doc l'équipe de Jean-François Bonastre au Laboratoire d'Informatique d'Avignon, où j'ai travaillé à la fois sur la reconnaissance du locuteur par l'humain et par la machine, et sur l'analyse automatique de la parole dysarthrique avec une approche à l'interface de la phonétique et du traitement automatique. En 2010, j'ai ensuite rejoint pour un second post-doc le Laboratoire de Phonétique et Phonologie (LPP), où je me suis consacré notamment à l'analyse multiparamétrique de signaux physiologiques tout en prolongeant les travaux amorcés auparavant sur la parole dysarthrique. Au début de l'année 2012, j'ai débuté un dernier post-doc au laboratoire LIMSI (devenu depuis LISN) dans l'équipe de Christophe d'Alessandro sur l'analyse et la synthèse de la parole expressive. C'est alors que j'ai été recruté comme Maître de conférences en sciences phonétiques au département ILPGA de l'université Sorbonne Nouvelle à partir de septembre 2012, ce qui m'a permis de retrouver cet environnement scientifique et humain particulièrement stimulant qu'est le LPP.

Depuis mon recrutement, j'ai développé une nouvelle facette de mon rapport à la phonétique et aux sciences de la parole sur lesquelles porte la plus grande partie de mon activité d'enseignement, qui auparavant concernait très majoritairement l'informatique. Si je reste également impliqué dans l'enseignement des statistiques ainsi que de l'algorithmique appliquée à l'automatisation des analyses acoustiques, j'ai ainsi été amené dès mon recrutement à reprendre la responsabilité du cours de phonétique en première année de licence et à proposer un cours de master sur la parole expressive, de même que d'autres cours axés sur la méthodologie expérimentale, la cognition et la psycholinguistique, ou plus récemment la prosodie.

Sur le plan scientifique, suite à mon retour au LPP en 2012 j'ai poursuivi la diversification de mes thématiques de recherche amorcée lors de mon premier post-doc, en m'inscrivant dans un premier temps dans la continuité de mes travaux précédents sur la parole expressive à travers la codirection avec Jacqueline Vaissière de la thèse de Charlotte Kouklia et des travaux ultérieurs sur l'expression d'émotions et d'autres affects. Une autre partie conséquente de mes travaux scientifiques a porté sur la variation entre locuteurs et entre styles de parole, principalement à travers la caractérisation de voyelles. Outre l'analyse des voyelles, j'ai progressivement élargi mon champ scientifique à l'étude des consonnes obstruantes afin d'apporter des éléments de réponse à diverses questions de recherche, principalement via l'analyse de grands corpus de parole continue. J'ai également étendu mon intérêt pour la phonétique clinique à travers des études portant sur la parole dysarthrique et la voix dysphonique, ainsi que les liens entre parole et vieillissement dont une meilleure compréhension est nécessaire à l'étude de pathologies de la voix et de la parole affectant les sujets âgés.

Ce sont ces travaux ainsi que ceux menés lors de mes trois années de post-doc précédant mon recrutement et qui ne s'inscrivaient pas dans la continuité de ma thèse que je récapitule dans les chapitres suivants de ce document de synthèse, après avoir développé dans les sections suivantes certains des aspects qui caractérisent selon moi mon approche de la recherche de façon transversale aux multiples thématiques scientifiques que j'ai abordées au cours des quatorze dernières années.

1.2 Implication dans les tâches collectives

Mon recrutement comme Maître de Conférences a également fait évoluer mon implication dans l'université Sorbonne Nouvelle, au-delà des rapports de voisinage amical que j'entretenais déjà lors de mon post-doc avec les membres du département ILPGA dédié aux Sciences du Langage, favorisés par la situation du LPP et de ce département dans les locaux historiques de l'Institut de Phonétique de Paris jusqu'à l'emménagement de la Sorbonne Nouvelle dans les locaux du campus Nation. Dès mon recrutement, j'ai été amené à assurer des responsabilités collectives d'enseignement et d'administration, d'abord modestes puis rapidement plus conséquentes. J'ai ainsi assumé la fonction de directeur-adjoint de département de 2014 à 2017, particulièrement chronophage en raison notamment du statut de site isolé du département mais également très formatrice, et celle de responsable du master parcours phonétique et phonologie depuis 2019 qui implique non seulement des tâches classiques de sélection des dossiers, d'information et de représentation mais aussi la coordination avec l'Université Paris Cité avec laquelle ce parcours est organisé en partenariat. De plus, je suis en charge des mobilités internationales entrantes et sortantes du département ILPGA depuis 2017, et je siège ou ai siégé en tant que membre élu ou nommé dans plusieurs conseils au niveau du département et de l'UFR.

Par ailleurs je me suis également impliqué dans l'animation de la recherche à travers ma participation à la société savante AFCP (Association Francophone de la Communication Parlée) qui fédère la communauté francophone de recherche en phonétique et en traitement automatique de la parole, l'organisation de colloques (tout particulièrement la conférence JEP-TALN-RECITAL en 2016 dont j'ai été l'un des principaux organisateurs), et la prise de responsabilités au sein du LabEx EFL (Empirical Foundations of Linguistics) que j'ai rejoint en tant que membre du LPP dès mon recrutement.

1.3 Une approche collaborative de la recherche

Comme en attestent mes publications, ma pratique de la recherche est éminemment collaborative. Ainsi, sauf rares exceptions mes publications sont toujours cosignées par les coauteurs ayant contribué à des degrés divers au travail présenté. Cela correspond aux pratiques auxquelles j'ai été habitué dès mes premiers pas dans la recherche, et qui sont largement partagées dans la communauté francophone et internationale de recherche en phonétique et en sciences de la parole. Dans quelques cas minoritaires l'inclusion de coauteurs peut relever d'une politique de copublication systématique dans le cadre de certains projets collaboratifs, ou correspondre à l'exploitation de données collectées par ces coauteurs pour les besoins d'autres études, mais dans la grande majorité des cas elle reflète leur implication effective dans les travaux publiés. Bien qu'on ne puisse en déduire directement une proportion du travail réalisé, l'ordre des auteurs dépend en règle générale de leur degré relatif d'implication dans la réalisation des travaux présentés et la rédaction de l'article. La seule exception récurrente concerne certaines publications dont la première auteure est une

étudiante en master ou en début de doctorat, avec souvent une implication plus importante des chercheurs permanents que ce que suggère l'ordre des auteurs, en raison d'une politique de la part de mes collègues et de moi-même de valoriser les travaux de recherche publiables des étudiants quand bien même la finalisation et la présentation des réalisations de jeunes chercheurs encore peu expérimentés nécessitent un travail plus conséquent de la part des chercheurs confirmés.

Outre ces usages en matière de publication, mon penchant pour cette approche collaborative de la recherche a par ailleurs été renforcé par ma participation à de multiples projets financés (pour la plupart par l'Agence Nationale de la Recherche), à la fois préalablement à mon recrutement et depuis celui-ci. En complément, j'ai également été amené à nouer des collaborations scientifiques plus informelles avec des collègues, au sein du Laboratoire de Phonétique et Phonologie mais aussi en dehors de celui-ci. Ces collaborations scientifiques se sont faites en majeure partie avec des collègues exerçant dans des laboratoires francophones, du fait des contacts établis pendant ma thèse et lors de mes séjours post-doctoraux et de mon implication dans les activités de l'association savante AFCP (Association Francophone de la Communication Parlée) et les conférences organisées par celle-ci (Journées d'Etudes sur la Parole et Journées de Phonétique Clinique) auxquelles je participe assidûment, depuis 2004 dans le cas des JEP. Au-delà des collaborations nationales et internationales amorcées lors de ma thèse et hormis le projet européen i-Treasures auquel j'ai contribué peu après mon recrutement, j'ai plus largement collaboré avec des collègues affiliés à des laboratoires français, probablement en partie en raison du rôle croissant joué par l'ANR dans le financement de la recherche publique en France qui tend à influencer par ricochet la structuration de la recherche. Au cours des dernières années, j'ai eu l'occasion de renforcer mes liens avec l'équipe de recherche en phonétique de l'université belge de Mons à travers un Partenariat Hubert-Curien (PHC) sur la thématique de la qualité de voix auquel j'ai participé activement, qui a donné lieu à des communications communes et surtout des collaborations scientifiques qui se poursuivent après la fin officielle de ce projet. J'ai également été récemment porteur pour la partie française d'un autre projet PHC avec l'université néerlandaise de Leeuwarden sur la thématique de la sociophonétique de laboratoire, qui a permis d'amorcer des travaux sur la variation de réalisation des consonnes labiodentales en français ainsi que de poser les jalons d'autres travaux collaboratifs sur le français et le frison.

Depuis mon recrutement à la Sorbonne Nouvelle, une part conséquente de mon activité de recherche s'exerce dans le cadre de l'encadrement d'étudiants de master (principalement en master Phonétique et phonologie, et dans une moindre mesure en TAL, en orthophonie et dans d'autres parcours de master Sciences du Langage) et de doctorat. Je me suis toujours fortement impliqué dans cette activité d'encadrement qui me tient particulièrement à cœur, au-delà de la définition de la problématique et des objectifs de recherche et de la relecture de versions intermédiaires des mémoires des étudiants. J'ai ainsi pris l'habitude, dès la première année de master et même dans le cadre de certains projets de troisième année de licence avec les étudiants les plus motivés d'assurer le suivi des travaux des étudiants à travers des entretiens réguliers, des conseils de lecture et une assistance apportée lorsque cela s'avérait nécessaire aux différentes étapes de conception des protocoles expérimentaux, de recueil, d'annotation et d'analyse des données.

Dans la présentation que je fais dans ce document de synthèse des travaux auxquels j'ai contribué et dont la majeure partie sont issus d'un travail collaboratif, je m'efforce de préciser la nature de ma contribution lorsqu'elle n'est que partielle, et au-delà de la contextualisation

de ces travaux et la présentation générale des questionnements scientifiques associés, de développer plus spécifiquement les aspects auxquels j'ai le plus contribué. De plus, au-delà de la liste des coauteurs associés aux publications, je précise le contexte collaboratif dans lequel le travail présenté s'inscrit, en indiquant qui étaient les principaux collaborateurs ainsi que leur affiliation lorsqu'il ne s'agit pas de collègues du Laboratoire de Phonétique et Phonologie.

1.4 Entre expérimentation contrôlée et phonétique de corpus

Bien que ciblant des expressions émotionnelles spontanées, le protocole de recueil de données que j'ai conçu et mis en place pour les besoins de ma thèse s'inscrivait dans une approche fortement expérimentaliste, avec la volonté de combiner la spontanéité des réactions émotionnelles et un fort contrôle expérimental permettant des comparaisons « toutes choses égales par ailleurs ». Si ce principe général s'applique à l'ensemble des sciences expérimentales, son application aux données linguistiques est sujette à interprétation, avec un degré de contrôle expérimental variable en fonction des disciplines et des questions de recherche. En effet, un contrôle expérimental strict implique le plus souvent des productions moins naturelles, ce qui peut conduire à s'éloigner de l'objet d'étude visé, notamment dans le cas de l'étude des expressions d'émotions ou d'autres affects. Classiquement, la phonétique expérimentale, développée entre autres dans le laboratoire dans lequel j'ai réalisé ma thèse et dans celui dans lequel j'exerce depuis 2010, a privilégié un fort degré de contrôle expérimental en raison de la complexité inhérente aux mécanismes de production et de perception de la parole à travers un recours massif à la « parole de laboratoire » ne répondant pas directement à un but communicatif, voire dénuée de sens dans le cas de la production de logatomes. Les travaux auxquels j'ai contribué au début de mon aventure scientifique au Laboratoire de Phonétique et Phonologie se sont largement inscrits dans cette approche expérimentale, indispensable à l'étude fine des mécanismes de production de la parole, tout particulièrement lorsque le recours à des mesures articulatoires est nécessaire pour accéder à la compréhension de ces mécanismes.

Par la suite, j'ai de plus en plus régulièrement été amené à faire évoluer mon approche vers celle de la phonétique de corpus, sans pour autant abandonner les expérimentations contrôlées ou les approches hybrides visant à combiner les données de parole produites dans un contexte écologiques et le contrôle expérimental afin de garantir l'interprétabilité des mesures effectuées. Les méthodes de la phonétique de corpus, qui concernent très majoritairement l'analyse acoustique de données de production en raison de la difficulté d'accès à grande échelle à d'autres types de données, s'appuient fortement sur les méthodes du traitement automatique et renforcent la convergence entre ce champ disciplinaire et les études phonétiques (voir par exemple le programme de l'école d'été CNRS « Big data & speech » dans laquelle plusieurs phonéticiens dont moi-même sommes intervenus au même titre que des collègues spécialistes du traitement automatique de la parole). Ces méthodes permettent à travers un changement d'échelle d'accéder à l'analyse de jeux de données non seulement plus grands, mais aussi plus proches de conditions naturelles de production de la parole. Ainsi, et bien que le contrôle expérimental en laboratoire reste compatible avec des productions plus proches de conditions naturelles sans se réduire à une version caricaturale de la parole de laboratoire, on peut qu'adhérer à ce constat de Mark Liberman :

« *College students reading citation forms in a sound booth yield very different measurements from people of all sorts conversing, discussing, orating, or even reading out loud in real life* » (M. Y. Liberman, 2019)

Au-delà de telles productions éminemment artificielles mais très représentées dans les données à l'origine des connaissances bien établies sur les mécanismes fondamentaux de production et de perception de la parole, la parole continue, en particulier lorsqu'elle est non-lue, peut englober de très nombreuses sous-catégories potentiellement distinctes les unes de autres, en fonction notamment du but communicatif ou de la tâche de production ainsi que des liens interpersonnels entre le locuteur et les interlocuteurs (voir par exemple Warner (2012) pour une réflexion sur les types variés de productions pouvant être qualifiés de « parole spontanée »). Ainsi, quels que soient les efforts déployés pour concevoir et mettre en place des protocoles expérimentaux visant à concilier contrôle expérimental et naturalité, le risque reste élevé que les échantillons recueillis s'éloignent de la parole que produiraient les mêmes locuteurs dans des contextes plus écologiques, par exemple dans la parole conversationnelle produite dans un cadre privé ou en lien avec un rôle sociétal ou professionnel particulier. La phonétique de corpus permet, lorsque de tels enregistrements sont réalisables sans que la seule présence d'un observateur et à plus forte raison de matériel d'enregistrement n'altère de façon importante la nature des données et que leur exploitation est compatible avec les exigences éthiques liées à la protection des données personnelles, d'accéder à l'analyse de telles données. Cet aspect explique en partie l'essor de cette approche dans le domaine de la sociophonétique (Kendall, 2013). De par le volume généralement important de données analysées, cette approche permet également de couvrir plus largement les phénomènes de variation qui touchent la parole que dans un cadre plus contrôlé qui contraint cette variation (voir par exemple les travaux d'Adda-Decker & Snoeren (2011) sur la réduction en parole journalistique et conversationnelle). Ainsi, en schématisant on peut considérer que la phonétique de corpus permet d'étudier la « vraie parole ».

L'utilisation en phonétique de ce type d'approche méthodologique a été rendue possible par les progrès techniques des dernières décennies, avec non seulement une puissance de calcul et une capacité de stockage importantes accessibles hors des unités de recherche et industries spécialisées en informatique, mais aussi des systèmes d'alignement automatique forcé de la parole (voir également en section 7.2.1 mes travaux sur l'évaluation d'un tel système sur la parole pathologique) désormais à la fois performants et accessibles aux chercheurs en phonétique à travers des outils tels que WebMAUS (Schiel, 1999; Kisler et al., 2017) ou Montreal Forced Aligner (McAuliffe et al., 2017). A condition que des transcriptions orthographiques fiables soient disponibles (ce qui en dépit des progrès récents observés également en matière de reconnaissance automatique de la parole implique généralement le recours à des transcriptions ou corrections manuelles), ces deux conditions rendent possible l'application à grande échelle de méthodes relativement classiques d'analyse acoustique de données de parole. L'analyse peut alors se faire à travers l'extraction et la comparaison de mesures de durée, de fréquence fondamentale, de formants ou d'autres mesures spectrales ou cepstrales sur des éléments sélectionnés parmi un grand nombre de tours de parole, d'unités inter-pausales, de mots, de syllabes, de segments ou autre unité jugée pertinente en fonction de leurs propriétés et des questions de recherche. Ainsi mes travaux qui s'inscrivent dans cette approche ont porté sur des dizaines voire des centaines de milliers d'éléments, pour lesquels une annotation manuelle aurait été inenvisageable.

En revanche, l'exploitation de corpus de parole continue non-scriptée implique une grande variabilité des contextes segmentaux, lexicaux, syntaxiques et prosodiques, voire des buts communicatifs et de l'implication des locuteurs. Si certaines de ces informations peuvent être extraites automatiquement sur des jeux de données de grande taille avec un degré de fiabilité

acceptable, cela se fait un prix d'une certaine perte de précision comparativement aux annotations beaucoup plus fines qui peuvent être réalisées manuellement sur un corpus de taille plus modeste. Ainsi, tout aussi performants soient-ils lorsqu'ils sont appliqués à des données de parole comparables aux données ayant servi à entrainer les modèles acoustiques, les systèmes d'alignement forcé ne sont pas exempts d'erreurs, tout particulièrement lorsque la phonétisation associée aux formes orthographiques transcrites ne correspond que partiellement à la prononciation effective de certains mots. De plus, la documentation des variations prosodiques qu'il est possible d'obtenir automatiquement de façon fiable est souvent limitée à un étiquetage en fonction de la position des pauses silencieuses, et les méthodes d'annotation syntaxique automatiques sont parfois prises en défaut sur les données de parole spontanée.

Au-delà de la question de l'annotation des contextes à différents niveaux pour permettre leur prise en compte dans l'analyse, la parole continue non-scriptée se caractérise également par un déséquilibre entre types d'unités et contextes dans lesquels ces unités apparaissent. Ainsi, certains segments sont beaucoup plus fréquents que d'autres dans la parole, et pour un segment donné certains contextes segmentaux sont également beaucoup plus fréquents que d'autres, en partie du fait des fréquences d'occurrences très élevées de certains mots. A l'inverse d'autres combinaisons de segments et de contextes sont beaucoup plus rares, et quand bien même des jeux de données de très grande taille sont utilisés, il n'est pas toujours possible d'extraire un nombre d'occurrences suffisant pour permettre des comparaisons pour l'ensembles des conditions et catégories visées. Bien que les modèles statistiques mixtes permettent de prendre en compte de façon robuste le déséquilibre entre catégories ou classes, ils ne permettent pas toujours de rendre comparables des sous-ensembles de données influencés par des facteurs de variation multiple, ce qui peut limiter la portée des analyses réalisées à partir de ces grands corpus.

En conséquence et bien que je reste convaincu de l'apport essentiel de la phonétique de corpus appliquée à des jeux de données de grande taille pour documenter la parole et son usage dans un contexte plus écologique que ce que permettent les données contrôlées, et que je compte continuer à inscrire une partie de mes travaux dans cette approche, je suis également convaincu que dans de nombreux cas de figure l'utilisation de données de parole plus contrôlée reste indispensable en complément pour valider les observations établies à partir de l'analyse de grands corpus non contrôlés.

1.5 L'usage des statistiques dans mes recherches

1.5.1 Du test de Student à la régression bayésienne…

Comparativement à d'autres domaines scientifiques s'inscrivant de façon plus univoque dans le champ très vaste des sciences expérimentales, le recours aux statistiques en linguistique a été relativement tardif. Bergougnoux (2016) date ainsi les premiers usages en France des statistiques appliquées à la linguistique à la fin des années 1950 avec les travaux de Pierre Guiraud. Leur usage s'est ensuite diffusé progressivement, de façon plus ou moins massive selon les domaines de la linguistique concernés. En sciences de parole, qui se sont toujours situées à la croisée des chemins de plusieurs domaines disciplinaires, cette diffusion a été plus précoce que dans d'autres domaines de la linguistique. L'usage des statistiques inférentielles s'est imposé de longue date, en grande partie sous l'influence de la psychologie cognitive et expérimentale, et est devenu incontournable pour valider les mesures

quantitatives extraites des données et évaluer dans quelle mesure les observations effectuées peuvent être généralisables à plus grande échelle, en inférant à partir de l'échantillon étudié les caractéristiques de la population dont cet échantillon est issu. Ainsi, hormis quelques domaines d'application dans lesquels les spécificités des données ne permettent pas toujours de recourir aux méthodes quantitatives (notamment les études de cas portant sur des pathologies rares en phonétique clinique ou certaines applications didactique), lorsque les recherches effectuées impliquent des comparaisons entre groupes ou entre conditions il est généralement inenvisageable de publier des résultats non-significatifs ou n'ayant pas été validés au préalable par des tests d'hypothèse ou des modèles de régression.

Lors de mes premiers pas dans la recherche, l'approche statistique la plus répandue en sciences de la parole consistait en l'application de tests paramétriques classiques (test t de Student et ANOVA) aux données continues, complétés par des tests non-paramétriques presque aussi classiques en cas de violation de la condition de normalité non résolue par une transformation de données, et de l'application du test du chi-deux aux dénombrements. Les années suivantes ont vu tout d'abord l'essor des ANOVA à mesures répétées, sur lesquelles j'ai suivi pendant ma thèse une formation organisée pour les chercheurs et doctorants du Gipsa-lab et dispensée par des collègues psychologues et que j'ai été conduit à appliquer à mes propres données.

Par la suite, le recours aux modèles avancés de régression (voir par exemple Sonderegger (2023) pour une présentation en profondeur de ces modèles et de leur application à des données linguistiques) et plus particulièrement aux modèles linéaires mixtes s'est rapidement généralisé, aidé en cela par le développement de paquets intégrables à l'environnement R qui permettent l'usage de ces modèles sans connaissances avancées en programmation. Le développement des modèles linéaires mixtes a été suivi dans une moindre mesure par la démocratisation des modèles additifs généralisés (Wood, 2017) GAM (ou GAMM dans le cas de modèles mixtes prenant en compte des facteurs aléatoires) qui permettent de modéliser l'évolution d'une mesure au cours du temps, ou en fonction des fréquences dans le cas de la modélisation de spectres, et s'avèrent donc particulièrement utiles pour l'application à des données de parole. Tout en continuant à utiliser en parallèle des outils simples tels que la corrélation, mes pratiques ont suivi cette évolution, et je me suis formé à la fois en assistant à des formations, comme par exemple celle que mes collègues du Laboratoire de Phonétique et Phonologie et moi avons organisée en invitant Bodo Winter à dispenser une formation intensive à destination des chercheurs et doctorants du laboratoire, et en me formant de façon autonome à travers mes lectures.

En dépit de l'évolution spectaculaire de l'utilisation des statistiques au cours des 15 dernières années avec un recours devenu courant à des modèles avancés, l'usage reste répandu en matière de statistiques d'effectuer des choix binaires à partir de la confrontation à un seuil prédéfini de la probabilité p de rejet à tort de l'hypothèse nulle (erreur de type I) estimée à partir d'un test d'hypothèse, afin de déterminer si un effet est significatif et peut être considéré ou non comme informatif.

Toutefois, ce type d'approche dans laquelle un effet peut basculer d'un statut insignifiant à celui de résultat important suscite beaucoup de critiques, tout particulièrement depuis quelques années. La pratique consistant à fonder les conclusions uniquement sur le caractère significatif ou non des effets étudiés est parfois désignée par le terme, péjoratif dans ce contexte, de « contemplation des étoiles » (*star gazing*) en référence aux étoiles souvent

associées aux effets significatifs dans les figures ou tableaux récapitulatifs. L'utilisation de ces seuls seuils de significativité pour tirer des conclusions d'études dans lesquelles la puissance statistique est trop faible peut être considérée comme l'une des causes de la fameuse « crise de la réplicabilité » qui fait l'objet de nombreuses discussions notamment en psychologie, avec un foisonnement d'études publiées faisant état d'effets significatifs mais que d'autres études reprenant une méthodologie comparable ne parviennent pas à répliquer, et constitue l'une des raisons qui ont conduit récemment divers auteurs à suggérer l'usage de modélisations bayésiennes en remplacement de l'approche plus classique dite « fréquentiste » (voir par exemple Maxwell et al. (2015)).

La régression bayésienne, dont les principes fondateurs sont établis de plus longue date, a été rendue plus accessible de façon récente à la fois via le développement de paquets R permettant aux chercheurs habitués à l'usage des modèles linéaires mixtes une transition relativement en douceur vers l'approche bayésienne, ainsi que du fait de l'augmentation de la puissance de calcul des ordinateurs personnels. En effet cette puissance de calcul rend réaliste l'utilisation d'ordinateurs d'usage courant pour ajuster ces modèles qui reposent sur plusieurs milliers de simulations afin d'obtenir des distributions de probabilité des effets étudiés interprétables de façon fiable sans nécessiter l'utilisations d'infrastructures dédiées au calcul scientifique. Outre les modèles GAM qui reposent en partie sur une modélisation bayésienne, l'application aux données phonétiques de l'analyse bayésienne a notamment été popularisée par Vasishth et al. (2018). Au-delà de leur robustesse, qui a constitué ma motivation première pour me tourner vers l'utilisation de ce type de modèles, l'approche bayésienne présente un certain nombre d'avantages comparativement à l'approche fréquentiste plus classique et notamment une interprétation plus directe et intuitive des résultats qui permet d'éviter les biais d'interprétation des statistiques inférentielles les plus courants (Kruschke & Liddell, 2018). Elle permet également de mieux rendre compte des propriétés des données analysées à travers une vaste gamme de types de distributions au-delà de distributions classiques comme la distribution normale ou la distribution binomiale.

Mon intérêt pour l'analyse bayésienne est récent ne s'est pour l'instant concrétisé que dans une étude publiée, dans laquelle je me suis appuyé sur une modélisation gaussienne des données, ainsi qu'une étude qui a donné lieu à un article actuellement en révision et dans laquelle j'ai eu recours à un modèle de régression beta ordonnée (Kubinec, 2023) afin de modéliser la distribution de mesures de taux de voisement qui se caractérisent par une surreprésentation des valeurs 0 et 1. En parallèle, j'ai continué à utiliser dans la majorité de mes travaux scientifiques des modèles mixtes de régression plus classiques. Toutefois, j'anticipe une extension dans les prochaines années de mon recours aux modèles de régression bayésienne.

Un autre point qui caractérise mon rapport aux statistiques dans mes travaux de recherche est le recours fréquent aux mesures de taille d'effet. C'est dans le cadre de ma thèse que j'ai commencé à prendre en considération ces mesures en complément de l'évaluation de la significativité des effets afin de pouvoir les hiérarchiser. J'ai par la suite eu recours régulièrement à ces mesures de taille d'effet, issues de modèles statistiques classiques tels que l'ANOVA, de modèles linéaires mixtes à travers des mesures de R^2 partiel ou plus récemment des estimations obtenues à partir de régressions bayésiennes, afin de comparer les tailles d'effets de différents facteurs au sein d'une même analyse ou l'effet du même facteur sur différentes variables dépendantes. Dans le cadre d'analyses dans une approche de phonétique de corpus effectuées dans un cadre fréquentiste, j'ai également eu recours aux tailles d'effet

afin de nuancer la significativité observée de ces effets, puisque du fait de l'influence de la taille de l'échantillon sur la probabilité de commettre une erreur de type I, l'analyse de larges jeux de données tend à rendre significatives des différences faibles et peu informatives, parfois même inférieures aux seuils de perception.

1.5.2 … sans négliger statistiques descriptives et figures

Un potentiel effet pervers du développement de modèles statistiques avancés, sur lequel je reviens dans la section 7.4.1 dans laquelle j'expose mes motivations pour le développement d'outils en ligne d'exploration interactive de données, est que la maîtrise de ces modèles et la veille imposée par leur évolution rapide nécessitent un investissement en temps conséquent, susceptible d'induire une relative perte de contact avec la réalité des données de parole qui ne sont plus vues qu'à travers la représentation fournie par les modèles statistiques. Outre le risque d'erreurs d'interprétations que cela implique en cas de choix de modélisations inadaptés à la réalité des données, les modèles statistiques, aussi fins et complexes soient-ils, proposent une représentation des données qui peut être considérée comme idéalisée. Si cette représentation présente l'avantage d'un potentiel de généralisation plus important, indispensable pour inférer des résultats à l'échelle de la population, elle est aussi susceptible de conduire à ignorer des éléments ne suivant pas la tendance majoritaire mais pouvant être très informatifs sur les phénomènes étudiés.

Je reste donc convaincu du rôle essentiel des statistiques descriptives et de l'inspection des données préalablement à toute approche modélisatrice, et mets régulièrement en garde les étudiants contre la tentation de ne considérer leurs données qu'à travers la modélisation qui en est faite, à la fois dans le cadre de mes enseignements en statistiques et en tant qu'encadrant de travaux scientifiques en licence, master et doctorat. Bien qu'une présentation descriptive de la distribution des données ne soit pas toujours incluse dans mes travaux publiés, souvent en raison du format court des articles qui impose des choix plus radicaux de sélection de l'information présentée, leur exploration fait partie intégrante de l'approche d'analyse des données que j'applique de façon systématique.

La représentation graphique des données afin de permettre au lecteur d'appréhender visuellement leurs principales caractéristiques constitue un aspect essentiel de mon approche des analyses quantitatives, et je m'efforce de transmettre dans mon activité d'enseignant et d'encadrant cette culture du graphique, qui à mon sens doit être autant que possible préférée à une présentation sous forme de tableaux et à plus forte raison à une présentation purement textuelle des résultats quantitatifs. J'applique largement ce principe dans mes travaux publiés, ainsi et bien que je n'aie sélectionné qu'une petite partie des figures produites, la majorité de mes travaux présentés dans ce document de synthèse sont accompagnés de figures sélectionnées qui illustrent les principaux effets mis en évidence. Pour autant, j'ai pu constater en me replongeant dans mes travaux moins récents à l'occasion de l'élaboration de ce document de synthèse que mon recours aux représentations graphiques s'est développé progressivement, ce qui m'a conduit à en développer de nouvelles pour les besoins de ce document. Ces représentations graphiques, qui pour la plupart portent sur des variables dépendantes continues ou sur des proportions prennent majoritairement une forme classique de boîtes à moustaches ou de graphe en barres pour la représentation de données univariées, de tracés de la densité de distribution ou de nuages de points pour la représentation de données bivariées, ou encore de courbes pour la représentation de mesures dépendantes du temps ou de la fréquence, ces éléments étant dans certains cas complétés par des

informations permettant de guider l'interprétation des résultats comme par exemple la représentation de seuils ou encore l'utilisation de gradients de couleur pour représenter l'amplitude d'une différence ou la taille des effets.

1.6 Le rôle de l'informatique et du traitement automatique de la parole

Ma formation initiale en ingénierie informatique m'est régulièrement utile pour automatiser et systématiser les différentes étapes du traitement des données de parole que je mets en œuvre dans le cadre de mes recherches, qu'il s'agisse de l'extraction d'informations à partir de l'annotation des corpus de parole, de l'automatisation des analyses acoustiques, de la modification de signaux de parole, du formatage des données ou encore de la mise en œuvre à grande échelle d'analyses statistiques, à travers l'écriture de scripts dont la majorité sont spécifiques à une étude donnée et n'ont pas vocation à être diffusés. Afin de faciliter l'accès à certaines de ces ressources par les étudiants et par les collègues moins à l'aise avec l'algorithmique et la programmation, j'ai été amené à diffuser certains scripts de traitement parmi les plus génériques, et à développer quelques applications en ligne dont certaines sont présentées en section 7.4 de ce document de synthèse.

J'ai également développé à de multiples reprises des interfaces de tests de perception en ligne pour les besoins d'études perceptives, afin d'assurer la collecte d'informations sur les auditeurs, la présentation de stimuli de parole en ordre aléatoire contrôlé, le recueil des réponses et le stockage de ces informations tout en permettant une diffusion plus large, notamment dans le cas de tests s'adressant à des auditeurs non-natifs pour lesquels une passation dans les locaux du laboratoire aurait été plus difficilement envisageable. Du fait de l'essor de la plateforme Psytoolkit (Stoet, 2010, 2017), dont l'usage à des fins académiques est gratuit et qui permet d'intégrer la plupart des paradigmes les plus courants en matière d'évaluation perceptive de stimuli de parole via un langage de scripts simplifié, j'ai depuis quelques années essentiellement recours à cette plateforme et je forme les étudiants à son usage pour des cas d'application simples (à travers notamment un tutoriel, accessible depuis ma page Web). Je n'ai donc plus recours au développement de tests perceptifs en ligne dans leur intégralité ou à l'aide d'outils nécessitant un recours plus important à la programmation tels que Jatos (Lange et al., 2015) que pour la mise en œuvre de paradigmes moins standard, comme par exemple le recueil de transcriptions orthographiques libres pour l'évaluation de l'intelligibilité.

Dans une partie de mes recherches, j'ai également été amené à utiliser plus directement des outils et méthodes de traitement automatique de la parole. Je fais le choix d'employer ici ce terme qui me semble mieux correspondre à mes pratiques, plutôt que celui d'intelligence artificielle dont une définition plus classique englobe ces méthodes mais dont l'acception courante a dévié depuis quelques années avec l'explosion des méthodes d'apprentissage automatique par réseaux de neurones profonds, et plus récemment l'intérêt croissant de la presse et du grand public pour l'intelligence artificielle générative dont l'avatar le plus répandu est GPT-4, plus connu à travers le nom de l'application associée ChatGPT. Du fait de la complexité et surtout de la variabilité extrême du signal de parole, la parole est encore peu concernée par l'intelligence artificielle générative au-delà de certains modèles de synthèse de la parole à partir du texte tels que Tacotron 2 (Shen et al., 2018) et de la combinaison d'outils de génération de texte avec des applications spécifiques à la parole (voir par exemple Yenduri et al. (2024) pour une revue récente des applications de l'intelligence artificielle générative au-delà de la parole). Si les avancées en matière d'intelligence artificielle générative appliquée à

la parole suscitent un intérêt certain de la part d'une partie de la communauté de recherche en traitement automatique de la parole, leur mise en œuvre reste limitée à des sous-domaines très spécifiques comme en témoigne le programme de la conférence Interspeech 2024 avec un nombre limité de sessions consacrées aux modèles génératifs.

Pour autant, l'usage des modèles de réseaux de neurones profonds s'est imposé au cours des dix dernières années, avec un usage qui couvre de nombreux domaines d'application du traitement automatique de la parole (voir Mehrish et al. (2023) pour une revue). Les gains de performance dans les tâches de classification automatique et de reconnaissance automatique de la parole se sont encore accrus avec l'avènement des modèles dits de bout-en-bout qui permettent de modéliser directement le signal acoustique sans nécessiter de sélection préalable de caractéristiques supposées pertinentes ni de prétraitement. Parmi de tels modèles, les plus largement utilisés à l'heure actuelle sont wav2vec2 (Baevski et al., 2020), applicable à des tâches de classification variées telles que la catégorisation automatique des émotions exprimées dans la parole (Sharma, 2022), et Whisper (Radford et al., 2023) dédié principalement à la reconnaissance automatique de la parole mais dont les représentations peuvent être appliquées à diverses tâches de classification de données de parole (voir par exemple Rathod et al. (2023) sur l'application à la détection automatique du degré de sévérité de la dysarthrie).

Si la capacité de tels modèles à apprendre des représentations pertinentes du signal de parole ne fait aucun doute et est à l'origine du niveau de performance atteint par ces modèles et de leur succès pour des applications commerciales ou cliniques, dans la majorité des cas ils ne permettent pas l'interprétation par les chercheurs de ces représentations en dépit d'un effort croissant vers l'interprétabilité des modèles par apprentissage profond (voir par exemple Abderrazek (2023) pour une application à l'intelligibilité de la parole pathologique, dans la communauté francophone on peut également mentionner la journée sur la thématique « Extraction de connaissances interprétables pour l'étude de la communication parlée » co-organisée en décembre 2023 par l'AFCP et l'AFIA-TLH). Ainsi, ces modèles peuvent être précieux pour fournir une référence des capacités de catégorisation ou d'identification de la machine pour la comparaison avec les performances de l'humain (voir par exemple Zellou et al. (2024) sur une tâche de catégorisation phonétique) ou encore avec une catégorisation effectuée à partir de mesures plus classiques et plus directement interprétables (voir par exemple Tirronen et al. (2023) sur l'application à la détection de la voix dysarthrique). En revanche ils ne permettent généralement pas d'extraire plus directement de connaissances sur les caractéristiques communes aux échantillons de parole analysés auxquelles les performances de classification pourraient être attribuées.

Les développements de l'interprétabilité des modèles dans ce domaine en évolution constante m'ont conduit récemment à m'en rapprocher, notamment à travers ma participation au projet ANR EVA (Explicit Voice Attributes) qui vise à établir un lien entre de tels modèles et des descripteurs interprétables de caractéristiques de la voix. A travers le recours à des modèles conçus pour intégrer des descripteurs plus explicites comme par exemple l'approche BA-LR appliquée à la comparaison de voix (Ben Amor & Bonastre, 2022), dans laquelle une voix est caractérisée par la présence ou l'absence d'un ensemble de descripteurs, les objectifs de ce projet convergent avec mes travaux dont une grande partie mettent en œuvre des mesures acoustiques destinées à capturer des corrélats de certaines dimensions de la variation de la voix et de la parole.

Toutefois, la majeure partie des méthodes issues du traitement automatique auxquelles j'ai recours dans certains de mes travaux de recherche sont des méthodes établies de plus longue date, qui pour la plupart visent à extraire des informations à partir des données via leur regroupement automatique en sous-ensembles cohérents ou en identifiant le poids relatif des variables candidates pour décrire un ensemble complexe de données. J'ai ainsi été amené dans certains cas à appliquer des méthodes de partitionnement de données à travers la méthode des k-means ou le regroupement hiérarchique. Pour d'autres études j'ai appliqué des méthodes de sélection de variables, par exemple dérivées de la technique des forêts aléatoires comme la méthode Boruta (Kursa & Rudnicki, 2010).

Par ailleurs, mon recours aux méthodes du traitement automatique se fait dans une large mesure en tant qu'utilisateur, via l'exploitation et dans certains cas l'adaptation d'outils développés par d'autres auteurs pour l'extraction de mesures acoustiques ou la modification du signal de parole. Au-delà de l'utilisation d'outils d'un usage courant en phonétique et notamment de Praat (Boersma, 2001) pour l'annotation de signaux de parole, l'extraction structurée de mesures acoustiques et dans certains cas la modification de la fréquence fondamentale et/ou des durées via l'utilisation de TD-PSOLA (Moulines & Charpentier, 1990), j'ai pu pour les besoins de différents projets m'appuyer sur mon expérience en informatique pour avoir recours à d'autres méthodes de traitement du signal. J'ai par exemple été amené à utiliser des méthodes alternatives d'extraction de la fréquence fondamentale à partir du signal acoustique telles que YIN (de Cheveigné & Kawahara, 2002), ou plus récemment FCN-f0 (Ardaillon & Roebel, 2019) pour l'analyse de la voix pathologique. De façon plus ponctuelle et bien que je privilégie en règle générale les méthodes d'analyse plus paramétrables et interprétables, j'ai également eu recours à OpenSMILE (Eyben et al., 2010) pour l'extraction d'un ensemble large de paramètres à des fins de comparaison avec d'autres méthodes d'analyse du signal de parole.

1.7 Liens entre mes activités d'enseignement et de recherche

Depuis mon recrutement comme Maître de Conférences en sciences phonétiques, j'ai la chance d'enseigner dans un département, l'ILPGA, dédié aux sciences du langage avec une part conséquente de la formation des étudiants consacrée à la phonétique et à la phonologie dès la licence, qui inclut par ailleurs une formation à la recherche via la réalisation d'un projet individuel encadré. De plus, ce département propose plusieurs formations de niveau master dont une formation spécialisée en phonétique et phonologie unique en France, qui a longtemps eu un statut de formation de DEA puis de master à part entière et conserve une grande autonomie depuis sa transformation en parcours du master de sciences du langage. Cette formation, dont j'assure la responsabilité depuis 2019, est fortement orientée vers la recherche en sciences de la parole et une grande partie des doctorants (et même de certains chercheurs confirmés) du Laboratoire de Phonétique et Phonologie en sont issus. Cet environnement est donc particulièrement propice à la convergence entre enseignement et recherche.

Mon activité d'enseignement, détaillée dans mon curriculum Vitæ, couvre principalement les trois grands champs thématiques suivants, certains cours se situant à l'interface de deux voire trois de ces champs :

- La phonétique articulatoire, acoustique et perceptive, avec un accent particulier sur la prosodie, l'acoustique, l'utilisation de l'instrumentation en phonétique expérimentale, et les questionnements spécifiques à l'étude de la parole expressive.
- La méthodologie de la recherche en sciences du langage et plus spécifiquement en sciences de la parole, s'inscrivant majoritairement dans les approches quantitatives. Ces enseignements vont de la formation théorique et pratique aux méthodes expérimentales et à l'analyse statistique en licence, jusqu'à la programmation pour l'automatisation des analyses acoustiques, aux méthodes de traitement du signal appliquées aux sciences de la parole et aux statistiques plus avancées en master, en passant par les méthodes de recueil de données avec un degré plus au moins important de contrôle expérimental à travers le cas des expressions d'émotions et d'attitudes.
- Dans une moindre mesure, les sciences cognitives avec des cours d'introduction aux structures anatomiques du système nerveux humain et leur fonctionnement, ainsi qu'à différents modèles permettant d'expliquer certains aspects de la production et de la perception du langage.

Comme tout enseignant-chercheur, je nourris mes enseignements de connaissances issues de la recherche, parfois simplifiées lorsque le niveau visé l'exige afin de les rendre accessibles. Les éléments sur lesquels je m'appuie sont dans certains cas issus de mes propres recherches, notamment dans le cadre du séminaire dédié à l'étude de la parole expressive dans lequel je présente aux étudiants certains de mes travaux issus de ma thèse et postérieurs à celle-ci sur l'expression dans la parole des émotions et attitudes, et les invite à travers la réalisation d'un dossier à s'initier à la collecte structurée et à l'analyse d'échantillons de parole expressive via le recours à un ancrage théorique et des méthodes, notamment d'analyse perceptive, qui font écho à certaines de mes orientations de recherche.

Dans mes cours à visée plus méthodologique et technique, j'ai recours à certains des concepts fondamentaux et certaines des méthodes qui sont à la base des analyses quantitatives mises en œuvre dans mes travaux de recherche, en m'efforçant de conserver un équilibre entre méthodes directement applicables dans le cadre d'un mémoire et invitation à la réflexion sur les apports mais aussi les limites de ces méthodes. C'est notamment le cas des cours de master les plus spécialisés en analyse de la parole parmi ceux que je dispense. Ainsi dans mon cours dédié à la programmation pour l'automatisation des analyses acoustiques à l'aide de Praat, si une part conséquente du programme abordé porte sur l'initiation aux notions fondamentales de l'algorithmique du fait de l'absence d'expérience préalable de la programmation de la majorité des étudiants concernés, je les forme également à des problématiques plus directement en lien avec mes pratiques de recherche, par exemple à travers l'adaptation du paramétrage des analyses à la nature des données de parole analysées et l'identification des erreurs de détection. Dans le cours consacré à l'application des réseaux de neurones aux données de parole que je partage avec Cédric Gendrot, j'initie les étudiants à diverses méthodes de représentation du signal de parole par un ensemble de valeurs numériques qui correspondent en partie à des méthodes employées dans mes travaux.

Une part importante des notions que j'enseigne dans mes cours de statistiques en licence et master sont plus simples que celles que je suis amené à mettre en œuvre dans mes travaux de recherche, et à quelques exceptions près je privilégie des exemples de données accessibles aux étudiants non-spécialistes de phonétique qui constituent la majeure partie du public de ces cours plus transversaux. Cependant, je m'appuie sur mes pratiques de recherche pour initier les étudiants à la structuration des données quantitatives préalable à l'analyse

statistique et à l'utilisation de fonctionnalités avancées des tableurs et de l'environnement R. Par ailleurs les applications en ligne présentées en section 7.4.2 de ce document de synthèse et qui sont désormais utilisées par des collègues dans le cadre de leurs activités de recherche ont été initialement développées dans le cadre de mes enseignements en statistiques afin de permettre aux étudiants d'appréhender plus directement les concepts fondamentaux relatifs à la distribution des données à partir de celles-ci en dépassant les difficultés purement techniques.

Dans le cadre de mes autres enseignements, outre le recours ponctuel à mes propres travaux à titre d'exemple pour rendre plus concret pour les étudiants la conception et la mise en œuvre d'un travail de recherche en linguistique, ainsi que la veille scientifique par le biais de lectures qui fait partie intégrante de mon activité de recherche, j'ai été à plusieurs reprises amené à intégrer des éléments issus de séminaires ou autres présentations scientifiques auxquelles j'ai assisté dans le cadre de conférences et colloques. Ainsi, pour la mise en place du cours de psycholinguistique en troisième année de licence lors de l'année universitaire 2014/2015, je me suis inspiré du cycle de conférences donné par Ellen Bialystok lors de sa venue en tant que professeure invitée par le LabEx EFL afin d'intégrer une séance de cours sur les liens entre bilinguisme et contrôle exécutif.

Par ailleurs en tant qu'encadrant de projets de recherche en troisième année de licence et en master, je suis régulièrement amené à proposer des sujets en lien avec mes activités de recherche ou s'inscrivant dans un champ disciplinaire et méthodologique proche de celles-ci. Bien entendu, les exigences de la recherche en sciences de la parole sont telles que seule une minorité de ces travaux d'étudiants donne lieu à des résultats exploitables, qui dans certains cas peuvent être retravaillés pour donner lieu à des publications. Aussi chronophage soit-elle, je considère néanmoins cette activité d'encadrement comme une composante essentielle de la formation par la recherche, nécessaire pour initier aux pratiques de recherche les étudiants envisageant une poursuite en doctorat et au-delà, mais aussi de façon plus générale pour apporter aux étudiants en linguistique se destinant à d'autres carrières une compréhension plus concrète des méthodes et enjeux de la recherche et ainsi une réflexion plus approfondie sur les objets d'étude de la linguistique.

Au-delà de mon activité statutaire d'enseignant en licence et master à l'université Sorbonne Nouvelle, j'ai également été amené à proposer des enseignements en lien avec mes recherches destinés aux doctorants et collègues. Ainsi, je suis intervenu dans une journée de formation en statistiques pour linguistes organisée en 2014 par l'école doctorale en sciences du langage de la Sorbonne Nouvelle (à l'époque ED268). Par la suite, j'ai été l'un des trois intervenants lors d'une journée de formation en statistiques destinée aux doctorants et chercheurs en phonétique clinique, organisée en 2017 par l'AFCP en marge de la conférence bisannuelle Journées de Phonétique Clinique. J'ai également proposé à deux reprises en 2016 et 2018 une formation à la programmation pour l'analyse acoustique semi-automatisée de données de parole et sa mise en œuvre avec l'outil Praat, à destination des membres du LabEx EFL. En outre, je suis intervenu dans l'école d'été CNRS « Big data & speech » en 2018, en concevant et animant un atelier dédié à l'extraction, l'analyse et la visualisation de mesures de formants à partir de grands corpus de parole continue en complément d'un cours assuré par mon collègue Cédric Gendrot.

1.8 La variation inter- et intra-locuteur comme fil conducteur

Bien que les questions liées à la variabilité entre locuteurs et au sein des productions d'un même locuteur aient suscité mon intérêt de longue date et constituent l'objet d'étude d'une partie de mes travaux que je présente dans le chapitre 2 de ce document de synthèse, dans une grande partie de ceux-ci j'ai cherché de façon plus classique à caractériser la voix et la parole de groupes de locuteurs supposés partager des caractéristiques communes, ou de modifications induites chez les mêmes locuteurs par un changement de condition de production. Dans ces études qui ont porté majoritairement sur l'analyse acoustique de données de production, les spécificités individuelles du locuteur sont considérées comme un facteur de variation ne pouvant être que très partiellement contrôlé en équilibrant autant que possible les groupes en fonction de variables telles que l'âge, le sexe ou encore le niveau de langue, et dans les approches statistiques reposant sur des modèles mixtes en considérant le locuteur comme un facteur de variation aléatoire afin de compenser les spécificités individuelles pour mieux accéder aux caractéristiques communes à l'échelle du groupe.

Sans remettre en question la validité d'une telle approche, mes travaux récents sur la variation entre locuteurs et au sein des productions d'un même locuteur m'ont amené à reconsidérer en partie certains de ces travaux et à m'interroger non seulement sur l'ampleur de la variation entre locuteurs au sein d'un même groupe et le lien avec les différences observées entre groupes comparés, qui sont prises en compte dans les modèles statistiques couramment utilisés mais tendent à être aussi en partie masquées par cette approche, mais aussi sur les informations plus fines que des analyses réalisées au niveau du locuteur considéré individuellement pourraient révéler.

En complément de la présentation de mes travaux scientifiques postérieurs à ma thèse et d'une réflexion sur les résultats obtenus, la rédaction de ce document de synthèse me fournit également une occasion d'approfondir une partie de ces travaux à travers de nouvelles analyses effectuées à partir des mêmes données que celles ayant donné lieu aux publications d'origine, mais centrées sur la variation individuelle. Ces nouvelles analyses, majoritairement descriptives en raison de la quantité de données disponible par locuteur, et recentrées dans certains cas sur le sous-ensemble des locuteurs les plus largement représentés dans les données, sont illustrées par des représentations graphiques et discutées à la suite de la présentation des travaux d'origine.

Du fait de la nature des données analysées, ces analyses complémentaires centrées sur le locuteur ont porté essentiellement sur la caractérisation des différences entre locuteurs, les données d'un même locuteur étant dans la majorité des cas issues d'une même session d'enregistrement, d'où une caractérisation possible de la variation intra-locuteur plus limitée voire impossible. Elles concernent essentiellement les travaux auxquels j'ai participé sur la variation vocalique, toutefois je serai également amené à revenir sur la question de la variation individuelle en complément de la présentation de travaux abordés dans d'autres chapitres de ce document de synthèse, en particulier le chapitre 6 qui regroupe mes travaux ne correspondant à aucun des autres chapitres thématiques.

1.9 Organisation de ce document de synthèse

Mes travaux scientifiques sur lesquels je reviens dans ce document de synthèse sont structurés en six chapitres thématiques numérotés de 2 à 7, chacun étant précédé d'un bref résumé encadré afin de faciliter la lecture du document dans son ensemble. Dans le chapitre

2 je reviens sur mes travaux dans lesquels la variation entre locuteurs et au sein des productions d'un même locuteur constituait directement l'objet d'étude, dans une optique orientée vers les applications technologiques ou criminalistiques ou avec une approche plus phonétique. Dans le chapitre 3 j'expose le cheminement scientifique qui a été le mien après la thèse en matière d'étude des affects, notamment à travers l'étude d'expressions d'attitudes en allant des corpus contrôlés vers des données plus naturelles. Dans le chapitre 4 je présente les travaux variés auxquels j'ai participé autour de la variation de la réalisation des voyelles, considérées dans leur ensemble en tant que système ou en ciblant des catégories vocaliques plus spécifiques. Le chapitre 5 est consacré à la présentation de mes travaux en phonétique clinique sur diverses pathologies de la voix et de la parole, ainsi que sur la thématique du vieillissement dans la voix et la parole dont l'étude est indispensable à une meilleure compréhension des pathologies de la voix et de la parole susceptibles d'affecter les sujets âgés. Dans le chapitre 6, je reviens sur l'autres travaux auxquels j'ai contribué et qui ne s'inscrivent pas dans l'une des autres thématiques, notamment ceux sur l'acquisition, sur le dimorphisme ainsi que d'autres travaux en phonétique de corpus. Enfin le chapitre 7 est consacré à mes principales contributions méthodologiques à travers le développement ou l'évaluation de méthodes et d'outils ainsi que le recueil de données.

Du fait du poids variable de ces thématiques dans ma carrière scientifique post-thèse, ces sept chapitres sont de volume inégal. Le degré d'homogénéité des travaux que j'ai choisi de regrouper au sein de chacun de ces chapitres est également variable, et pour cette raison il serait dans certains cas artificiel de présenter un cadre théorique commun à des études parfois très diverses du point de vue des objectifs scientifiques et du contexte dans lequel elles s'inscrivent. Ainsi, si j'ai cherché dans chacun des chapitres à contextualiser mes recherches qui y sont présentées, cette contextualisation est souvent présentée en plusieurs étapes, du général au particulier. En outre, le choix de catégoriser certains de mes travaux comme relevant d'une thématique plutôt que d'une autre comporte une part d'arbitraire. Ainsi, certaines études que j'ai fait le choix de présenter dans le chapitre 5 dédié à la phonétique clinique et au vieillissement pourraient être considérées comme relevant du chapitre 4 sur la variation vocalique. Il en va de même de certains des travaux que je présente dans le chapitre 2 dédié à la variation inter- et intra-locuteur. Par ailleurs, une part conséquente de mes contributions méthodologiques, que j'ai fait le choix de rassembler dans le chapitre 7, ont été développées en lien avec des questions scientifiques traitées par ailleurs et je suis donc amené régulièrement à faire référence dans le texte de ce chapitre à d'autres sections, et inversement.

J'ai fait le choix de ne revenir dans ce document de synthèse que sur mes travaux ayant donné lieu à des publications évaluées par les pairs, ou au minimum à des communications dans des colloques ou conférences suite à un processus d'évaluation anonyme sur résumé, toutefois la présentation que je propose de ces travaux ne se limite pas au contenu de ces publications et dans de nombreux cas je présente conjointement plusieurs publications qui se rapportent à une thématique commune.

Dans les chapitres 2 à 7, je précise au début des sections correspondantes les publications et communications associées aux travaux présentés sous forme d'un encadré sur fond grisé, en reprenant la numérotation utilisée dans mon Curriculum Vitæ ainsi que dans le volume de travaux publiés (Volume 2 de ce dossier d'habilitation), qui combine le type de publication et une numérotation en ordre chronologique inversé. La liste de mes publications n'étant pas reprise dans la bibliographie listée en fin de volume afin d'éviter les redondances, je reprends également cette numérotation lorsque ces publications sont citées dans le texte, notamment

dans le cas de figures reprises de mes publications ou générées pour les besoins de ce document de synthèse à partir des données présentées dans les publications. En l'absence d'indication explicite d'une référence bibliographique comme source des figures présentées dans le texte, ces figures sont issues de nouvelles analyses réalisées pour les besoins de ce document de synthèse. Lors de l'insertion de ces figures je me suis efforcé d'appliquer le principe selon lequel la légende d'une figure doit permettre de l'interpréter sans nécessairement avec pris connaissance de l'ensemble du texte.

2 Variation inter- et intra-locuteur

Résumé du chapitre 2

Dans ce chapitre, je commence par introduire les enjeux de l'étude de ces deux grands types de variation en phonétique et en criminalistique, ainsi que les questions liées à la variabilité des différentes classes de segments phonémiques et à la capacité d'auditeurs à identifier perceptivement une voix, avant de présenter mes travaux sur la variation inter- et intra-locuteur.

Une partie de ces travaux se sont appuyés sur l'analyse des performances de systèmes automatiques de vérification du locuteur en fonction de la nature des données utilisées par ces systèmes, en les croisant avec des analyses phonétiques d'une partie de ces données. Cette série d'études a montré que bien que certains segments soient porteurs de plus d'informations idiosyncrasiques que d'autres, les différences de contenu segmental ne suffisent pas à expliquer les variations de performances observées entre locuteurs et entre extraits produits par un même locuteur. La confrontation des performances en vérification du locuteur d'un système automatique à celles d'auditeurs a montré que si les performances du système automatique sont globalement supérieures, les extraits ambigus pour ce système sont mieux reconnus par les auditeurs humains.

Dans une série d'études menées à partir d'enregistrements acoustiques de parole lue produites par les mêmes locuteurs francophones lors de sessions multiples plus ou moins espacées dans le temps, la variabilité inter-locuteurs de mesures de voix et de parole a été comparée à la variabilité intra-locuteur, en considérant notamment des unités d'environ dix syllabes et la modulation entre unités consécutives. Ces comparaisons ont montré que les mesures qui distinguent le mieux les locuteurs les uns des autres, en premier lieu liées à la voix plutôt qu'à l'articulation segmentale, sont aussi les plus discriminantes entre sessions, mais que les patrons individuels de variabilité sont relativement stables entre unités consécutives.

La comparaison entre locuteurs de la coarticulation labiale anticipatoire dans la production de la consonne /s/ a mis en évidence une variation inter-individuelle en termes à la fois de degré de coarticulation, d'empan temporel de cette coarticulation et de stabilité entre sessions d'enregistrement. Enfin, l'analyse comparée des réalisations vocaliques individuelles en parole lue et spontanée a montré une différence entre locuteurs plus importante en parole lue, et portée plus largement par les voyelles nasales que par les voyelles orales avec une influence moindre du style de parole pour les nasales.

2.1 Enjeux théoriques et applicatifs

2.1.1 Variation inter- et intra-locuteur et phonétique

Pour l'essentiel, les connaissances sur la parole se sont construites à partir de la comparaison entre données produites par différents locuteurs ou groupes de locuteurs ou des donnée moyennées entre locuteurs. Par conséquent, nous savons relativement peu de choses sur la variabilité de la parole au sein d'un même individu. Pourtant, nous savons que les locuteurs s'adaptent aux contextes linguistiques et à la situation de communication et modifient leurs productions en fonction de facteurs tels que leur état émotionnel ou biologique. Cependant, on ne sait pas exactement dans quelle mesure les locuteurs varient d'une session d'enregistrement à l'autre, sur quelles caractéristiques de la parole, et dans

quelle mesure les enregistrements recueillis à des jours, des mois ou des années d'intervalle sont variables au-delà des facteurs identifiables, et pour certains quantifiables. Si la variabilité de la parole est omniprésente, elle est aussi, du moins en partie, structurée et contrainte. Par conséquent, la documentation de la variation de la parole et l'explication de l'origine de cette variation revêtent un caractère fondamental pour la recherche en phonétique.

Deux types de variabilité peuvent être distinguées : la variabilité entre locuteurs et la variabilité au sein des productions d'un même locuteur (Wright, 2006). La variabilité inter-locuteurs, porteuse d'informations indexicales à travers un faisceau de caractéristiques de voix et de parole, relève à la fois de spécificités individuelles au niveau physiologique et anatomique, et de caractéristiques régionales et sociales qui vont contribuer aux différences entre groupes de locuteurs (Foulkes & Docherty, 2006).

La variabilité intra-locuteur provient quant à elle de sources multiples, dont Bürki (2018) a proposé une revue. Outre les effets relativement bien documentés du contexte phonémique ou prosodique sur la réalisation des segments, une part plus difficile à définir et modéliser de cette variabilité pour un même locuteur est liée au style de parole, notion elle-même très vaste qui englobe l'adaptation du locuteur à la fois au contexte communicatif, aux caractéristiques de l'interlocuteur ou encore aux relations interpersonnelles (voir section 4.2 pour un développement sur les liens entre style de parole et réalisation des voyelles). La variabilité intra-locuteur dépend également de l'état affectif du locuteur au moment de la production de parole (voir par exemple Barett & Paus (2002)), ou encore de son état de santé ou de fatigue, susceptible non seulement de varier d'un moment à l'autre mais aussi d'évoluer entre le début et la fin d'un enregistrement (Gelfer et al., 1991).

Du fait de la difficulté à caractériser certains de ces facteurs de variation et surtout leur interaction faute de critères clairs pour délimiter ce qui relève du style de parole ou de l'état du locuteur, la variabilité intra-locuteur est couramment considérée comme un facteur de variabilité aléatoire. On considère classiquement en phonétique qu'un énoncé voire un son ne sera jamais prononcé deux fois de façon rigoureusement identique, mais l'ampleur de cette variation et les facteurs qui la conditionnent sont relativement peu documentés. On sait que l'incidence du délai entre les enregistrements est non négligeable (voir par exemple Campbell et al. (2009) sur l'impact considérable que peut avoir ce délai sur les performances d'un système automatique de vérification du locuteur), mais que l'ampleur de cette variation ne dépend pas exclusivement de ce facteur : des voyelles isolées produites le même jour à des moments différents de la journée peuvent ainsi être plus différentes entre elles que celles produites au même moment des jours différents (Heald & Nusbaum, 2015). Par ailleurs, la multiplicité des sources de variation intra-locuteur et les fluctuations de l'amplitude de la variabilité qu'elles sont susceptibles d'induire dans les productions d'un même locuteur rendent incertaine la quantification de la variation inter- et intra-locuteur, essentielle pour l'estimation de l'importance des distinctions entre locuteurs ou groupes de locuteurs. Une meilleure documentation de la variation intra-locuteur s'avère donc essentielle pour mieux comprendre la variation inter-locuteurs.

2.1.2 Identification du locuteur et criminalistique

La question des différences entre locuteurs et entre productions d'un même locuteur revêt une importance particulière dans le domaine criminalistique, dans lequel des échantillons de voix doivent être comparés afin de déterminer s'ils ont ou non été produits par la même

personne. Dans un cadre de police scientifique, il s'agit à la demande d'une autorité judiciaire de confronter une « pièce de question », qui consiste en un enregistrement produit par un locuteur dont l'identité est incertaine, à une « pièce de comparaison » qui consiste en un enregistrement attribuable de façon certaine à un locuteur connu. Outre les conditions d'enregistrement des pièces de question souvent très éloignées des conditions optimales classiquement associées aux enregistrements acoustiques destinés à l'analyse phonétique, de nombreux facteurs de variation sont susceptibles de renforcer les différences entre pièce de question et pièce de comparaison et donc de compliquer cette tâche : délai parfois important entre les enregistrements, situation de communication et donc styles de paroles très différents, voire manque de coopération des locuteurs au moment de la production de la pièce de comparaison.

Bien que la demande judiciaire soit forte, la variabilité intra-locuteur inhérente à la parole remet en question la possibilité d'utiliser la voix comme un marqueur biométrique au même titre que l'ADN ou que les traces papillaires (plus couramment appelées empreintes digitales). Ce constat a été à l'origine d'une motion votée en 1990 par la communauté scientifique francophone de recherche sur la parole à travers l'association savante AFCP (Association Francophone de la Communication Parlée, à l'époque GFCP), considérant que l'identification du locuteur à partir de sa voix était un problème non-résolu. Cette motion a été suivie en 1997 d'un moratoire sur la participation de la communauté scientifique à des expertises vocales, ses membres n'intervenant qu'au titre de témoins scientifiques ou « sachants » afin de présenter dans le cadre judiciaire l'état des connaissances scientifiques sur l'identification du locuteur et les conclusions pouvant être tirées ou non des expertises, mais en aucun cas en tant qu'experts judiciaires. Ce moratoire a toutefois eu un effet pervers et a laissé le champ libre à des pratiques relevant du charlatanisme (voir Bonastre (2020) pour un retour sur l'historique des comparaisons de voix judiciaires en France), conduisant des membres de l'AFCP à devoir argumenter devant les tribunaux pour faire valoir le caractère irrecevable voire frauduleux de certaines expertises.

Bien que le moratoire de l'AFCP reste d'actualité dans la sphère francophone, la position de la communauté scientifique sur l'identification d'un locuteur à partir de sa voix a peu à peu évolué, du fait des progrès observés en matière d'identification du locuteur et du resserrement des liens entre chercheurs et membres de la police technique et scientifique (PTS) qui est à l'origine du projet ANR VoxCrim réunissant laboratoires de recherche et services de PTS auquel j'ai participé au même titre que d'autres membres du Laboratoire de Phonétique et Phonologie. Néanmoins, les acteurs scientifiques du domaine s'accordent à considérer que la comparaison de voix appliquée à la criminalistique nécessite une grande prudence (Campbell et al., 2009), et que les comparaisons de voix ne peuvent pas donner lieu à une réponse tranchée, autrement dit qu'il est impossible de répondre de façon certaine par oui ou non à la question de savoir si deux enregistrements ont été produits par la même personne, mais uniquement à des probabilités exprimées sous forme de rapports de vraisemblance. Parmi les défis majeurs posés à la communauté scientifique auxquels le projet VoxCrim a tenté d'apporter des éléments de réponse, on peut identifier notamment l'effort nécessaire d'interprétabilité et de reproductibilité des critères appliqués en matière de comparaisons de voix, et d'explication des limites de ces comparaisons auprès des acteurs du monde judiciaire.

2.1.3 Quels segments et dimensions sont les plus variables entre locuteurs ?

Une question connexe dont les implications concernent à la fois les recherches en phonétique sur la variation inter-individuelle et les applications à la comparaison de voix est celle de la variabilité comparée des différentes unités segmentales. Cette question a été principalement abordée à travers le prisme de la discrimination entre locuteurs. En japonais, Amino et al. (2006) ont montré une plus importante discrimination entre locuteurs à partir de syllabes incluant une consonne nasale qu'à partir de syllabes incluant une consonne orale, attribuée aux variations individuelles de morphologie de la cavité nasale et de l'impossibilité de moduler cette cavité par des mouvements articulatoire au-delà du degré d'abaissement du vélum. Ainsi en anglais britannique, de Jong et al. (2007) ont montré à partir d'une analyse des fréquences formantiques que la voyelle /ɔː/, soumise à un changement diachronique et pour cette raison supposée plus variable entre locuteurs que d'autres voyelles plus stables, permettait effectivement mieux de discriminer entre locuteurs. Plus récemment en français, Ajili et al. (2016, 2017) ont conclu à partir de l'analyse comparative des performances d'un système automatique de vérification du locuteur appliqué à différentes classes phonémiques que sur leurs données, la discrimination était la meilleure sur les voyelles orales, suivies des voyelles nasales, les consonnes occlusives ayant le pouvoir discriminant le plus faible.

Au-delà de la comparaison entre phones ou classes segmentales, une autre question est celle des dimensions de la voix et la parole les plus dépendantes du locuteur parmi toutes celles susceptibles de varier. Du fait de la paramétrisation du signal acoustique qui englobe l'ensemble de ces dimensions de façon conjointe, que cette paramétrisation soit effectuée préalablement à l'entrainement des systèmes automatiques via par exemple le calcul de coefficients cepstraux ou directement par des modèles de bout en bout, les approches qui se fondent sur l'utilisation de systèmes automatiques de reconnaissance du locuteur ne permettent pas de comparer directement le poids relatif de ces dimensions dans la discrimination entre locuteurs. Cependant quelques études ont proposé de telles analyses à partir de méthodes plus classiques d'analyse phonétique. Ainsi, McDougall (2006) a mis en évidence une grande variabilité inter-locuteurs dans les transitions formantiques. Par ailleurs les analyses menées par van Dommelen (1987) et par Nolan (2001) ont suggéré que si la forme des contours et le registre de fréquence fondamentale ne suffisent pas à discriminer entre locuteurs, ces paramètres peuvent contribuer à affiner les performances de l'identification du locuteur.

2.1.4 Peut-on identifier perceptivement un locuteur ?

Une autre question qui fait écho à celle de l'analyse de la variabilité inter- et intra-locuteur à partir de données de production est celle de la capacité de l'humain à identifier perceptivement un échantillon de voix comme étant produit par une personne particulière. Cette capacité tend à être surévaluée par les sujets naïfs qui font régulièrement l'expérience de l'identification d'une personne connue à partir de sa voix. Outre l'exploitation probable d'indices contextuels, la capacité d'auditeurs à identifier dans un protocole contrôlé les voix de personnes inconnues est en effet beaucoup plus faible que les performances mesurées dans le cas de personnes familières (Hollien et al., 1974; Van Lancker et al., 1985).

Parmi les autres facteurs susceptibles d'expliquer la variabilité des performances d'auditeurs dans une tâche d'identification de voix, on peut également mentionner la longueur des énoncés à identifier. Blatchford & Foulkes (2006) ont ainsi mis en évidence des

performances sensiblement inférieures sur des énoncés de deux syllabes que sur des énoncés plus longs. En revanche le contenu lexical ne semble pas avoir un effet notable sur les performances de discrimination (Yarmey, 2001). Les résultats de la littérature relatifs à l'effet du niveau d'expertise des auditeurs sont contrastés, certaines études concluant à de meilleures performances d'auditeurs expérimentés (N. O. Schiller & Köster, 1998) tandis que d'autres ne mettent pas en évidence de différences entre groupes (Reich & Duke, 1979).

2.2 Identification du locuteur par l'humain et la machine

2.2.1 Variabilité phonétique et vérification automatique du locuteur

Publications et communications associées :

[ACTI48] Kahn, J., **Audibert, N.**, Rossato, S., & Bonastre, J.F. (2011). Inter and intra-speaker variability in French: an analysis of oral vowels and its implication for automatic speaker verification. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong-Kong, Chine, pp. 1002-1005.

[ACTI52] Kahn, J., **Audibert, N.**, Rossato, S., & Bonastre, J.F. (2010). Intra-speaker variability effects on Speaker Verification performance. *Proceedings of the 7th International Workshop on Speaker and Language Recognition (Odyssey 2010)*, Brno, République Tchèque, actes CD-ROM.

[ACTI57] Kahn, J., **Audibert, N.**, Rossato, S., & Bonastre, J.F. (2010). Modéliser un locuteur : Influence des signaux d'apprentissage sur les performances d'un système de RAL. *Actes des 18èmes Journées d'Études sur la Parole (JEP 2010)*, Mons, Belgique, pp. 109-112.

En collaboration avec Juliette Kahn dans le cadre de sa thèse (Kahn, 2011) réalisée au Laboratoire d'Informatique d'Avignon sous la direction de Jean-François Bonastre et Solange Rossato, j'ai contribué à l'étude de la variabilité inter et intra-locuteur dans de grands corpus de parole pas ou peu contrôlés dans une démarche à l'interface de la phonétique et du traitement automatique de la parole. L'objectif principal des travaux auxquels j'ai participé, principalement à travers le prétraitement des données pour obtenir une segmentation exploitable, l'extraction de mesures acoustiques et l'analyse statistique des résultats, était de mieux comprendre l'incidence de cette variabilité sur les performances de systèmes de vérification du locuteur dans lesquels la machine doit déterminer à partir d'un échantillon de parole inconnu si cet échantillon a été produit ou non par un locuteur cible, le contenu segmental des données de parole utilisées pour l'entraînement et l'évaluation des systèmes n'étant pas contrôlées. Pour cela, les performances d'un système de vérification du locuteur ont été évaluées sur un ensemble d'échantillons de parole produits ou non par le même locuteur, les caractéristiques phonétiques de ces échantillons étant ensuite confrontées aux performances du système afin d'évaluer dans quelle mesure elles permettent d'expliquer les variations de performances.

Le système d'identification utilisé dans ces travaux était ALIZE/SpkDet (Bonastre et al., 2008), fondé sur un principe de modélisation de la voix et de la parole d'un locuteur par mélange de gaussiennes (UBM/GMM) à partir d'une paramétrisation cepstrale du signal, et dont les performances dans les campagnes internationales d'évaluation mises en place par l'organisme américain NIST étaient proches de l'état de l'art jusqu'à l'avènement des approches par apprentissage profond. Les performances d'un tel système, et plus généralement d'un système biométrique, sont évaluées à partir d'un ensemble de comparaisons dites « cible »

dans laquelle l'échantillon à vérifier a bien été produit par le locuteur cible, et d'un ensemble de comparaisons dites « imposteur » dans laquelle l'échantillon à vérifier a été produit par un locuteur qui n'est pas le locuteur cible. Les comparaisons « cible » permettent de calculer un taux de faux rejet correspondant aux cas dans lesquels l'échantillon est considéré à tort comme n'ayant pas été produit par le locuteur cible, et les comparaisons « imposteur » un taux de fausse acceptation qui correspond aux cas dans lequel l'échantillon est considéré à tort comme ayant été produit par le locuteur cible. En pratique, pour chaque comparaison effectuée le système attribue un score, qui n'est pas nécessairement interprétable directement comme une probabilité mais qui peut être considéré comme un degré de proximité entre l'échantillon de parole à vérifier et la modélisation du locuteur cible par le système, et la décision est prise à partir d'un seuil : pour une comparaison donnée, si le score attribué par le système atteint ce seuil, l'échantillon est considéré comme ayant été produit par le locuteur cible. En conséquence, la définition du seuil a une incidence directe sur les taux de faux rejet et de fausse acceptation, puisqu'un seuil plus élevé tend à faire baisser le taux de fausse acceptation mais à faire augmenter en contrepartie le taux de faux rejet, et inversement. Plutôt que de se contenter d'un seuil prédéfini, les performances des systèmes biométriques sont évaluées à partir de la distribution des scores attribués pour chaque comparaison, via des courbes DET (*Detection Error Tradeoff*) dans l'espace à deux dimensions du taux de fausse acceptation et du taux de faux rejet qui représentent l'évolution du compromis entre les deux types d'erreur en fonction du seuil fixé. Diverses métriques peuvent être dérivées de ces courbes, l'une d'elles étant le taux d'égale erreur (*Equal Error Rate*, ci-après EER) qui correspond au point de la courbe DET pour lequel le taux de fausse acceptation est égale au taux de faux rejet et que nous avons retenu comme métrique pour l'évaluation des performances de la vérification automatique du locuteur. Si elle ne peut être calculée qu'a posteriori et ne correspond donc pas à un usage en conditions réelles d'un système biométrique, pour un système et un ensemble de données on peut donc considérer l'EER comme le meilleur compromis possible entre un système trop strict (taux de faux rejet élevé) et un système trop permissif (taux de fausse acceptation élevé).

Pour les évaluations effectuées, les conversations téléphoniques en anglais d'une durée de 2,5 minutes issues de la campagne d'évaluation NIST-SRE 2008 (A. F. Martin & Greenberg, 2009) ont été utilisées avec une sélection de 171 locuteurs masculins représentés par un total de 816 modèles, ainsi que les données de parole lue en français du corpus BREF 120 (Lamel et al., 1991) qui comprend 64 femmes et 43 hommes. Pour chaque locuteur, 39 extraits d'au moins 30 secondes de parole utile après élimination des portions détectées automatiquement comme silence ou bruit ont été sélectionnées dans le corpus BREF 120 afin de faire varier les signaux d'apprentissage et de test pour un même locuteur dans des proportions comparables. Dans ces deux corpus, les performances du système de vérification ont été évaluées en fonction des extraits choisis pour représenter chaque locuteur. Une très importante fluctuation des performances a été relevée en fonction des extraits sélectionnés, avec pour les données de NIST-SRE 2008 une EER moyenne optimale (obtenue en sélectionnant le modèle le plus performant pour chaque locuteur) de 4,1% tandis le choix des modèles les moins performants aboutissait à une EER de 21,9%. L'application de la même méthodologie aux données lues du corpus BREF 120 ont abouti à une EER encore plus variable en fonction des extraits sélectionnés que sur les données de NIST-SRE 2008, allant de 1,0% à 33,0% pour les hommes, et de 1,1% à 28,5% pour les femmes. A titre de comparaison, une sélection aléatoire a également été prise en compte sur les deux jeux de données, ainsi que la sélection par défaut

proposée dans l'évaluation NIST-SRE 2008 et des extraits de 2,5 minutes pour les données de BREF 120, aboutissant à un niveau de performance intermédiaire proche de l'état de l'art.

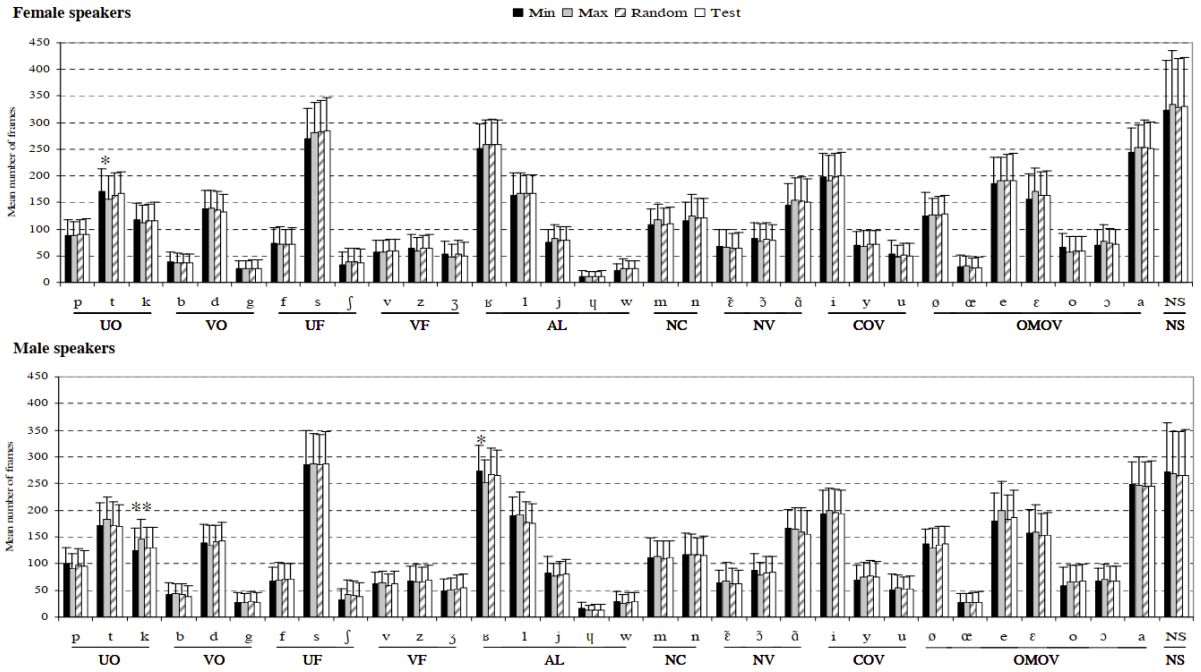


Figure 1 : Distribution de l'information segmentale dans les quatre sous-ensemble de données issus du corpus BREF 120 (Lamel et al., 1991) évalués pour les femmes (en haut) et les hommes (en bas). Les jeux de données *Min* et *Max* correspondent respectivement à la meilleure performance (EER minimale) et à la pire performance à partir d'échantillons de 30 secondes, *Random* à une sélection aléatoire de ces échantillons et *Test* aux échantillons de 2,5 minutes. Les barres d'erreur représentent l'écart-type, les étoiles au-dessus des barres indiquent les cas dans lesquels le nombre de trames incluses dans le jeu de données associé à la meilleure performance est significativement différent du nombre de trames incluses dans le jeu de données associé à la pire performance (* : $p < .05$; ** : $p < .01$). Les segments sont regroupés par classe phonémique, la catégorie NS correspondant aux trames considérées par le système comme n'étant pas des sons de parole. D'après Kahn et al. (2010, [ACTI52]).

En raison du niveau de bruit plus élevé et surtout de l'absence de transcription des données de NIST-SRE 2008, la transcription par un autre moyen qu'un système de reconnaissance automatique de la parole étant contraire aux conditions d'utilisation des enregistrements mis à disposition par l'organisme NIST dans le cadre de ces campagnes d'évaluation, seules les données du corpus BREF 120 ont fait l'objet d'une analyse plus approfondie afin de tenter d'identifier les liens entre contenu phonétique et performance des systèmes au-delà de ce constat d'une variabilité extrêmement importante à la fois sur des données de parole conversationnelle et de parole lue. Une première analyse a consisté en la comparaison des distributions segmentales entre les jeux de données aboutissant à différents niveaux de performance, comptabilisées en nombre de trames de durée fixe incluses dans ces jeux de données. En effet, certaines études comme celle d'Amino et al. (2006) sur les sons nasals ont mis en évidence une plus grande variabilité inter-individuelle dans la réalisation de certains segments. Comme illustré par la Figure 1, cette analyse a révélé des distributions très similaires entre jeux de données, conformes aux fréquences d'occurrences segmentales décrites dans la littérature pour le français (Tranel, 1987), avec toutefois des fluctuations importantes entre locuteurs. Les rares différences relevées entre le jeu de données *Min* associé aux meilleures

performances et le jeu de données *Max* associé aux pires performances consistaient pour les femmes en une représentation de /t/ significativement plus importante dans le jeu de données *Min*, et pour les hommes en une représentation de /k/ significativement plus faible et une représentation de /ʁ/ significativement plus élevée dans le jeu de données *Min*.

Ces différences de distribution segmentale, limitées et associées à des tailles d'effet modestes comme l'illustre la Figure 1, ne suffisent pas à expliquer l'ampleur des variations de performance observées. En complément, une analyse comparative entre jeux de données des mesures cepstrales utilisées par le système de vérification a été effectuée. Ces mesures consistent pour chaque trame de 10 ms de signal acoustique en un ensemble de 20 coefficients LFCC dont un correspondant au niveau d'énergie dans la trame. Les coefficients LFCC sont analogues aux coefficients MFCC (Davis & Mermelstein, 1980) plus couramment appliqués au traitement automatique de la parole, mais avec un banc de filtre linéaire plutôt que défini sur l'échelle fréquentielle logarithmique Mel afin d'accorder le même poids aux différentes plages de fréquence. En complément de ces 20 coefficients par trame, pour chacune de ces 20 dimensions la dérivée première Delta (différence entre trames consécutives) et la dérivée seconde DeltaDelta (différence entre valeurs consécutives de Delta) sont intégrées pour représenter la dynamique temporelle de l'information acoustique, selon une méthode couramment utilisée en traitement automatique de la parole. Chaque trame de 10 ms de signal acoustique est donc représentée par un ensemble de 60 valeurs numériques.

La comparaison statistique multivariée entre jeux de données *Min* et *Max* des coefficients LFCC, des valeurs de Delta et de celles de DeltaDelta, séparément pour chaque catégorie phonémique, a indiqué une différence significative des LFCC entre *Min* et *Max* pour toutes les classes phonémiques a l'exception des /v/ produits par les femmes, et sur les valeurs des dérivées Delta pour certaines classes phonémiques mais de façon peu consistante entre hommes et femmes, tandis que les effets sur les dérivées secondes DeltaDelta étaient presque inexistants. Ces résultats suggèrent que si la distribution du contenu segmentale est globalement comparable entre les échantillons qui aboutissent à la meilleure performance et ceux qui aboutissent à la pire performance, la réalisation de ces segments est directement en lien avec la performance du système de vérification du locuteur. Quand bien même elles ne touchent pas spécifiquement les voyelles (ce qui pourrait être lié à la relative imprécision de l'alignement automatique utilisé pour apparier les trames de signal acoustique et l'information segmentale, la résolution d'un tel système d'alignement étant au minimum de 10 ms dans le cas idéal), les différences relevées sur les valeurs de la dérivée première Delta pourraient être liées aux transitions formantiques, identifiées par McDougall & Nolan (2007) comme porteuses d'informations sur le locuteur.

Une telle analyse, si elle présente l'avantage de permettre un lien plus direct avec les performances du système de vérification du locuteur en s'appuyant sur les mêmes mesures que celles fournies en entrée du système, souffre toutefois de possibilités d'interprétations limitées. En effet les variations des coefficients cepstraux tels que les LFCC utilisés ici ne peuvent pas être directement reliées à mouvements articulatoires ou à des différences acoustiques perceptibles. Afin d'obtenir des informations plus directement interprétables sur les caractéristiques des jeux de données associés aux performances plus ou moins bonnes du système de vérification du locuteur, une analyse des fréquences des quatre premiers formants a été effectuée, dans un premier temps sur l'ensemble des voyelles orales produites par les 107 locuteurs du corpus BREF 120 (173 728 voyelles produites par les 64 femmes, 154 288 voyelles produites par les 43 hommes). On peut noter que cette première partie de l'analyse,

visant à documenter la variabilité intra- et inter-locuteurs, rejoint par ces méthodes et objectifs certains des travaux que j'ai menés plus récemment.

L'analyse descriptive a indiqué une variabilité entre locuteurs fortement dépendante de la voyelle, avec par exemple sur F1 un écart-type de l'ordre de 30Hz pour la voyelle /i/ et proche de 100Hz pour la voyelle /ø/. Par ailleurs sur les données du corpus BREF 120, la variabilité intra-locuteur est supérieure à la variabilité inter-locuteur pour l'ensemble des voyelles et des formants à l'exception du F3 des /œ/ produits par les femmes. L'analyse statistique multivariée prenant en compte conjointement les fréquences des quatre premiers formants a indiqué une taille d'effet du locuteur variable en fonction de la voyelle, avec un faible potentiel discriminant pour les voyelles fermées arrondies /y/ (taille d'effet η^2 de 10% pour les femmes et 15% pour les hommes) et /u/ (η^2 de 11% pour les femmes et 13% pour les hommes) ainsi que dans une moindre mesure pour la non-arrondie /i/, et plus important pour /œ/ (η^2 de 37% pour les femmes et 30% pour les hommes) ainsi que dans une moindre mesure pour /a/, /ɛ/, et /e/. La Figure 2 récapitule les tailles d'effet mesurées pour chacune des dix voyelles orales analysées à partir de cette analyse multivariée. De plus une analyse univariée effectuée séparément sur chacun des quatre premiers formants a indiqué des effets légèrement supérieurs sur F3 et F4 que sur F1 et F2. En complément des segments nasals considérés dans la littérature comme plus variables entre locuteurs que les segments oraux, cette analyse suggère donc que lorsque les extraits utilisés pour les comparaisons de voix peuvent être sélectionné, les voyelles moyennes et ouvertes auraient un potentiel de discrimination plus important que les voyelles fermées, et que l'analyse formantique dans une optique de comparaison entre locuteurs ne doit pas négliger les formants supérieurs.

L'analyse des caractéristiques des jeux de données *Min* et *Max* aboutissant respectivement à la meilleure et à la moins bonne performance du système de vérification du locuteur a été étendue au-delà de l'analyse des fréquences des formants à l'aire du triangle vocalique et à des dimensions décrites dans la littérature comme variables entre locuteurs, principalement le registre de fréquence fondamentale (Van Dommelen, 1987), et la coarticulation à travers la dynamique des formants (McDougall, 2006), le locus étant estimé comme l'évolution de F2 entre le début et le milieu de la voyelle. Toutefois, hormis une différence modérée de f0 entre *Min* et *Max* pour les hommes, ces mesures ne révèlent pas de différence entre les deux jeux de données. Par ailleurs des mesures de jitter et de shimmer ont également été incluses mais n'ont pas révélé de différence significative entre jeux de données, et doivent par ailleurs être interprétées avec prudence lorsqu'elles sont extraites sur la parole continue en raison de l'influence des variations locales de f0 sur ces mesures. La comparaison des valeurs formantiques entre les deux jeux de données a indiqué des différences significatives mais d'amplitude modérée et ne concernant que certaines voyelles et certains formants (F1 de /ɛ/ pour les femmes, F2 de /e/ pour les hommes, F4 de /œ, y, u/ pour les femmes et de /e/ pour les hommes), sans que ces différences soient consistantes entre les données des femmes et celles des hommes. Ainsi, aucune des mesures candidates considérées n'a fourni d'explication satisfaisante des importantes différences de performance relevées entre jeux de données et reflétées par les différences entre coefficients LFCC pour tous les segments sauf /v/.

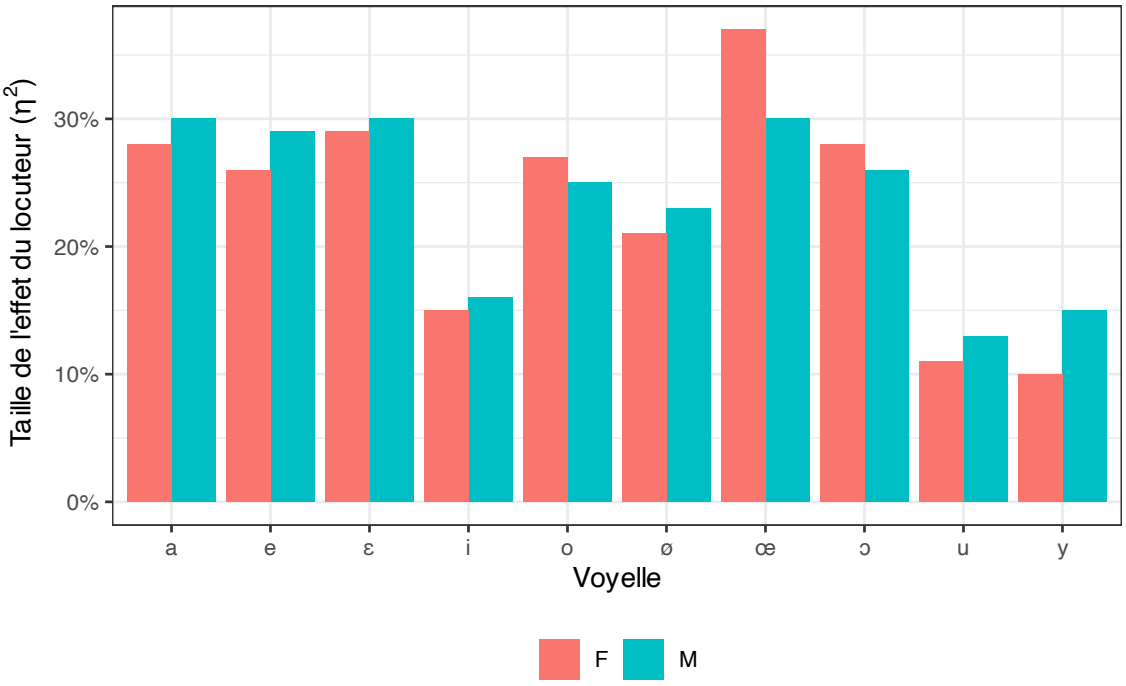


Figure 2 : Taille de l'effet du locuteur sur les quatre premiers formants pour chacune des dix voyelles orales analysées dans le corpus BREF 120 (Lamel et al., 1991), séparément pour les 64 femmes et les 43 hommes. Adapté des données présentées dans Kahn et al. (2011, [ACTI48]).

2.2.2 Comparaison des performances de l'humain et de la machine

Publications associées :

[ACTI49] Kahn, J., **Audibert, N.**, Rossato, S., & Bonastre, J.F. (2011). Speaker verification by inexperienced and experienced listeners vs. speaker verification system. *Proceedings of the 2011 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, République Tchèque, pp. 5912-5915.

[ACTI51] **Audibert, N.**, Larcher, A., Lévy, C., Kahn, J., Rossato, S., Matrouf, D., & Bonastre, J.F (2010). LIA human-based system description for NIST HASR 2010. *Proceedings of the NIST 2010 Speaker recognition evaluation workshop (Odyssey 2010 satellite)*, Brno, République Tchèque, actes CD-ROM.

Lors de la campagne d'évaluation NIST-SRE 2010, une nouvelle tâche annexe a été introduite en complément de la plus classique évaluation des performances de systèmes automatique de vérification du locuteur : la tâche *Human Assisted Speaker Recognition* (Greenberg et al., 2010), ci-après HASR, conçue pour évaluer dans quelle mesure l'intervention d'auditeurs peut améliorer ou compléter les performances du système dans des conditions particulièrement défavorables pour les systèmes automatiques. Cette tâche reposait sur des règles similaires à celle de l'évaluation générale ne faisant intervenir que des systèmes automatiques, avec un ensemble d'échantillons de parole inconnus parmi lesquels certains étaient produits par le même locuteur, mais dans laquelle un ou plusieurs auditeurs pouvaient intervenir dans le processus de décision, seul(s) ou en complément des systèmes automatiques. Ces échantillons consistaient en des dialogues en anglais entre un locuteur à identifier et une seconde personne considérée comme intervieweur, enregistrés dans des

conditions non-optimales et donc susceptibles d'être fortement bruités voire d'avoir subi des distorsions liées aux moyens d'enregistrement utilisés et non-documentés dans les données fournies par l'organisme NIST (qui précisaient uniquement s'il s'agissait d'appels téléphoniques ou d'entretiens), les conditions d'enregistrements n'étant pas toujours homogènes entre éléments à comparer. Deux ensembles d'extraits de taille plus réduite (respectivement 15 paires d'échantillons d'une durée de 2,5 minutes pour HASR1 et 150 paires pour HASR2 pouvant donner lieu à une comparaison, l'ensemble HASR1 étant inclus dans HASR2) que ceux utilisés dans la campagne d'évaluation dédiée aux systèmes automatiques ont ainsi été présélectionnés par les organisateurs de la campagne d'évaluation en raison des difficultés posés par ces extraits aux systèmes automatiques de vérification du locuteur à la fois dans des comparaisons « cible » et « imposteur », la sélection finale étant recentrée sur des extraits aboutissant également à des confusions de la part d'un panel auditeurs sollicités dans le cadre d'un prétest. La tâche proposée était donc choisie pour être particulièrement complexe. De même que dans la campagne d'évaluation des systèmes automatiques, les échantillons étaient répartis en deux catégories : modèle (locuteur cible à identifier) ou test (échantillons pouvant avoir été produits ou non par le locuteur cible, cette information n'étant divulguée que postérieurement à la campagne d'évaluation).

Le choix du Laboratoire d'Informatique d'Avignon pour la soumission initiale a été de proposer une contribution dans laquelle j'ai été largement impliqué. La décision sur l'ensemble HASR2 composé de 150 paires d'échantillons était prise par un vote majoritaire à partir des décisions prises indépendamment par trois auditeurs francophones natifs (deux femmes de 25 et 36 ans, un homme de 31 ans), tous trois expérimentés en analyse de la parole. Afin d'obtenir une distribution de scores continue directement comparable avec celle issue du système automatique de vérification du locuteur au-delà des quatre valeurs possibles en fonction des décisions prises par chacun des trois auditeurs, un appariement avec la distribution des scores du système automatique était effectué a posteriori.

Le protocole d'évaluation par les auditeurs reposait sur la sélection d'extraits courts présentés aux auditeurs. En effet, dans une tâche de comparaison de voix les auditeurs humains prennent leur décision dans un délai beaucoup plus court que 2,5 minutes (Blatchford & Foulkes, 2006). Des extraits de six secondes ont donc été sélectionnés et concaténés afin de constituer les stimuli évalués par les auditeurs, présentés sous forme de comparaisons (*trials* selon la terminologie utilisée en vérification du locuteur) composées alternativement d'extraits issus de l'échantillon modèle et d'extraits issus de l'échantillon test à partir desquels les auditeurs devaient déterminer si le modèle et le test avaient été produits ou non par le même locuteur. Bien que cette information ne fasse pas partie des résultats soumis, pour chacune de ces comparaisons il était demandé aux auditeurs d'auto-évaluer sur une échelle de Likert à six points leur degré de confiance dans leur décision.

Les extraits de six secondes ont été sélectionnés afin de maximiser la proportion de trames de 10 ms considérées par le système comme étant des échantillons de parole, sans inclure de productions de l'intervieweur qui ne devaient pas être prises en compte dans la comparaison. Afin de signaler sans ambiguïté aux auditeurs la transition entre un extrait de six secondes issu de l'échantillon modèle à un extrait issu de l'échantillon de test et inversement, les extraits consécutifs concaténés dans une même comparaison étaient séparés par un « bip » de 50 ms constitué d'un son pur à 1000 Hz, précédé et suivi d'un silence de 75 ms. Préalablement à leur concaténation, l'ensemble des extraits combinés dans une même comparaison étaient normalisés à la même intensité acoustique.

Selon les contraintes imposées par le protocole HASR, les auditeurs devaient effectuer les comparaisons non seulement dans le même ordre, mais aussi de façon coordonnée puisque l'accès à la comparaison suivante était conditionné par la soumission de la décision prise sur le précédent. En conséquence, il était impossible de randomiser l'ordre de présentation des comparaisons comme cela est couramment fait pour les évaluations perceptives. L'évaluation a été effectuée dans un environnement calme, les auditeurs étant équipés d'un casque audio fermé. Les auditeurs avaient la possibilité de s'appuyer sur leur expérience en analyse de la parole pour compléter leur écoute par une inspection spectrographique des signaux de parole à évaluer, et éventuellement dans le cas des signaux les plus bruités, notamment en basses fréquences, de procéder à un filtrage après inspection de l'enveloppe spectrale. L'évaluation d'une comparaison pouvait être interrompue dès la décision prise, sans nécessité de l'écouter en intégralité. En pratique, les auditeurs ont mis entre 12 secondes (soit le délai minimum absolu pour écouter un extrait de six secondes du modèle et un extrait de six secondes du test) et 180 secondes pour prendre leur décision, pour une durée moyenne de 66 secondes, soit un temps nettement inférieur à la durée des échantillons de 2,5 minutes combinés pour constituer la comparaison d'origine. En combinant via un vote majoritaire les décisions des trois auditeurs, le taux de fausse acceptation était de 29% et le taux de faux rejet de 12%. Cette performance est à comparer à celle obtenue sur les mêmes données par un système automatique fondé sur une classification par Support Vecteur Machine (SVM), avec un taux de faux rejet identique de 12% mais un taux de fausse acceptation sensiblement plus faible de 13%. Par ailleurs à une exception près, la performance des auditeurs a été comparable à celle obtenue par les autres équipes ayant participé à l'évaluation, avec dans certains cas des taux de fausse acceptation plus faibles mais compensés par des taux de faux rejet plus élevés, et inversement.

Le panel d'auditeurs a ensuite été élargi à un ensemble plus étendu de 29 auditeurs naïfs, tous francophones natifs (20 femmes et 9 hommes âgés en moyenne de 29 ans) n'ayant pas d'expérience en matière d'analyse de la parole et ayant étudié l'anglais en moyenne pendant 9 ans sans avoir vécu plus d'un an dans un pays anglophone. Pour les besoins de cette évaluation, le protocole expérimental a été adapté, avec la randomisation de l'ordre de présentation des stimuli et l'équilibrage du nombre de comparaisons « cible » et « imposteur » en sélectionnant un sous-ensemble des comparaison incluses dans HASR2. Pour cela, les 51 comparaisons « cible » de l'ensemble HASR2 ont été complétées par une sélection de 51 comparaisons « imposteur » composée des neuf comparaisons « imposteur » de l'ensemble HASR1 complétés par les 42 de l'ensemble HASR2 les plus problématiques pour le système de vérification automatique du locuteur (c'est-à-dire les comparaisons « imposteur » donnant lieu aux scores les plus élevés, d'où un taux important de fausses acceptations). On peut noter en revanche qu'il n'a pas été possible de concilier ce critère avec un équilibre entre comparaisons portant sur des voix d'hommes et de femmes, les voix de femmes étant surreprésentées dans la sélection avec 75 comparaisons sur un total de 102 en raison des relatives meilleures performances du système automatique sur les voix d'hommes et de femmes.

Une autre évaluation a ciblé un ensemble de 18 auditeurs phonéticiens experts francophones natifs, en se limitant toutefois à un sous-ensemble de neuf comparaisons « cible » et neuf comparaisons « imposteur » en raison des difficultés de recrutement d'auditeurs experts pour un test aussi chronophage que celui proposé aux auditeurs naïfs. Je ne présente pas ici de façon détaillée les résultats issus de l'évaluation menée auprès des auditeurs experts sur ces 18 comparaisons, en revanche on peut noter que sur les

comparaisons évaluées à la fois par les experts et les naïfs, la performance moyenne mesurée comme la somme du taux de fausse acceptation et du taux de faux rejet n'est que légèrement meilleure pour les auditeurs expérimentés comparativement aux auditeurs naïfs, en raison d'un taux de fausse acceptation moindre (22% contre 28% pour les naïfs).

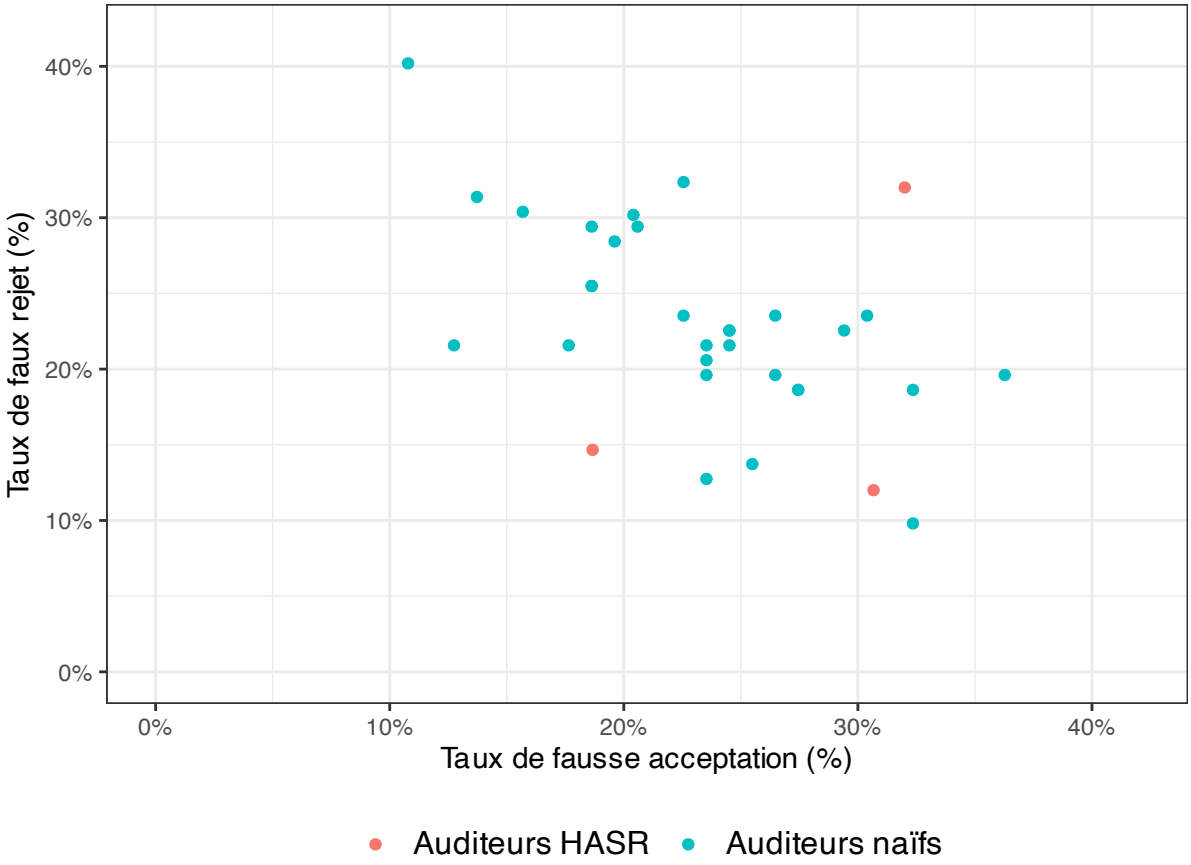


Figure 3 : Taux de fausse acceptation et de faux rejet obtenu sur les 102 comparaisons incluses dans le protocole élargi par les 29 auditeurs naïfs, comparé au taux de fausse acceptation et de faux rejet obtenu sur les mêmes comparaisons par les trois auditeurs expérimentés ayant pris part à l'évaluation de l'ensemble HASR2. Adapté de Kahn et al. (2011, [ACTI49]).

Bien que les comparaisons sélectionnées pour le protocole HASR l'aient été en raison de leur difficulté supposées, le niveau de difficulté pour les auditeurs a été extrêmement variable d'une comparaison à l'autre comme en témoigne la variabilité très importante des performances (10% à 87% de taux de réponses correctes pour les comparaisons « cible » et 3% à 90% pour les comparaisons « imposteur »). Comme illustré par la Figure 3 qui présente également les taux de fausse acceptation et de faux rejet issus des décisions prises par les trois auditeurs expérimentés de l'évaluation HASR sur les mêmes 102 comparaisons, des performances globales variables entre auditeurs naïfs ont été observées, avec des taux de fausse acceptation FA de 11% à 36% et des taux de faux rejet FR de 10% à 40%. Bien qu'également fortement variable, le taux d'erreur cumulé (FA+FR) est légèrement plus homogène entre auditeurs (34% à 56% pour les naïfs), du fait de la tendance de certains auditeurs à massivement considérer que les deux extraits ont été produits par le même locuteur d'où un taux de faux rejet plus faible mais en contrepartie un taux de fausse acceptation plus élevé, ou inversement. On peut noter par ailleurs que parmi les 29 auditeurs naïfs, seuls quatre ont identifié les voix avec une performance significativement supérieure au

hasard au sens de la métrique d-prime (Swets, 1964) qui permet de mesurer la performance humaine dans des tâches de décision binaire en compensant la tendance de certains juges à privilégier l'une des deux réponses. On peut noter également que les performances des auditeurs experts ayant participé à l'évaluation HASR avec la possibilité d'inspecter les spectrogrammes et éventuellement de filtrer les signaux les plus bruités n'obtiennent pas globalement de meilleures performances que les naïfs. En effet si l'un de ces trois experts identifie les voix présentées légèrement mieux que le meilleur des auditeurs naïfs avec un taux d'erreur cumulé FA+FR de 33%, une autre obtient la moins bonne performance avec un taux cumulé FA+FR de 64%. Les performances des auditeurs, qu'ils soient naïfs ou experts, sont donc globalement faibles sur cette tâche perceptive de vérification du locuteur, ce qui peut s'expliquer au moins en partie par la sélection effectuée, qui consiste en des extraits particulièrement problématiques à la fois pour les systèmes automatiques de vérification du locuteur et pour les auditeurs. En outre, les échantillons de parole à comparer étaient tous produits par des locuteurs inconnus des auditeurs, qui ne pouvaient donc pas tirer parti de l'avantage conféré par la familiarité avec les locuteurs à reconnaître (Hollien et al., 1974; Van Lancker et al., 1985).

De plus, les auditeurs ayant participé à ces évaluations n'étaient pas des anglophones natifs, ce qui a pu rendre la tâche plus complexe. Au vu des performances des autres équipes de recherche ayant pris part à l'évaluation HASR et qui à une exception près ont obtenu des résultats comparables à ceux du Laboratoire d'Informatique d'Avignon, les différences éventuelles entre la langue maternelle des locuteurs à reconnaitre et celle des auditeurs n'apparaissent pas comme un facteur évident pour expliquer les différences de performance. Toutefois, le fait de partager la même langue que les locuteurs ou à défaut une plus grande aisance dans cette langue pourrait constituer un facteur facilitant pour une telle tâche d'identification perceptive. En effet, le taux de réponses correctes données par les auditeurs naïfs était modérément corrélé au nombre d'années d'étude de l'anglais auto-déclaré par ces auditeurs.

La confrontation pour chacune des 102 comparaisons des performances des auditeurs et de celles du système de vérification automatique à base de SVM utilisé comme référence pour l'évaluation dans le cadre de la campagne HASR est illustrée par la Figure 4, dans laquelle les comparaisons « imposteur » sont représentées à gauche et les comparaisons « cible » à droite. La décision prise par les auditeurs (H sur la figure) correspond ici à un vote majoritaire à partir des décisions des 29 auditeurs naïfs et des trois experts. Comme on peut le voir, les divergences entre les décisions prises par les auditeurs et celles du système automatique sont nombreuses, ce qui est confirmé par la faible corrélation ($r = .21$) entre le taux de réponses correctes des auditeurs et les scores attribués par le système.

Bien que le taux de fausse acceptation obtenu par le système dans la campagne d'évaluation HASR ait été sensiblement plus bas que celui des auditeurs, ce qui se traduit également dans la Figure 4 par la proportion plus importante de réponses correctes du système pour les comparaisons « imposteur », 22% des comparaisons « cible » et 14% des comparaisons « imposteur » sont traités correctement par les auditeurs mais pas par le système. A condition de disposer d'un critère permettant d'identifier les comparaisons problématiques pour le système, ce constat pourrait ouvrir la voie à une amélioration des performances de comparaison de voix pour les applications criminalistiques via l'intervention ciblée d'auditeurs dans certains cas. Sur ces données, nous avons pu constater que les comparaisons aboutissant à une décision incorrecte du système automatique mais à une

décision correcte des auditeurs étaient pour en grande partie associées à un score attribué par le système automatique proche du seuil de décision. Cela nous a conduit à suggérer une approche hybride dans laquelle seules les comparaisons obtenant un score proche de ce seuil seraient réévaluées par un ou plusieurs auditeurs. Une simulation sur ces données en appliquant cette approche aux 10% des scores les plus proches du seuil, soit 11 comparaisons sur les 102 évaluées, conduit en effet bien à une amélioration des performances de vérification du locuteur.

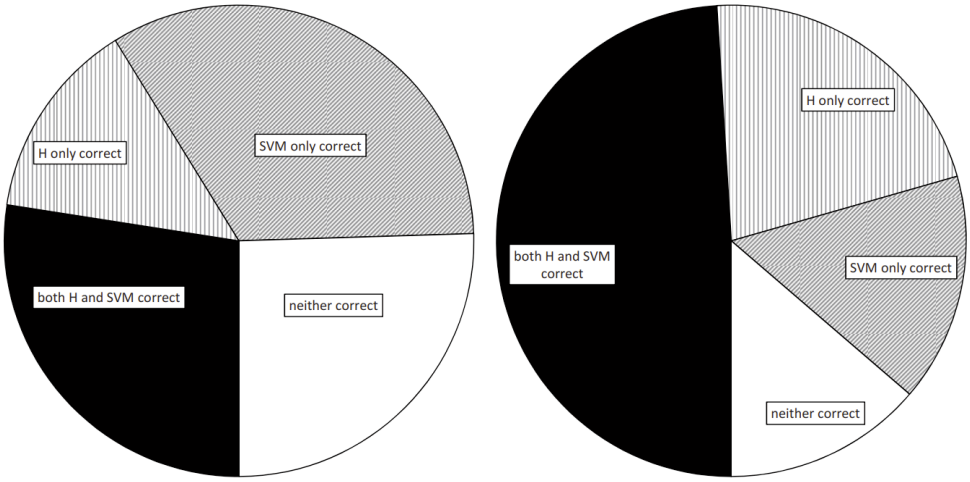


Figure 4 : Distribution des décisions correctes ou non, comparées entre système automatique à base de SVM et auditeurs étiquetés H pour « *Human* » (expérimentés dans le cadre de la campagne HASR et inexpérimentés confondus, la décision pour chaque comparaison étant prise par un vote majoritaire), pour les comparaisons « imposteur » dans lesquels l'échantillon de test n'est pas produit par le locuteur cible à gauche, et pour les comparaisons « cible » dans lesquels l'échantillon de test est produit par le locuteur cible à droite. D'après Kahn et al. (2011, [ACTI49]).

2.3 Variabilité entre sessions d'enregistrement

Dans le cadre du projet ANR Voxcrim mené en partenariat avec le Laboratoire d'Informatique d'Avignon, la Police Technique et Scientifique d'Ecully et le Laboratoire Parole et Langage d'Aix-en-Provence, les travaux du Laboratoire de Phonétique et Phonologie se sont concentrés plus particulièrement sur la caractérisation phonétique de la variabilité inter- et intra-locuteur à travers un ensemble de mesures ciblant diverses dimensions segmentales et suprasegmentales. Les études dans lesquelles j'ai été impliqué se sont concentrées tout particulièrement sur la modulation des informations rythmiques et acoustiques à court et à moyen terme.

La plupart de ces études se sont appuyées sur les corpus PATATRA et/ou PATAFreq (Fougeron et al., 2022, [ACTI16]) présentés en section 7.3.3 qui incluent de multiples sessions d'enregistrement des mêmes locuteurs à intervalles plus ou moins rapprochés, tandis que la dernière en date a exploité les données du corpus PTSVox (Chanclu et al., 2020) recueilli dans le cadre du projet VoxCrim.

2.3.1 Modulation rythmique et acoustique entre unités consécutives

Publications et communications associées :

[ACTI14] **Audibert, N.**, & Fougeron, C. (2022). Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population. *Proceedings of Interspeech 2022*. Incheon, Korea, pp. 4755-4759.

[ACTI15] Fougeron, C., & **Audibert, N.** (2022). Variabilité intra-individuelle en parole lue. *Actes des 34e Journées d'Études sur la Parole (JEP2022)*. Noirmoutier, France, pp. 145-153.

[ACTI24] Chardenon, E., Fougeron, C., **Audibert, N.**, & Gendrot, C. (2020). Dis-moi comment tu varies ton débit, je te dirai qui tu es. *Actes des 33èmes Journées d'Études sur la Parole*, 2020, Nancy, France, pp. 82-90.

[ACTI30] Gendrot, C., Chignoli, G., **Audibert, N.**, & Fougeron, C. (2018). Variabilité inter et intra locuteurs de mesures spectrales et prosodiques en parole lue. *Actes des 32èmes Journées d'Études sur la Parole*, Aix-en-Provence, France, pp. 46-54.

[INV3] Fougeron, C., & **Audibert, N.** (2023). Intraspeaker and interspeaker variability. *Invited plenary talk. SpeakVar Workshop*. 2-4 octobre 2023, Budapest, Hongrie.

[COM3] **Audibert, N.**, Fougeron C., & Chardenon, E. (2021). Do you remain the same speaker over 21 recordings? *XVII° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV)*, Zürich (en ligne), Switzerland.

Dans une série de travaux en collaboration avec Cécile Fougeron, amorcés dans le cadre du mémoire de master d'Estelle Chardenon que nous avons coencadré, j'ai exploré le rôle dans la variation inter- et intra-locuteur des modulations acoustiques entre unités successives, avec une granularité plus large que les variations locales entre trames consécutives de durée classiquement intégrées dans les paramétrisations du signal de parole pour le traitement automatique à travers les dérivées premières et secondes, en complément de mesures statiques plus classiques.

Auparavant, j'avais contribué à une première étape dans cette direction à travers ma contribution à une étude menée principalement par Cédric Gendrot et Gabriele Chignoli. Dans cette étude, les métriques rythmiques destinées à capturer les variations de durée entre unités vocaliques et/ou consonantiques consécutives (Grabe & Low, 2002; Dellwo, 2006), plus couramment appliquées à la comparaison inter-langues, ont été extraites pour les dix locuteurs du corpus PATATRA à l'aide de l'outil Correlatore (Mairano & Romano, 2010) sur la partie finale de la lecture de la fable « La bise et le soleil », et comparées entre sessions et locuteurs. Comme illustré par la Figure 5, cette analyse des métriques rythmiques a suggéré une distinction entre la plupart des locuteurs de ce corpus dans l'espace en deux dimensions des mesures normalisées de variabilité de la durée d'unités consécutives nPVI mesurées respectivement sur les voyelles (VnPVI) et sur les consonnes ou séquences de consonnes consécutives (CnPVI). Par ailleurs, cette étude a également confirmé à partir de mesures acoustiques statiques incluant notamment les quatre premiers moments spectraux sur les fricatives et de la paramétrisation cepstrale extraites de la lecture des listes de mots incluse dans le corpus PATATRA que la taille d'effet associée à la différence entre locuteurs était plus

importantes sur les fricatives et les nasales (mais aussi sur les sonantes) que sur les occlusives, conformément aux observations antérieures de Kahn (2011).

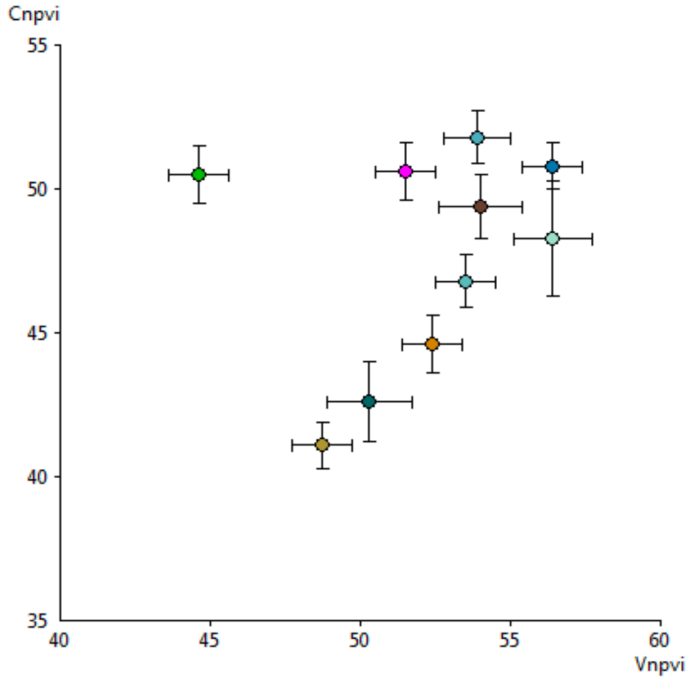


Figure 5 : Représentation en deux dimensions obtenue à l'aide de l'outil Correlatore (Mairano & Romano, 2010) des valeurs moyennes par locuteur et de la variation entre sessions pour un même locuteur des mesures normalisées de variabilité locale de la durée nPVI mesurées respectivement sur les voyelles (Vnpvi) et les séquences de consonnes consécutives (Cnpvi), pour les dix locuteurs du corpus PATATRA (Fougeron et al., 2022, [ACTI16]), chacun représenté par une couleur différente. D'après Gendrot et al. (2018, [ACTI30]).

Nos travaux suivants ont porté également sur les tâches de lecture de textes du corpus PATATRA ou du corpus PATAFreq, mais se sont concentrés sur la variation entre unités consécutives plus longues d'environ dix syllabes chacune désignées ci-après par le terme anglais de « *chunk* », sélectionnées afin d'être directement comparables entre locuteurs et sessions d'enregistrement et de permettre des choix de segmentation reproductibles. Les frontières de début et de fin de chaque chunk ont été annotées manuellement, ainsi que la position de l'ensemble des pauses silencieuses perceptibles par les annotateurs, les pauses situées entre deux chunks étant par convention reliées au premier des deux. Outre l'encadrement du mémoire d'Estelle Chardenon qui a constitué le point de départ de ces recherches, nous avons pour la réalisation de ces tâches d'annotation mis à contribution plusieurs groupes d'orthophonistes en cours de formation dont nous avons encadré les stages d'initiation à la recherche, dont une partie consistait en une contribution à cette tâche d'annotation. Dans ce cadre, certains enregistrements faisaient l'objet d'annotations multiples afin de pouvoir évaluer la consistance inter-annotateur en termes de décalage des frontières temporelles attribuées et de choix d'annoter ou non comme pause les portions silencieuses les plus courtes.

Bien qu'elle ne permette pas directement une analyse acoustique fine au niveau segmental, la segmentation en chunks ainsi réalisée et l'annotation des pauses silencieuses peuvent donner lieu à l'extraction d'un certain nombre de mesures au niveau de chaque

chunk, dont nous dérivons ensuite des mesures de variabilité entre chunks consécutifs notées d, calculées comme suit pour chaque chunk d'indice i à l'exception du premier :

$$d_i = \frac{|chunk_i - chunk_{i-1}|}{(chunk_i + chunk_{i-1})/2}$$

Ainsi la mesure de variabilité chunk-à-chunk du niveau moyen de fréquence fondamentale d(f0) est calculée à partir du niveau moyen de f0 extrait de chaque chunk, d'où un nombre de valeurs de d(f0) correspondant au nombre de chunks moins un.

Dans une première étude exploitant la lecture de la fable « La bise et le soleil » découpée en dix-huit chunks par les huit locuteurs francophones natifs du corpus PATATRA à raison de trois lectures par an pendant sept ans, nous nous sommes concentrés sur les mesures temporelles. Les mesures retenues dans un premier temps ont été des mesures temporelles de débit de parole pauses incluses et de débit articulatoire, mesuré sur chaque chunk comme la durée du chunk dont était retranchée la durée cumulée des pauses, divisée par le nombre de syllabes dans le chunk, ainsi que et sur la variation locale de ces mesures. Nous avons ensuite étendu l'analyse de ces données à un ensemble de mesures acoustiques compatibles avec des extraits de parole d'environ dix syllabes, supposées capturer différentes dimensions de variation de la voix et de la parole et pouvant être extraites automatiquement à plus grande échelle dans l'optique d'une extension de ce protocole d'analyse. Parmi les mesures relatives à l'organisation temporelle de la parole, outre le débit articulatoire et le débit de parole pauses incluses, nous avons extrait le rapport entre la durée détectée comme voisée et la durée totale du chunk. Cette dernière mesure, fréquemment appelée v-ratio et utilisée au niveau segmental comme corrélat acoustique du degré de voisement (Snoeren et al., 2006), a été retenue afin de fournir indirectement des informations sur la répartition entre segments voisés et non-voisés dans les chunks analysés. La fréquence fondamentale moyenne dans chaque chunk ainsi que sa variabilité (exprimée en demi-tons pour la comparabilité entre registres plus ou moins élevés, notamment entre hommes et femmes) ont été extraites comme estimation de l'activité laryngée et des variations intonatives. Enfin les caractéristiques générales de qualité de voix ont été estimées à partir du spectre moyen à long terme LTAS sur l'ensemble des portions détectées comme voisées dans le chunk par la pente calculée comme la différence entre l'énergie dans la bande de fréquence 0-1kHz et l'énergie dans la bande de fréquence 1-4kHz. De même que précédemment, les mesures de variabilité chunk-à-chunk ont été extraites pour chacune de ces six variables, soit un total de douze variables.

L'effet sur chacune de ces douze variables de la session d'enregistrement, du locuteur et de leur interaction a été évalué par un modèle de régression linéaire mixte, le chunk étant considéré comme un facteur aléatoire. Cette analyse a montré un effet significatif de la session et de l'enregistrement sur l'ensemble des mesures extraites sur chaque chunk, et un effet du locuteur sur les mesures de variabilité chunk-à-chunk à l'exception de la pente du LTAS, suggérant que la magnitude de la variation acoustique entre sessions d'enregistrements serait dépendante du locuteur, avec toutefois une variabilité intra-session plus constante entre sessions.

Afin de décrire plus spécifiquement la variabilité en permettant une comparaison entre mesures exprimées dans des unités différentes, après standardisation en z-scores nous avons calculé pour chaque mesure et pour chaque locuteur l'écart-type entre sessions d'enregistrement, interprétable comme un indice normalisé de variabilité. La comparaison

entre locuteurs et entre mesures des valeurs de cet indice de variabilité est récapitulée par la Figure 6, sous forme d'un profil de variation pour chaque locuteur.

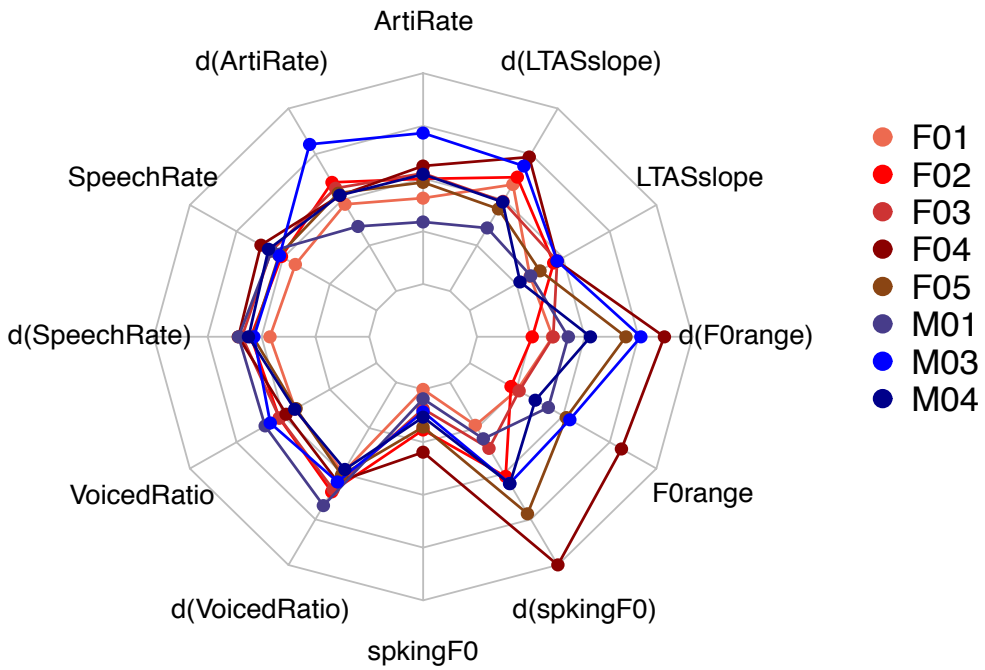


Figure 6 : Profils de variation entre sessions d'enregistrement de la lecture de texte des huit locuteurs francophones natifs du corpus PATATRA, pour les six variables mesurées sur les 18 chunks du texte et la variation chunk-à-chunk de chacun de ces variables notée d(variable). Les valeurs pour chacune de ces douze mesures sont exprimées sous forme d'écart-type normalisé afin de comparer leur variabilité sur une échelle commune. Les femmes sont désignées par les codes en F et les hommes par les codes en M. D'après Audibert et al. (2021, [COM3]).

Ainsi, la fréquence fondamentale moyenne, notée ici *spkingF0*, est plus stable entre sessions d'enregistrement que les autres variables prises en compte dans notre analyse. Dans une moindre mesure, c'est également le cas de la pente du spectre moyen voisé à long terme notée *LTASslope*. Cette observation pourrait suggérer une certaine stabilité entre sessions d'enregistrement des caractéristiques de qualité de voix, mais doit être nuancée du fait des propriétés spectrales suprasegmentales également capturées par une mesure aussi globale que la pente du spectre moyen à long terme.

A l'inverse, la variation entre chunks consécutifs du niveau moyen de fréquence fondamentale *d(spkingF0)* et de la pente du LTAS voisé *d(LTASslope)*, la plage de variation de la fréquence fondamentale *F0range* et sa variation entre chunks consécutifs *d(F0range)* et le débit articulatoire *ArtiRate* et sa variation entre chunks *d(ArtiRate)* sont les mesures les plus variables entre sessions d'enregistrement, avec toutefois d'importantes différences entre locuteurs.

Bien que certains locuteurs soient globalement plus variables que d'autres entre sessions d'enregistrement, avec des différences plus importantes pour les femmes que pour les hommes sur la plupart des dimensions acoustiques considérées, l'analyse des profils de variabilité ne fait pas émerger de patron consistant entre hommes et femmes. De plus les différences interindividuelles dans les profils de variation ne peuvent pas se réduire à un niveau de variabilité plus ou moins important, avec par exemple des locuteurs comme M03 qui varient plus largement leur débit articulatoire et la variation de leur débit entre chunks

consécutifs (ce qui par contraste avec les variations du débit de parole peut être interprété comme une fluctuation plus importante de la gestion des pauses silencieuses), et d'autres comme notamment F03 qui varient plus largement leurs caractéristiques locales de fréquence fondamentale.

Nous avons par la suite mené une série d'autres analyses sur les données du corpus PATAFreq, qui comprend neuf locuteurs enregistrés lors de six à dix sessions (8,9 en moyenne) moins espacées dans le temps que celles du corpus PATATRA. Pour cela, les lectures des deux textes spécifiques au corpus PATAFreq (à l'exclusion de la lecture de la fable « La bise et le soleil ») ont été retenues. Outre les mesures temporelles et acoustiques considérées précédemment, nous avons également inclus la mesure HNR de rapport entre l'énergie harmonique et l'énergie du bruit, en considérant à la fois sa valeur moyenne dans le chunk et la variabilité dans le chunk, et la variabilité à l'intérieur du chunk de l'intensité acoustique dans quatre bandes de fréquence : les bandes 0-1kHz, 1-2.5kHz et 2.5-4kHz considérées comme fortement influencées par le contenu segmental en traitant séparément les plages fréquentielles supposées capturer respectivement l'énergie des trois premiers formants, et la bande 4-8kHz considérée comme moins dépendante du contenu segmental et donc potentiellement plus dépendante du locuteur. En outre, la mesure de débit de parole pauses incluses a été remplacée par une mesure de temps de pause cumulé afin d'obtenir une mesure indépendante du débit articulatoire et complémentaire de celui-ci. Selon le même principe que celui appliqué dans l'étude précédente, une mesure de variation chunk-à-chunk a été dérivée de chacune de ces mesures, à l'exception de la variabilité intra-chunk du rapport entre énergie harmonique et bruit HNR, pour un total de 21 variables prises en compte.

Afin de déterminer l'importance relative des différentes variables dans la distinction entre locuteurs et entre sessions d'enregistrement pour un même locuteur, nous avons adopté une procédure de sélection de variable à partir d'une tâche de classification, selon le principe général des forêts aléatoires (Cutler et al., 2012). Ce type d'approche est plus couramment appliquée en traitement automatique afin d'identifier les variables les plus efficaces pour une tâche de classification sur un ensemble de données afin de réduire la dimensionnalité des données et optimiser ainsi l'apprentissage. Cependant elle peut aussi être utilisée plus directement pour l'analyse de données complexes afin de déterminer parmi un ensemble de variables lesquelles permettent le mieux de classifier entre catégories (voir par exemple Al-Tamimi (2017) pour une application à des données de parole du principe des forêts aléatoires conditionnelles). Notre choix s'est porté sur l'algorithme implémenté dans le package R Boruta (Kursa & Rudnicki, 2010), qui évalue chaque variable à partir de la baisse des performances de classification induite par la permutation aléatoire entre éléments, et lui attribue ainsi un score indépendant de l'échelle selon laquelle cette variable est exprimé, ce qui permet la comparaison entre variables. Cette méthode, déjà appliquée à des données phonétiques par Klein et al. (2019), permet en outre d'obtenir une distribution des scores d'importance à partir de simulations multiples. Le score étant toutefois dépendant du nombre de classes et d'observations, nous avons retenu comme métrique le rang de chaque variable classée de la plus efficace à la moins efficace dans la tâche de classification. Nous avons appliqué la même procédure à deux tâches de classification : la classification en locuteurs et la classification entre sessions d'un même locuteur (la procédure étant dans ce cas appliquée indépendamment pour chaque locuteur). Dans une première version de cette étude (Fougeron & Audibert, 2022, [ACTI15]) nous avons également considéré la classification entre les deux textes lus, toutefois cette partie de l'analyse s'est révélée peu informative et nous nous sommes recentrés ensuite

sur une analyse des productions des locuteurs pour les deux textes confondus, avec donc un nombre plus élevé de chunks analysés par locuteur et par session.

La Figure 7 présente les variables classées par ordre d'importance pour la séparation en sessions au sein des productions de chaque locuteur et pour la séparation en locuteurs. Une série de variables qui capturent principalement les variations liées à la voix, essentiellement en basses fréquences, peuvent être identifiées parmi les variables qui discriminent le mieux entre sessions d'un même locuteur et sont donc plus instables. La valeur moyenne de fréquence fondamentale et la variation intra-chunk de l'énergie en basses fréquences (0-1kHz) figurent ainsi parmi les cinq variables les plus discriminantes entre sessions pour tous les locuteurs, à l'exception d'une locutrice pour chacune de ces deux variables. De plus la pente du spectre moyen voisé à long terme et le rapport entre énergie harmonique et bruit comptent parmi les cinq variables les plus discriminantes entre sessions pour sept des neuf locuteurs.

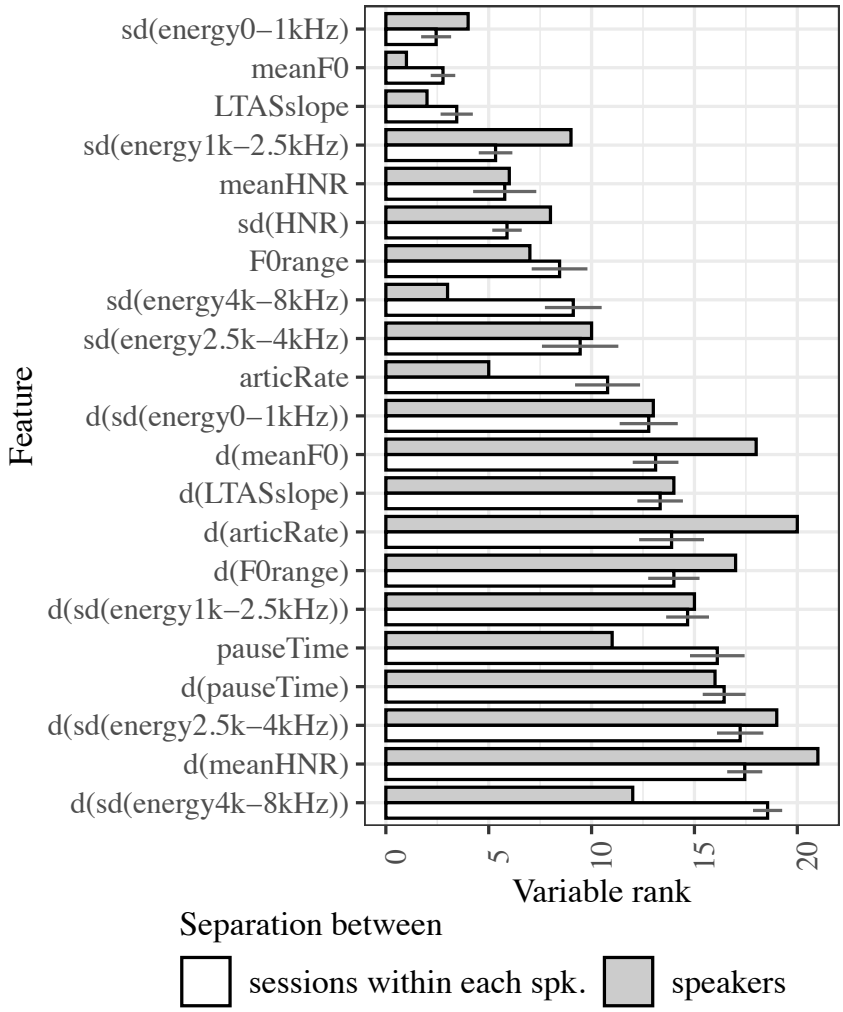


Figure 7 : Classement de l'importance attribuée à chaque variable par l'algorithme de Boruta (Kursa & Rudnicki, 2010) pour la séparation en locuteurs sur l'ensemble des données, et pour la séparation en sessions pour chaque locuteur (classement moyen sur les 9 locuteurs, les barres d'erreur représentent l'erreur standard). Les valeurs les plus faibles représentent les variables jugées les plus importantes pour la tâche de discrimination considérée. Les variables sont présentées de haut en bas par ordre décroissant d'importance moyenne (et donc par ordre croissant de rang) dans la tâche de classification entre sessions. D'après Audibert & Fougeron (2022, [ACTI14]).

Comme le suggère la forme générale de la Figure 7, les variables qui discriminent le mieux entre sessions ont tendance à être les mêmes que celles qui discriminent le mieux entre nos neuf locuteurs. On retrouve en effet la fréquence fondamentale moyenne, la pente du spectre moyen voisé à long terme et la variation intra-chunk de l'énergie en basses fréquences (0-1kHz) parmi les cinq variables qui discriminent le mieux entre locuteurs. Cette tendance est confirmée par la corrélation de Spearman relativement élevée (ρ=.81) entre le rang pour la séparation en locuteurs et le rang moyen pour la séparation en sessions au sein des productions de chaque locuteur.

On peut observer que globalement, le débit articulatoire discrimine mieux entre locuteurs qu'entre sessions, mais avec une variabilité entre sessions fortement dépendante du locuteur. On peut également noter que comme attendu, la variabilité de l'intensité en hautes fréquences (4-8kHz), supposée plus dépendante du locuteur, discrimine mieux entre locuteurs qu'entre sessions. Cependant les valeurs de cette variable fluctuent également entre sessions, et l'inspection des valeurs au sein d'une même session suggère que la variation segmentale capturée par cette mesure est plus importante que prévue.

Enfin, on peut relever que les mesures de modulation chunk-à-chunk préfixées par « d() », bien que peu discriminantes entre locuteurs, sont également parmi les plus stables entre sessions. Cela suggère que l'introduction de mesures définies sur des empans temporels plus longs que ceux classiquement considérés en traitement automatique pourrait avoir un intérêt pour les tâches de classification qui reposent sur des descripteurs qui ne sont pas définis exclusivement sur des trames locales.

2.3.2 Stratégies individuelles de coarticulation labiale

Publication associée :

[ACTI12] Guitard-Ivent, F., Wohmann-Bruzzo, L., **Audibert, N.**, & Fougeron, C. (2023). Speaker-specific anticipatory labial coarticulation in French. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic. pp. 654-658.

Dans une étude menée en collaboration avec Fanny Guitard-Ivent et qui s'appuyait en partie sur le travail réalisé dans le mémoire de master de Louise Wohmann-Bruzzo que j'ai coencadré avec Cécile Fougeron, nous avons cherché à caractériser la variabilité inter-indviduelle dans les stratégies de coarticulation labiale anticipatoire. Nous nous sommes concentrés pour cela sur la syllabe /sy/ en raison de la proximité articulatoire entre les deux sons qui composent cette syllabe, qui favorise la coarticulation (Yeni–Komshian & Soli, 1981), l'arrondissement du /y/ et l'allongement de la cavité antérieure pouvant être anticipés par le locuteur pendant la production du /s/ avec des conséquences directes sur le spectre du bruit produit. La fréquence dans laquelle se concentre l'énergie du bruit, mesurée par le centre de gravité spectral (CoG) ou par la fréquence du pic d'énergie spectrale, baisse en effet dans le contexte d'une voyelle arrondie (voir par exemple Jongman et al. (2000)).

Pour cela, nous avons analysé les productions extraites de deux corpus de 14 locuteurs francophones (11 femmes et 3 hommes) ayant produit de multiples occurrences de /s/ suivi de /y/ ou d'une voyelle antérieure non-arrondie (/i/ ou /e/ selon le corpus utilisé) utilisée comme base de comparaison. Le premier corpus consistait en un sous-ensemble du corpus PATAFreq (Fougeron et al., 2022, [ACTI16]) qui comprend 6 femmes et 3 hommes ayant produit 24 à 60 /sy/ et 24 à 62 /se/ par locuteur dans la lecture de texte lors de 6 à 10 sessions

d'enregistrement, et correspond aux locuteurs étiquetés dans la Figure 8 par un code composé de F ou H en fonction du sexe suivi de deux chiffres. Le second corpus, conçu et enregistré par D'Alessandro (2022), comprend 5 femmes ayant produit 135 /sy/ et 108 à 134 /si/ par locutrice dans une condition de lecture de phrases lors de 5 sessions, et correspond aux codes à deux ou trois lettres.

Le degré de coarticulation labiale anticipatoire et son évolution dans le temps ont été caractérisés pour chaque locuteur à partir de la mesure du CoG sur chaque occurrence de /s/ en dix points équidistants, les spectres FFT dont sont tirés chaque mesure de CoG étant calculés sur une fenêtre de Hanning de 10ms et filtrés pour retenir les fréquences de 350 Hz à 18 kHz afin de limiter l'effet éventuel du voisement résiduel, notamment au début et/ou à la fin de la production du /s/. Les trajectoires en fonction du temps normalisé du centre de gravité spectral entre le début et la fin du /s/ ont fait l'objet d'une modélisation par un modèle GAMM prenant en compte la durée segmentale, afin de comparer les trajectoires-types en fonction du contexte vocalique (arrondi ou non-arrondi), du locuteur et de la session d'enregistrement. Les trajectoires ainsi modélisées pour chaque locuteur sont présentées dans la Figure 8.

A l'exception de l'un des trois hommes, les valeurs de CoG sont globalement distinctes entre contexte arrondi et non-arrondi, avec toutefois d'importantes variations inter-individuelles, à la fois en termes de degré de coarticulation, matérialisé par l'amplitude de la différence entre trajectoires, mais aussi d'étendue temporelle de la coarticulation anticipatoire, les réalisations spectrales du /s/ de certains locuteurs divergeant plus tôt que d'autres sans que cela soit lié de façon systématique à l'amplitude de cette divergence.

Une autre question à laquelle nous avons cherché à répondre dans cette étude a été celle de la stabilité de la coarticulation anticipatoire entre sessions pour un même locuteur, en nous concentrant sur le cas des neuf locuteurs qui coarticulent le plus et en considérant comme métrique pour estimer le degré de coarticulation l'écart cumulé entre trajectoires pour chaque locuteur et chaque session. Les résultats de cette analyse ont mis en évidence d'importantes différences individuelles en termes de stabilité des productions, avec des locuteurs très cohérents dans leurs schémas coarticulatoires et d'autres beaucoup plus variables. Ces différences semblent relever de la variation individuelle et non du sexe des locuteurs ou de la différence entre les tâches de production dans les deux corpus considérés, avec des patrons de variabilité différents observées à la fois pour les hommes et les femmes dans les données des deux corpus. Elles n'apparaissent pas non plus comme liées à l'ampleur de la coarticulation, puisque parmi les locuteurs qui coarticulent le plus on observe à la fois des locuteurs très stables et d'autres très variables d'une session à l'autre. En dépit de ces variations, les différences inter-individuelles restent néanmoins supérieures aux différences intra-individuelles dans ces données, la majorité des locuteurs montrant des patrons de coarticulation consistants entre sessions, ce qui rejoint les résultats obtenus par Whalen & Chen (2019) à partir de nombreuses répétitions produites par un même locuteur.

Les variations inter-individuelles de patrons de coarticulation peuvent s'expliquer par des spécificités physiologiques et anatomiques (Fuchs & Toda, 2010), mais aussi par des stratégies individuelles liées à la gestion du débit comme suggéré par les résultats de D'Alessandro & Fougeron (2021). Cependant dans les données que nous avons analysées, certes restreintes à un nombre de locuteurs qui limite la portée d'une telle interprétation, la durée segmentale utilisée comme approximation du débit de parole n'apparaît pas comme liée au degré de coarticulation observé ni à son empan temporel.

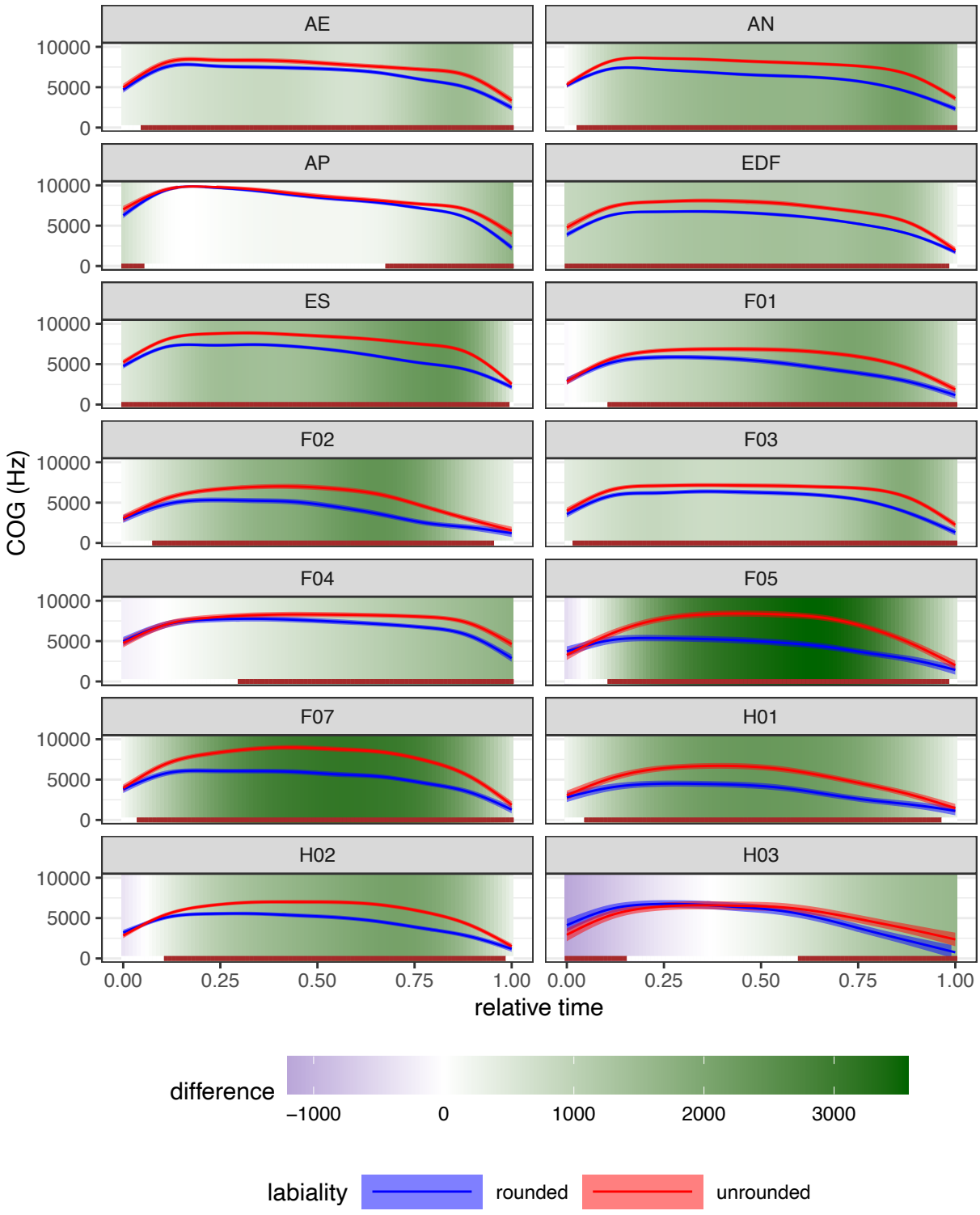


Figure 8 : Trajectoire de l'évolution pour chaque locuteur de la fréquence du centre de gravité spectral CoG entre le début et la fin du /s/ modélisée par un modèle GAMM en fonction de la présence d'un /y/ (courbe bleue) ou d'une voyelle non-arrondie (rouge) en contexte droit, toutes sessions confondues. L'enveloppe autour des courbes représente l'intervalle crédible à 95%. La couleur de fond reflète l'amplitude de la différence entre les deux trajectoires, le surlignage rouge foncé de l'axe des abscisses indique une différence entre trajectoires pouvant être considérée comme significative. D'après Guitard-Ivent et al. (2023, [ACTI12]).

2.3.3 Variation inter- et intra-locuteur de l'espace vocalique

Publication associée :

[ACTI1] **Audibert, N.**, Fougeron, C., & Meunier, C. (2024). Do Speaker-dependent Vowel Characteristics depend on Speech Style? *Proceedings of INTERSPEECH 2024*, Kos Island, Greece, pp. 3669-3673.

Dans une étude menée en collaboration avec Cécile Fougeron et Christine Meunier (LPL Aix-en-Provence), nous avons cherché à établir dans quelle mesure les différences individuelles de réalisation des voyelles, documentées comme à la fois variables entre locuteurs et entre styles de parole pour un même locuteur (cf. à ce sujet les développements présentés en section 4.2 de ce document de synthèse), sont transposables d'un style de parole à l'autre. Outre les questions plus générales relatives à une meilleure compréhension de ces deux facteurs de variation dans la parole, l'interaction entre les variations de la réalisation des voyelles liées au style de parole et celles dépendantes du locuteur peut également avoir une incidence pour des problématiques de police scientifique. En effet, dans ce cadre les réquisitions portent régulièrement sur des comparaisons de voix dans lesquelles la pièce de comparaison consiste en des extraits de parole lue par la personne suspectée tandis que la pièce de question, issue par exemple d'interceptions téléphoniques, correspond à un style de parole très différent, généralement spontané. Documenter la variation entre styles de parole comparativement à la variation entre locuteurs peut ainsi contribuer à mieux cerner les limites de telles comparaison au-delà de celles induite par la qualité des enregistrements traités.

Bien que cet aspect ait été relativement peu étudié, quelques études précédentes ont suggéré que les variations liées au style de parole pourraient impacter les performances de l'identification du locuteur. Ainsi, Dankovičová & Nolan (1999) ont conclu à une amélioration des performances d'identification du locuteur en suédois lorsque les données d'entraînement incluait une variation systématique des styles de parole pour chaque locuteur. A partir de productions de locuteurs anglophones produisant des variations de style sur la parole lue et notamment une condition de lecture rapide en opposition à la tâche de lecture à vitesse confortable, Grimaldi & Cummins (2009) ont conclu à un effet modéré du changement de style sur les performances d'identification et de vérification du locuteur lorsque les conditions d'enregistrement ne changement pas entre styles de parole. Plus spécifiquement sur les voyelles, Cavalcanti et al. (2023) ont comparé la capacité de discrimination des propriétés spectrales des voyelles produites par des locuteurs du portugais brésilien entre deux styles de parole non-scriptés (une situation de dialogue entre locuteurs familiers et une interview par une personne non-familière), et ont conclu que les fréquences formantiques et particulièrement F3 et F4 permettaient de discriminer entre locuteurs avec une meilleure performance en condition de dialogue.

Pour notre part, nous avons pris en compte 385 724 voyelles produites par 23 locuteurs (11 femmes et 12 hommes, âgés de 18 à 24 ans et élèves-policiers) du corpus PTSVox (Chanclu et al., 2020), ayant tous été enregistrés lors d'un minimum de deux sessions dans deux styles de parole : la lecture de la fable « La bise et le soleil », et un entretien libre dans lequel les locuteurs étaient interrogés sur leurs études et leurs loisirs. Bien que le corpus PTSVox comprenne également des enregistrements réalisés par téléphone à des fins de comparaison, nous nous sommes concentrés ici sur les enregistrements réalisés à l'aide d'un microphone et d'une carte son dans un environnement calme afin de documenter la variation acoustique de

réalisation des voyelles dans des conditions optimales, et avons utilisé une segmentation en phones corrigée manuellement suite à une première passe d'alignement forcé automatique. Afin de disposer d'un nombre d'exemplaires suffisant par locuteur, par session et par style de parole et en raison de la confusion fréquente avec le schwa en français, les voyelles /ø/ et /œ/ n'ont pas été incluses. En revanche, afin de prendre en compte les voyelles nasales pour lesquelles la détection des formants est particulièrement problématique, nous avons opté pour une description de la réalisation acoustique des voyelles en un point temporel centré sur le milieu de la voyelle à partir de douze coefficients cepstraux MFCC (voir section 7.2.3.3 pour des précisions sur cette approche appliquée à la description de l'espace vocalique), le coefficient 0 correspondant au niveau global d'énergie étant exclu de l'analyse. Nous avons donc pris en compte les six catégories de voyelles orales /i, y, u, E, O, a/ (E et O représentant les archiphonèmes respectivement des paires de voyelles /e, ɛ/ et /o, ɔ/), et les trois catégories de voyelles nasales /ɛ̃, ɔ̃, ɑ̃/.

Afin de caractériser la variation entre locuteurs, nous avons calculé une distance entre locuteurs dans l'espace des coefficients MFCC pour chaque paire de locuteurs, au sein de chaque style de parole et séparément pour les femmes et les hommes. Cette distance inter-locuteurs permet ainsi d'estimer dans quel mesure le locuteur est distinct des autres locuteurs de même sexe pour chacun des deux styles de parole. Chaque locuteur est également caractérisé par une distance intra-locuteur, qui capture la distance acoustique entre ses voyelles en parole lue et en parole spontanée. Afin de limiter le biais lié à la variabilité des contextes segmentaux dans lesquels les voyelles apparaissent, les distances inter et intra-locuteur ont été calculées entre voyelles d'une même catégorie apparaissant dans un même contexte segmental, après recodage des contextes en fonction du lieu d'articulation des consonnes et de l'antériorité/postériorité des voyelles (les pauses étant considérées comme une catégorie distincte) et élimination des contextes regroupant moins de huit exemplaires de la voyelle considérée. Ces distances sont alors calculées comme des distances euclidiennes entre centroïdes pour chaque catégorie vocalique et chaque contexte, dans l'espace en 12 dimensions des coefficients MFCC. En conséquence, dans chaque style de parole chaque locuteur est caractérisé par un ensemble de 993 à 2 182 distances entre locuteurs correspondant aux neuf catégories de voyelles multipliées par le nombre de locuteurs de même sexe moins un, multiplié par le nombre de contextes suffisamment représentés pour donner lieu à des comparaisons. De même, pour chaque locuteur l'écart de réalisation des voyelles entre parole lue et spontanée est caractérisé par une distance pour chaque catégorie de voyelle et pour chacun des contextes comparables entre parole lue et parole spontanée.

La Figure 9 représente la distribution des distances inter-locuteurs dans chacun des deux styles de parole et des distances intra-locuteur, pour chaque locuteur et chaque catégorie de voyelle, séparément pour les 11 femmes et les 12 hommes. Les distances inter-locuteurs ont été comparées entre styles de parole, locuteurs et catégories de voyelles à l'aide de modèles linéaires mixtes tenant compte du contexte, séparément pour les femmes et pour les hommes, les interactions entre ces facteurs étant également prises en compte. Afin de tenir compte du déséquilibre observé, avec des contextes représentant 0,007 % à 19 % du nombre total de voyelles en parole lue et de 0,001 % à 18 % en parole spontanée, les distances ont été pondérées par le nombre d'exemplaires pris en compte dans chaque contexte. De façon similaire, les distances intra-locuteur ont été comparées entre locuteurs et catégories de voyelles en prenant également en compte l'interaction entre le locuteur et la catégorie de voyelle.

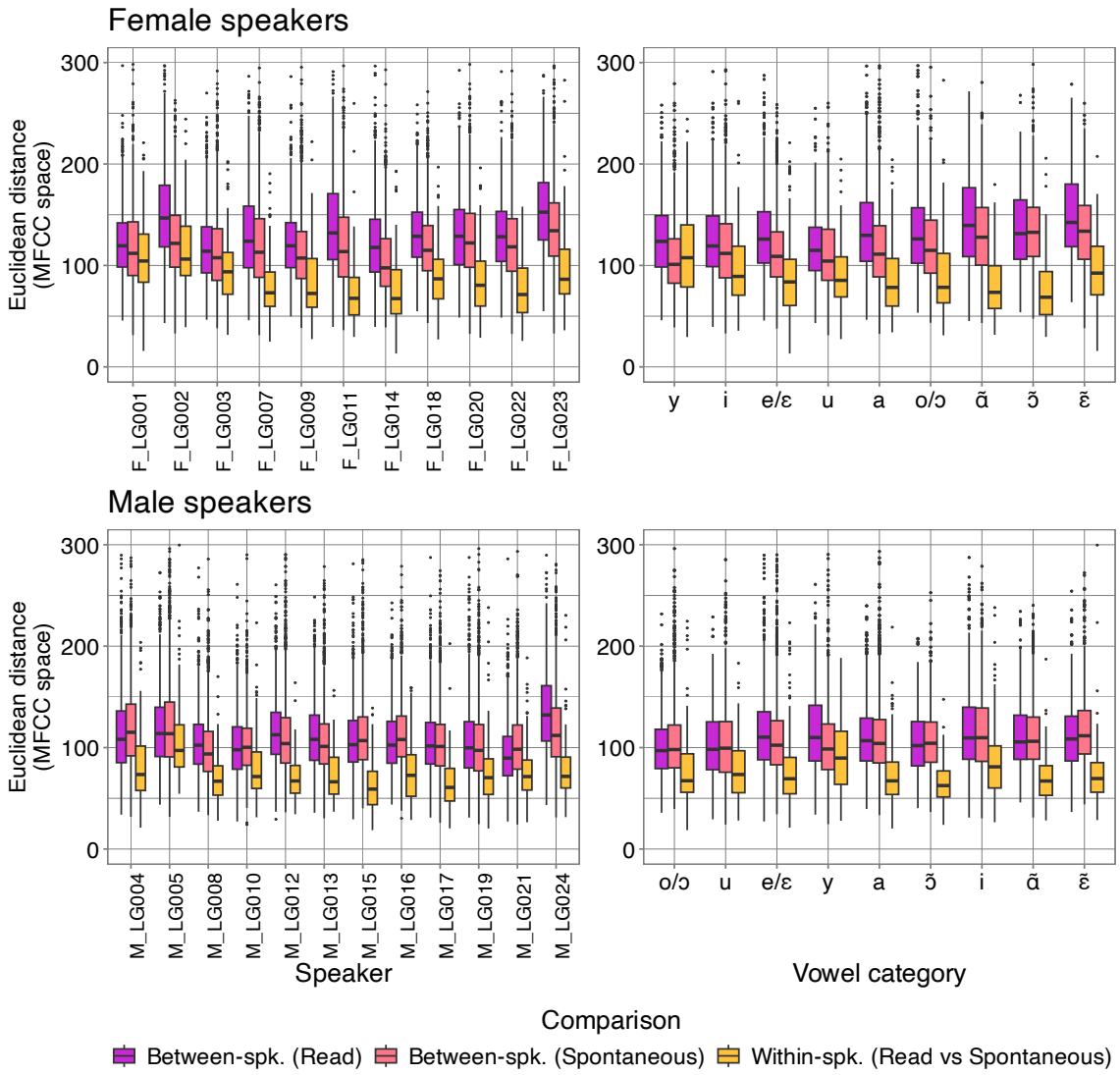


Figure 9 : Distribution des distances euclidiennes dans l'espace à 12 dimensions des coefficients MFCC pour les femmes (haut) et les hommes (bas), par locuteur (gauche) et par voyelle (droite). Pour chaque locuteur ou voyelle, les deux boîtes à moustaches de gauche représentent la distance entre locuteurs (c'est-à-dire la spécificité du locuteur au sein du groupe de locuteurs) en lecture (violet) et spontané (rose). En orange, la distance intra-locuteur (c'est-à-dire la distinction entre le style lu et le style spontané). Sur la droite, les catégories de voyelles sont classées de gauche à droite par ordre croissant des distances entre locuteurs (en moyenne entre lecture et parole spontanée), séparément pour les hommes et les femmes. D'après Audibert et al. (2024, [ACTI1]).

Ces analyses ont montré que les distances entre locuteurs étaient influencées par le style de parole, avec une interaction significative entre style et locuteur : si pour l'ensemble des femmes les distances inter-locuteurs sont plus grandes en lecture qu'en parole spontanée, l'amplitude de la différence est variable d'une locutrice à l'autre. Pour les hommes la différence n'est significative que pour sept locuteurs sur douze, et on observe même une tendance inverse significative pour trois locuteurs qui diffèrent plus des autres en parole spontanée qu'en lecture. Les catégories de voyelles les plus distinctives entre locuteurs dépendent à la fois du style de parole et du sexe des locuteurs, toutefois les trois voyelles nasales émergent comme globalement plus discriminantes que les voyelles orales, la seule exception étant la voyelle /ɔ̃/ en parole lue pour les hommes.

Nous avons également montré dans cette étude que la variabilité en fonction du style de parole, estimée par les distances intra-locuteur, dépend quant à elle à la fois du locuteur et de la voyelle comme l'indique l'interaction significative entre ces deux facteurs. Par ailleurs cette variabilité dépend plus du locuteur que de la voyelle, comme l'indique la comparaison des tailles d'effet estimées par les rapports de vraisemblance. On peut aussi noter qu'à l'exception d'un homme et d'une femme sur les 23 locuteurs pris en compte pour lesquels ce constat ne s'applique qu'en parole lue, la variation entre styles de parole est plus faible que la variation entre locuteurs, à la fois en parole lue et en parole spontanée. Enfin, la variabilité entre styles de parole est inégale selon les voyelles, avec une hiérarchie entre voyelles différente entre hommes et femmes. On retrouve toutefois la voyelle /y/ parmi les voyelles les plus variables aussi bien pour les femmes que pour les hommes.

Dans cette étude, nous avons donc pu montrer à partir de l'analyse de près de 400 000 voyelles produites par 23 locuteurs que les locuteurs sont plus distincts les uns des autres en parole lue qu'en parole spontanée sur le plan de la réalisation des voyelles, avec toutefois d'importantes différences interindividuelles d'adaptation aux changements de styles de parole, qui fait écho à la variabilité des stratégies d'adaptation déjà relevée par Ferguson & Kewley-Port (2007). Bien que la différence de réalisation des voyelles entre styles de parole ne puisse se réduire à une question d'expansion plus moins importante de l'espace vocalique, elle-même liée seulement partiellement au débit de parole (voir par exemple en section 4.2.1 la présentation de résultats plus spécifiques à cette question), cette distinctivité accrue en parole lue pourrait être liée à l'hyperarticulation associée à la parole lue et en partie favorisée par le débit plus lent dans ce style de parole. On peut également s'interroger sur le lien possible entre les distances acoustiques plus importantes observés chez les femmes et la taille plus importante de leur espace acoustique (Whiteside, 2001; Weirich & Simpson, 2013).

Nos travaux ont également confirmé que certaines voyelles reflètent mieux que d'autres les spécificités du locuteur, et qu'en dépit des fluctuations observées entre hommes et femmes et entre styles de parole, les voyelles nasales apparaissent comme relativement plus stables et constitueraient donc de bonnes candidates pour le français et potentiellement dans d'autres langues incluant des voyelles nasales. Ce résultat ne confirme que partiellement les observations d'Ajili et al. (2016, 2017) qui ont conclu à partir de l'analyse des performances d'un système automatique de vérification du locuteur que les voyelles permettaient une meilleure performance de discrimination que les consonnes, mais de façon moindre pour les voyelles nasales comparativement aux voyelles orales.

3 Parole expressive : des émotions aux affects sociaux

Résumé du chapitre 3

Dans ce chapitre je reviens sur les travaux sur la parole expressive auxquels j'ai contribué, à la fois sur les expressions d'attitudes et sur les expressions émotionnelles.

Une étude expérimentale de l'effet Kuleshov à partir d'extraits de parole exprimant de la colère ou neutres délexicalisés avec ou sans préservation de la qualité de voix a montré qu'une expression faciale de colère présentée préalablement influence le traitement affectif des expressions vocales lorsque celles-ci sont ambiguës.

La comparaison des jugements dimensionnels attribués à des expressions neutres contextualisées ou non a montré que les expressions non-contextualisées sont jugées différemment de celles produites en contexte en raison notamment d'un débit plus lent. Une étude du degré d'hésitation exprimé dans la parole spontanée a montré que l'hésitation perçue dépend principalement du nombre de pauses pleines et l'allongement des voyelles, mais avec une importante variabilité individuelle. L'évaluation de la perception par des apprenants francophones d'expressions audiovisuelles d'attitudes du chinois mandarin a suggéré un recours plus important aux indices visuels que chez les auditeurs natifs, en particulier dans le cas d'expressions de politesse.

Une étude d'expressions d'attitudes agressives simulées par des acteurs sur des énoncés contrôlés a conclu à une augmentation de f0 et du F1 de /a/ avec le degré d'agressivité perçu indépendamment du contexte consonantique, mais des stratégies individuelles divergentes en termes de tension laryngée. Dans le cadre de la thèse de C. Kouklia, l'analyse d'attitudes hostiles produites dans une situation de débat politique et comparées à une version de contrôle produite a posteriori par les locuteurs a suggéré quant à elle que l'hostilité perçue dépendrait à la fois d'écarts locaux au modèle d'accentuation attendu en français et de l'insertion de pauses silencieuses.

Suite aux travaux que j'ai réalisés dans le cadre de ma thèse et dans le prolongement de celle-ci sur les expressions des émotions dans la parole, mes travaux sur la parole expressive n'ont représenté qu'une part relativement modeste de mon activité scientifique, principalement à travers la codirection avec Jacqueline Vaissière de la thèse de Charlotte Kouklia et mes collaborations avec elle sur des thématiques liées aux expressions d'affects. Ces travaux m'ont conduit progressivement à passer des émotions aux expressions d'attitudes, de nature plus linguistique que les émotions qui sont de nature plus biologiques (voir par exemple Daneš (1994)), et autre affects dits « sociaux » (Buck, 1999) dont l'expression est spécifique aux interactions entre humains. De telles expressions sont beaucoup plus fréquentes dans la communication parlée que les expressions émotionnelles à proprement parler, considérées par la plupart des auteurs en psychologie des émotions comme issues de l'évolution (voir Sander & Scherer (2009) pour une discussion approfondie sur ce point) et souvent très informatives lorsqu'elles surviennent mais aussi plus rares dans la communication face-à-face que ne le sont les affects sociaux.

Plus récemment, dans le cadre de collaborations avec Caterina Petrone et d'autres collègues du Laboratoire Parole et Langage d'Aix-en-Provence, je suis revenu à des questions liées aux expressions d'émotions et notamment à celle de la colère, qui incluent également une étude préliminaire menée auprès de patients fortement dysphoniques que je présente en

section 5.2.2 dans le chapitre de ce document dédié à la phonétique clinique. Je propose ici une organisation thématique qui part de mes travaux les plus directement liés aux expressions émotionnelles avant d'ouvrir vers l'expression d'autres types d'affects.

3.1 Interaction entre expression faciale et prosodique de la colère

Publication associée :

[ACL2] Petrone, C., Carbone, F., **Audibert, N.**, Champagne-Lavau, M. (2024). Facial cues to anger affect meaning interpretation of subsequent spoken prosody. *Language & Cognition.* Published online 2024:1-24. doi:10.1017/langcog.2024.3

Dans une étude en collaboration avec Caterina Petrone et Maud Champagne-Lavau, ainsi que Francesca Carbone qui était post-doctorante au Laboratoire Parole et Langage d'Aix-en-Provence lors de la mise en place de la majorité des étapes de conception et d'analyse, nous avons cherché à déterminer dans quelle mesure l'expression faciale influence le traitement affectif de la parole perçue. Pour les besoins de cette étude, j'ai développé une méthode de délexicalisation de la parole expressive visant à préserver autant que possible les caractéristiques prosodiques et spectrales des énoncés originaux, présentée en section 7.2.4.

L'étude que nous avons proposé s'appuie sur les travaux qui montrent que la perception d'une expression faciale associée à une émotion conditionne les attentes de l'auditeur en termes d'expression vocale de l'émotion (Jessen & Kotz, 2013), ainsi que sur « l'effet Kuleshov », nommé ainsi d'après l'œuvre du réalisateur russe Lev Kuleshov dans laquelle des expressions faciales neutres ont été associées à des scènes émotionnelles de nature diverses afin d'influencer la perception par le public des émotions exprimées par les visages. Cet effet consiste en une exploitation plus importante des indices contextuels afin de parvenir à une interprétation émotionnelle lorsque l'expression émotionnelle produite par l'interlocuteur, ou plus généralement le sujet humain en présence duquel l'observateur se trouve, est ambiguë. L'effet Kuleshov a été mis en évidence dans la littérature en psychologie des émotions à partir d'expressions faciales dégradées ou neutres (voir par exemple Calbi et al. (2017)) mais qui n'avait pas été évalué jusqu'alors sur des expressions vocales ambigües.

La série d'expériences que nous avons mises en place pour évaluer dans quelle mesure cet effet peut également être observé lorsque des expressions vocales ambigües en termes de valence émotionnelle s'est appuyé sur le paradigme classique de l'amorçage intermodal, dans lequel l'amorce et la cible sont présentées selon deux modalités distinctes, en l'occurrence visuelle et auditive. Contrairement aux études précédentes de l'effet Kuleshov mettant en œuvre des expressions vocales d'émotions, dans lesquelles la clarté de l'émotion associée à l'expression faciale a été manipulée expérimentalement en maintenant une expression vocale clairement émotionnelle, nous avons évalué dans quelle mesure l'influence de l'expression faciale présentée auparavant augmente lorsque l'émotion véhiculée par le stimulus audio présenté ensuite est moins clairement identifiable.

Nous avons sélectionné des expressions de colère, considérées dans une perspective évolutionniste comme à même d'informer l'entourage d'un danger potentiel et ainsi d'attirer leur attention vers un élément menaçant, à la fois volontairement et involontairement (Aue et al., 2011), ainsi que des expressions jugées neutres produites par les mêmes locuteurs dans un contexte comparable. Ces expressions ont été sélectionnées dans des films français et validées perceptivement après délexicalisation selon les principes de la méthode présentée en

section 7.2.4, afin de s'assurer qu'elles soient bien perçues comme exprimant de la colère ou comme neutres. Neuf stimuli exprimant de la colère, appariés avec neuf stimuli neutres produits par les mêmes acteurs, ont ainsi été sélectionnés à partir d'un ensemble plus large. La procédure de délexicalisation a été appliquée en deux versions : en appliquant uniquement la première étape d'analyse-resynthèse par diphones avec substitution de phones et copie prosodique selon la méthode de Pagel et al. (1996), pour générer la condition étiquetée ci-après Morphing–, ou en utilisant la méthode complète dans laquelle une partie de l'information spectrale est reconstruite par morphing vocal suite à cette première étape (condition Morphing+). L'efficacité de la délexicalisation ayant par ailleurs été validée sur ces stimuli selon les principes exposés en section 7.2.4, leur utilisation permettait de s'assurer que l'attribution émotionnelle ne repose que sur les informations prosodiques et de qualité de voix ainsi que sur le biais éventuel induit par l'information visuelle, mais pas sur des indices lexicaux.

Deux expériences ont été menées pour cela à partir des stimuli audio sélectionnés et délexicalisés, et d'expressions faciales issues de la banque d'images standardisée « Karolinska Directed Emotional Faces database » (Lundqvist et al., 1998). Dans la première expérience, des expressions vocales de colère et des expressions neutres ont été présentées seules ou à la suite d'une expression faciale congruente, sélectionnée en fonction du sexe de l'acteur ayant produit le stimulus audio. La tâche soumise aux 200 participants francophones natifs était d'évaluer les stimuli audio présentés selon les deux dimensions les plus couramment utilisée pour l'annotation dimensionnelle des expressions émotionnelles (Russell, 2003) : une échelle de valence sur laquelle les participants devaient juger les expressions comme plus ou moins positives ou négatives, et une échelle d'intensité de l'émotion, analogue dans ce contexte à une échelle d'activation, sur laquelle les expressions devaient être jugées comme exprimant une émotion plus ou moins intense.

L'hypothèse était que les expressions de colère précédées d'une expression faciale de colère seraient jugées comme plus négatives et plus intenses comparativement à la présentation isolée des mêmes stimuli audio, avec un effet renforcé dans le cas de stimuli audio dans lesquelles l'information émotionnelle est appauvrie, qui conduirait les sujets à exploiter plus activement les informations issues des expressions faciales (de Gelder et al., 2006). Cette hypothèse a été partiellement validée, l'effet de l'amorce visuelle sur les jugements de valence et d'intensité émotionnelle n'étant présent que dans la condition Morphing– dans laquelle seule une partie des indices acoustiques sont conservés, confirmant ainsi le rôle prépondérant des informations spectrales de qualité de voix pour véhiculer l'information émotionnelle (voir par exemple Gobl & Ní Chasaide (2003)). En effet dans cette tâche, les stimuli présentés en condition Morphing+ comprenaient suffisamment d'informations spectrales permettant une attribution émotionnelle sans ambiguïté pour « bloquer » l'effet Kuleshov dans les conditions incluant la présentation de l'amorce visuelle.

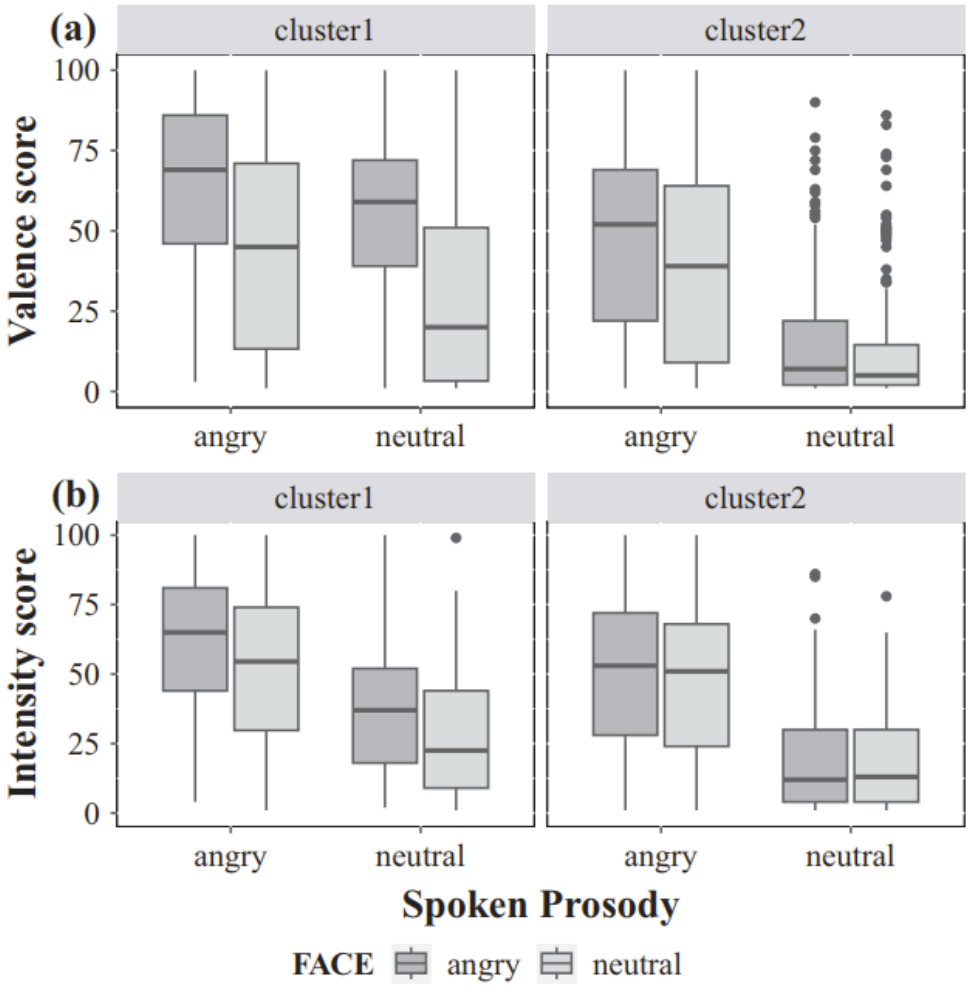


Figure 10 : Distribution des scores de valence (a) et d'intensité émotionnelle (b) attribués par les 67 participants à la deuxième expérience lors de l'évaluation de stimuli audio délexicalisés en condition Morphing-, exprimant de la colère ou une expression neutre et précédés de l'image d'une expression faciale de colère ou neutre, dans chacun des deux groupes (étiquetés cluster1 et cluster2) obtenus par regroupement hiérarchique. D'après Petrone et al. (2024, [ACL2])

Dans la seconde expérience à laquelle 67 participants francophones natifs ont participé, la tâche a été recentrée sur l'utilisation de la condition Morphing– qui était la seule à donner lieu à un effet de l'amorce visuelle. Par ailleurs l'amorce était toujours présentée mais consistait soit en une expression faciale congruente avec la catégorie émotionnelle (colère ou neutre) du stimulus audio, soit en une expression faciale incongruente. Afin d'explorer les éventuels patrons de variabilité entre auditeurs en termes d'exploitation des informations tirées des expressions faciales, un regroupement hiérarchique a été effectué à partir des jugements de valence attribués aux stimuli audio supposés neutres, susceptibles de donner lieu à plus de variation inter-individuelle d'après les résultats d'une étude précédente (Mullennix et al., 2019). Le nombre optimal de deux clusters sur ces données a été déterminé à l'aide de la statistique Gap (Tibshirani et al., 2002) qui cherche à minimiser la variabilité intra-classe via un ensemble de simulations considérant plusieurs nombres de classes candidats. L'appartenance à l'un de ces deux clusters, qui comprenaient respectivement 31 et 36 participants et étaient équivalents en âge et relativement à la répartition entre hommes et femmes, a été prise en compte dans l'analyse des résultats à des fins de comparaisons. La Figure 10 présente la

distribution des jugements de valence et d'intensité émotionnelle en fonction de la nature de l'amorce et de la cible dans chacun de ces deux clusters.

Comme illustré par cette figure, les participants du premier cluster ont été plus sensibles à l'effet Kuleshov et ont jugé non seulement les stimuli audio de colère comme plus négatifs et exprimant une émotion plus intense lorsqu'ils étaient précédés de l'image d'une expression de colère, mais également les stimuli audio neutres, tandis que seuls les jugements de valence des participants du second cluster ont été influencés par les expressions faciales de colère présentés avant une expression vocale de colère. Globalement, les résultats obtenus dans ces deux expériences ont donc confirmé les résultats d'études précédentes (voir par exemple Garrido-Vasquez et al. (2018)) avec une influence d'autant plus importante de l'expression faciale présentée comme amorce qu'amorce et cible sont émotionnellement congruent, mais ont aussi mis en évidence des patrons variables entre groupes d'auditeurs, avec un groupe pour lequel l'effet est tout aussi fort en condition incongruente.

3.2 Quel statut pour l'expression émotionnellement « neutre » ?

Publication et communication associée :

[ACTI39] Kouklia, C., & **Audibert, N.** (2013). Expressivity conveyed by contextualized vs. non-contextualized "neutral" acted speech: which control condition for expressive speech modeling? *Proceedings of the First International Workhop on Affective Social Speech Signals* (online proceedings).

[AFF1] Kouklia, C., **Audibert, N.**, Hallé, P., & Isel, F. (2013). Influence du contexte de production sur l'évaluation multidimensionnelle d'énoncés de parole émotionnellement « neutres ». Quelle condition de contrôle ? *Séminaire de l'IUPDP*, Boulogne-Billancourt, France.

En collaboration avec Charlotte Kouklia dans un travail commun réalisé en marge des problématiques de sa thèse, nous nous sommes penchés sur la question de l'attribution émotionnelle, c'est-à-dire le processus par lequel les auditeurs associent à une expression vocale ou faciale une évaluation émotionnelle ou plus généralement affective, dans le cas particulier de l'expression supposée neutre. En effet, dans de nombreux travaux sur la parole expressive, une condition neutre est incluse comme référence à des fins de comparaison acoustique ou perceptive avec des expressions d'émotions ou autres affects, comme cela a par exemple été le cas dans l'étude présentée dans la section précédente. Par ailleurs bien que cette approche soit désormais moins massivement utilisée avec l'avènement des systèmes de synthèse de la parole reposant sur une modélisation par réseaux de neurones, une telle condition neutre est souvent également considérée pour certaines applications des technologies de parole expressive qui reposent sur des méthodes de conversion de voix, dans lesquelles la modélisation porte généralement sur les modifications du signal de parole qui permettent le passage d'une voix neutre à une cible expressive. Dans ces différentes approches, une définition claire de ce qu'est une expression neutre sur le plan émotionnel ou affectif, ou à défaut une certaine stabilité parmi les échantillons de parole considérés comme relavant de cette catégorie, est nécessaire.

Pourtant et en dépit de la généralisation des approches basées sur des scénarios (Enos & Hirschberg, 2006) pour améliorer le naturel des productions des acteurs fréquemment sollicités pour obtenir des échantillons de parole expressive en contrôlant le contenu des énoncés et les conditions d'enregistrement (voir Bänziger & Scherer (2007) ou Busso et al.

(2008) pour deux exemples de corpus de parole émotionnelle ayant eu recours à ce type de méthode et à la base de nombreux travaux), le cas de l'expression neutre est resté assez largement négligé. La méthodologie typique d'élicitation de l'état neutre consiste à demander aux acteurs de produire de façon neutre les phrases utilisées pour les productions émotionnelles, sans définir de scénarios dédiés à un état neutre. Cette méthodologie peut induire un biais dans la collecte de la parole neutre, car les acteurs qui suivent les instructions ont tendance à produire un style de parole très contrôlé et peu naturel, s'apparentant plus à la parole lue de laboratoire qu'à la parole telle qu'elle est produite dans un contexte écologique d'interaction (Bänziger et al., 2012). En raison de ce constat, Bänziger & Scherer (2007) sont allés jusqu'à recommander de ne pas utiliser de référence neutre dans les études portant sur la parole expressive, ce qui n'est toutefois pas toujours applicable.

Dans cette étude, nous avons cherché à évaluer l'incidence de la contextualisation ou non de productions affectivement neutres sur le plan à la fois perceptif et acoustique. Nous avons utilisé pour cela des productions de la phrase « Il est huit heures », contextualisée en s'inspirant des contextes proposés dans l'étude originale de Fónagy & Bérard (1972), complétés en suivant les préconisations de Scherer (2003) afin d'obtenir des degrés d'activation variables pour chacune des quatre classes d'émotions considérées : peur, tristesse, joie et colère. En complément, deux versions de l'expression neutre ont été produites, dans une condition interactive de dialogue simulé en réponse à la question « Quelle heure est-il ? », considérée ci-après comme condition contextualisée, et dans une condition non-contextualisée proche de celle fréquemment utilisée dans les corpus de parole expressive incluant une condition neutre, dans laquelle la seule instruction donnée aux locuteurs était simplement « d'être neutre ». Quatre acteurs semi-professionnels francophones natifs (trois hommes et une femme, âgés de 23 à 32 ans) ont été recrutés pour produire ces énoncés avec les différentes cibles expressives. Deux semaines avant la session d'enregistrement réalisée en chambre sourde, un livret de courts scénarios a été remis aux acteurs, afin de leur permettre de préparer leur performance pour chaque catégorie émotionnelle devant être produite. Lors de la séance d'enregistrement chaque scénario a été relu à voix haute, avant de laisser les acteurs produire autant de répétitions que souhaité et d'en sélectionner deux jugées satisfaisantes pour chacun des scénarios proposés.

Bien que des niveaux d'activation plus importants aient également été produits pour chacune des quatre classes émotionnelles, pour les besoins de cette étude nous avons sélectionné les versions faiblement activées, désignées comme anxiété, joie calme, irritation et dépression, ainsi que les deux versions de l'expression neutre. Les deux répétitions sélectionnées par les acteurs ont été retenues pour chaque acteur et chacune des six catégories. Ces 48 stimuli ont été évalués par 23 auditeurs francophones natifs sur les trois dimensions classiquement utilisées pour la description dimensionnelle des émotions (Fontaine et al., 2007) : la valence qui correspond à la positivité/négativité de l'affect via une échelle malheureux-heureux (d'après Pereira (2000)), l'activation qui dans le contexte de l'évaluation perceptive est fortement liée à l'intensité perçue de la réaction émotionnelle selon une échelle calme-agité, et la dimension dite de « pouvoir » ou de dominance qui désigne le contrôle que le locuteur a de la situation selon une échelle soumis-dominant.

Les résultats, illustrés par la Figure 11, ont indiqué que si les deux versions de l'expression neutre sont jugées avec un niveau équivalent de valence et d'activation, tandis que les expressions relevant des quatre classes émotionnelles se distinguent entre elles et par rapport aux expressions neutres par au moins deux dimensions, la version non-contextualisée de

l'expression neutre est jugée significativement plus soumise que la version contextualisée, à un niveau équivalent à celui de l'expression de dépression.

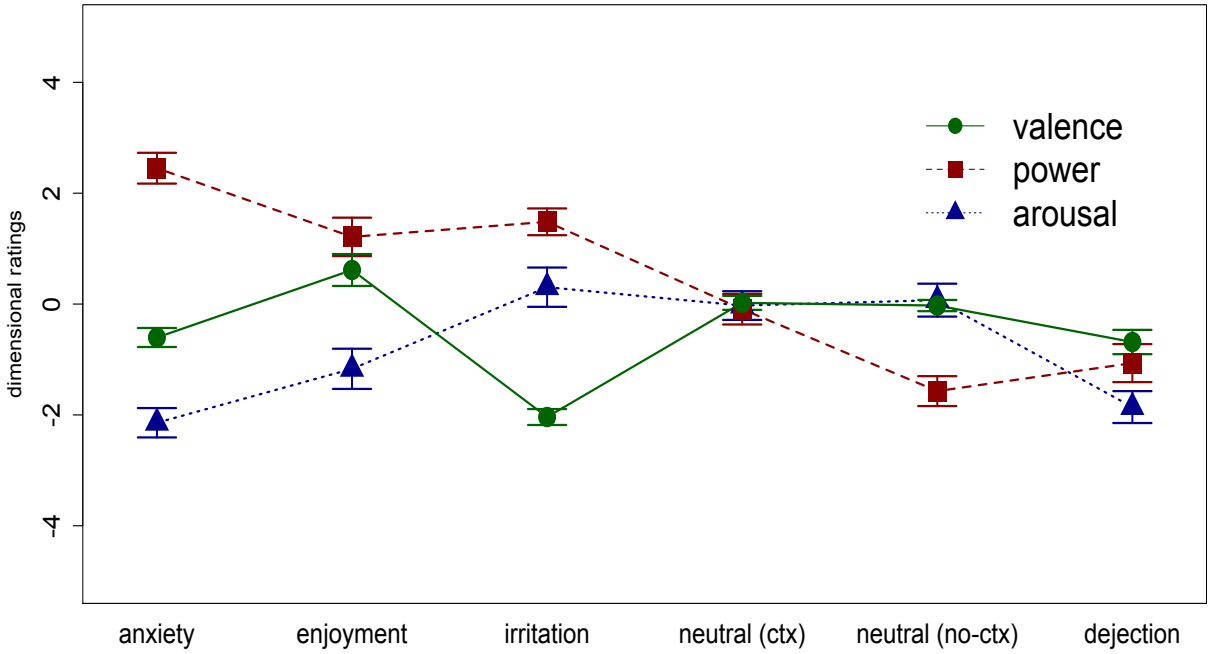


Figure 11 : Jugements moyens et intervalles de confiance à 95% attribués aux productions des quatre acteurs par les 23 auditeurs ayant participé à l'évaluation des 48 stimuli, pour chacune des six attitudes considérées, selon les dimensions affectives de valence, dominance et excitation. D'après Koulia & Audibert (2013, [ACTI39]).

Après une étape de segmentation en phones corrigée manuellement, un ensemble de mesures acoustiques ont été extraites de chacun des énoncés, et comparées entre les deux versions de l'expression neutre afin d'évaluer quelles dimensions acoustiques permettent le mieux d'expliquer les différences observées. Les variations de fréquence fondamentale ont ainsi été caractérisées par des mesures de niveau moyen dans l'énoncé, d'étendue de variation et de pente entre le début et la fin de l'énoncé obtenu au moyen d'une régression linéaire, et complétées par une mesure de débit de parole (équivalent dans ces données au débit articulatoire en l'absence de pause) et du coefficient Varco appliqué aux durées des voyelles suivant White & Mattys (2007). Parmi ces mesures, le débit de parole s'est avéré être la seule mesure acoustique significativement distincte entre les deux versions de l'expression neutre, avec un débit plus lent pour la version non-contextualisée qui reflète les similitudes entre cette version et la parole lue. Bien que la pente de la déclinaison de la fréquence fondamentale dans l'énoncé n'ait pas été significativement différente entre l'expression neutre contextualisée ou non, nous avons pu noter qu'elle était légèrement plus abrute dans la version non-contextualisée, ce qui est également consistant avec les descriptions dans la littérature des différences entre parole lue et parole conversationnelle (Laan, 1997; Meunier & Espesser, 2011). On peut également noter que le débit observé dans l'expression neutre non-contextualisée est équivalent à celui de l'expression de dépression. La mise en relation du débit de parole avec les jugements moyens de dominance attribués à chacun des énoncés supposés neutres indique de plus une forte corrélation entre débit de parole et jugements de dominance parmi les productions de chaque locuteur, jugées d'autant plus soumises qu'elles sont produites avec un débit lent.

Bien qu'on puisse supposer que dans une évaluation en choix forcé les deux versions de l'expression neutre auraient été catégorisées comme neutre, l'attribution émotionnelle dimensionnelle à un énoncé supposé neutre est donc fortement influencée par le caractère interactif ou non du cadre dans lequel cet énoncé est produit. Lorsque l'objectif du recueil de données de parole affectivement neutre est d'obtenir des échantillons aussi représentatifs que possible de la parole conversationnelle considérée comme neutre ou de comparer ces échantillons à des productions expressives produites en contexte, une recommandation directe tirée de nos résultats serait donc d'étendre l'utilisation de scénarios d'interaction aux productions neutres. Cette conclusion converge avec les impressions des acteurs recueillies à la suite de l'enregistrement, en effet les quatre acteurs ont considéré la condition contextualisée qui leur permettait de s'appuyer sur une situation de communication réaliste comme une condition d'élicitation plus naturelle.

3.3 Expression de l'hésitation en parole spontanée

Publication associée :

[ACTI25] Wottawa, J., Tahon, M., Marin, A., & **Audibert, N.** (2020). Towards Interactive Annotation for Hesitation in Conversational Speech. *Proceedings of LREC 2020*, Marseille, France, pp. 1526-1532.

En collaboration avec Jane Wottawa et Marie Tahon du Laboratoire d'Informatique de l'Université du Mans avec qui j'ai coencadré le mémoire de master d'Apolline Marin, nous avons abordé la question des corrélats acoustiques de l'expression de l'hésitation en parole spontanée, à partir de données issues du corpus de parole conversationnelle NCCFr (Torreira et al., 2010). Plus spécifiquement, nous objectif a été de positionner des extraits de parole sur un continuum entre l'hésitation et la confiance en soi, conceptuellement proche de la dimension de contrôle telle que définie par Scherer (2005).

Parmi les études de l'hésitation dans la parole spontanée, une grande part se sont concentrées sur la distribution et la durée des pauses silencieuses et des pauses pleines (voir par exemple Maclay & Osgood (1959) sur l'anglais américain). Dans une moindre mesure, l'allongement syllabique a également été décrit comme un corrélat acoustique de l'hésitation en parole spontanée, par exemple par Duez (2001) sur le français. Les résultats relatifs aux liens entre la perception de l'hésitation et les autres dimensions de la réalisation acoustique de la parole sont moins tranchés. Ainsi en allemand, Mixdorff & Pfitzinger (2005) n'ont relevé aucune incidence des hésitations marquées par la présence de pauses pleines sur la modélisation du contour de fréquence fondamentale au niveau de l'énoncé. A l'inverse, Carlson et al. (2006) ont mis en évidence en suédois un effet modéré de la pente de la fréquence fondamentale et de la présence de voix craquée sur l'hésitation perçue, et en anglais américain Pon-Barry & Schieber (2011) ont conclu à un lien entre pente et étendue de la fréquence fondamentale d'une part et évaluation de l'hésitation d'autre part, quoique plus faible que le lien avec la durée des pauses silencieuses et le débit de parole. En français spontané, l'hésitation a été décrite comme se manifestant notamment par l'insertion de pauses pleines généralement réalisées comme [œ] ou [ø] (Vasilescu et al., 2004), notées « euh » ci-après, et par un allongement porté par la syllabe finale de mot mais qui peut toucher plusieurs mots au sein d'une même séquence (Candea, 2000).

Une annotation manuelle a été réalisée sur un sous ensemble des productions de 32 locuteurs du corpus NCCFr de parole conversationnelle (Torreira et al., 2010), alignées automatiquement en phones et segmentées en unités inter-pausales en considérant les pauses d'une durée supérieure à 200 ms comme délimitant ces unités, la segmentation étant ensuite corrigée manuellement si nécessaire. Pour chacun de locuteurs sélectionnés, les dix premières minutes d'enregistrement ont été annotées, représentant un total de 5 834 unités inter-pausales. Le degré d'hésitation a été évalué par une annotatrice sur une échelle de Likert à sept points allant de « très sûr de soi » (-3) à « très hésitant » (+3). Sur la base de la distribution des degrés d'hésitation attribués, ceux-ci ont été regroupés en trois niveaux d'hésitation (sûr, neutre et hésitant) pour certaines analyses, ces trois niveaux étant respectivement représentés en vert, bleu et orange dans les figures ci-dessous. Les dimensions émotionnelles de valence, activation et dominance ont également été cotées sur des échelles de Likert à cinq points afin d'évaluer une éventuelle interaction entre affects exprimés et degré d'hésitation, la formulation des échelles émotionnelles étant inspirée de Perreira (2000), toutefois ces dimensions n'ont pas été exploitées en raison d'un faible degré de consistance dans les choix de cotation de l'annotatrice.

De multiples paramètres (232 au total) ont été extraits automatiquement sur chaque unité inter-pausale, à partir à la fois d'informations symboliques et de durée issues de l'alignement forcé, et du signal acoustique. Ces paramètres incluaient notamment des mesures temporelles et rythmiques, dont certaines relatives aux pauses silencieuses et aux pauses pleines étiquetées comme telles dans la transcription orthographique du corpus NCCFr, des mesures dérivées de la fréquence fondamentale et de sa distribution dans chaque unité inter-pausale, ainsi que des mesures spectrales au niveau de l'ensemble des segments voisés de l'unité inter-pausale ou spécifiques à certains phones, supposées capturer des corrélats acoustique de la qualité de voix ou de l'articulation segmentale.

Les principaux résultats mis en évidence par la mise en relation de l'annotation du degré d'hésitation et les mesures acoustiques concernent le lien important entre la présence de pauses pleines transcrites comme « euh » et le degré d'hésitation, comme illustré par la Figure 12, la durée moyenne de ces pauses pleines étant par ailleurs d'autant plus importante que le degré d'hésitation perçu était élevé.

Bien qu'aucune distinction n'ait été notée entre le niveau « sûr » (vert) et le niveau « neutre » (bleu), les degrés d'hésitation regroupés dans le niveau « hésitant » (orange) sont associés à une importante augmentation de la durée des voyelles qui ne s'explique pas uniquement pas celle des pauses pleines, de façon relativement consistante entre locuteurs comme illustré par la Figure 13. Dans une moindre mesure, le débit articulatoire est également lié au degré d'hésitation avec une tendance au ralentissement du débit dans les unités inter-pausales jugées comme exprimées avec un degré important d'hésitation, avec toutefois une plus grande variabilité inter-individuelle et un biais lié à la longueur des unités inter-pausales, les plus courtes tendant à être produites avec un débit relativement lent sans pour autant être associées à l'hésitation. En revanche l'analyse de la durée des pauses silencieuses précédant les unités inter-pausales ne permet pas de tirer de conclusions quant aux liens entre pauses silencieuses et hésitation, contrairement à certains résultats de la littérature.

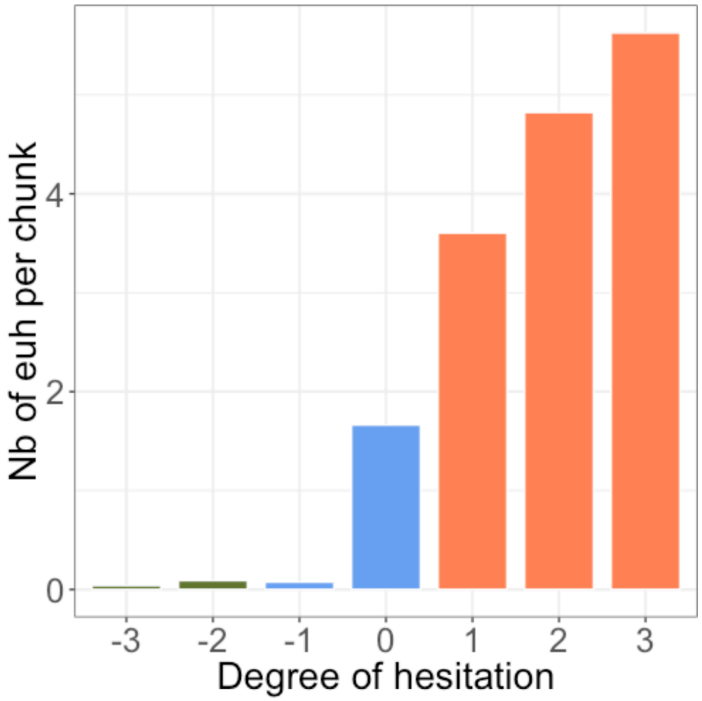


Figure 12 : Distribution du nombre de pauses pleines étiquetées « euh » en fonction du degré d'hésitation annoté dans chacune des 5834 unités inter-pausales. D'après Wottawa et al. (2020, [ACTI25]).

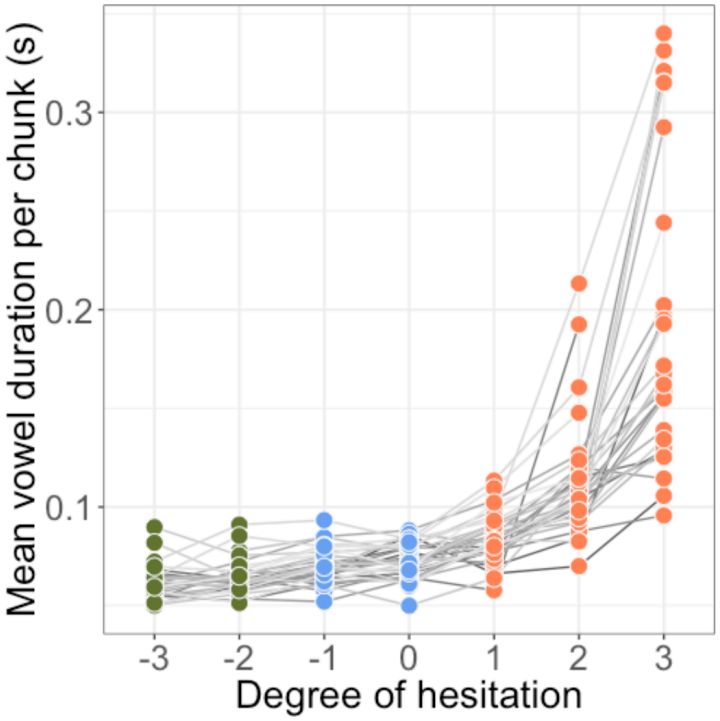


Figure 13 : Distribution des mesures de durée moyenne des voyelles en secondes en fonction du degré d'hésitation annoté dans chacune des 5 834 unités inter-pausales. Chaque ensemble de points connecté par des segments représente l'un des locuteurs du corpus NCCFr analysés. D'après Wottawa et al. (2020, [ACTI25]).

La fréquence fondamentale produite sur la voyelle de la syllabe finale tend à être plus basse dans les productions jugées plus hésitantes, toutefois cet effet reste modéré et variable entre locuteurs. Si cette tendance semble se vérifier dans les énoncés déclaratifs, dans l'optique d'une application à l'annotation automatique de l'hésitation et en l'absence d'une référence permettant de distinguer entre modalités voire entre attitudes (certaines pouvant être associées à une f0 montante, voir par exemple Morlec (2001)), cette mesure apparait donc comme insuffisamment robuste pour fournir des résultats directement interprétables. Les autres mesures acoustiques considérées n'ont pas permis d'identifier de liens avec le degré d'hésitation annoté.

L'applicabilité à l'annotation automatique du degré d'hésitation de ces mesures extraites du signal acoustique a été évaluée à l'aide de modèles de régression, choisis en raison de la possibilité de les entraîner sur un jeu de données de taille modeste comme celui dont nous disposions, en considérant les trois classes de degré d'hésitation obtenues par regroupement. Pour cela, les unités inter-pausales annotées de 30 des 32 locuteurs ont été utilisées comme base d'apprentissage, celles des deux locuteurs restants étant utilisées comme base de test afin d'évaluer la possibilité de généraliser à des locuteurs inconnus les patrons acoustiques associés à l'expression de l'hésitation. Les mêmes modèles de régression ont été entraînés à partir de l'ensemble de mesures dérivées de l'alignement et du signal acoustique, et à partir de coefficients MFCC extraits sur des trames successives de signal acoustique, complétés par leur dérivée première, et de leur distribution à l'intérieur de chaque unité inter-pausale.

Les performances estimées à partir de l'erreur RMSE ont été sensiblement meilleures pour les modèles entraînés sur les mesures extraites de l'annotation et du signal, confirmant dans une telle tâche l'intérêt de ce type de mesures, qui présentent également l'avantage d'être plus directement interprétables. L'application d'une méthode de sélection de variables a indiqué que les mesures qui contribuent le plus à la classification étaient la durée de la voyelle la plus longue dans l'unité, la proportion de voyelles /œ/ et /ø/, la durée du premier phone, et dans une moindre mesure le rapport entre énergie harmonique et bruit HNR sur les voyelles et la fréquence fondamentale de la dernière voyelle. Ces résultats ont donc majoritairement convergé, quoique parfois indirectement, avec les résultats de la première analyse.

Bien que cette mesure ne soit pas extractible automatiquement, on peut également noter que la divergence entre unités inter-pausales segmentées automatiquement et la correction manuelle de ces unités inter-pausales effectuée à partir de l'écoute de ces extraits a également contribué de façon conséquente à la classification, ce qui suggère une interaction entre la perception de l'hésitation et la perception de ce qui constitue une pause silencieuse dans la parole.

On bien enfin noter que bien les performances obtenues avec le meilleur modèle (qui s'appuie sur le principe du Support Vecteur Machine, qui reste largement utilisé par ailleurs pour la classification automatique des émotions) soient prometteuses, une approche indépendante du locuteur de l'annotation de l'hésitation ne semble pas optimale, et qu'une approche tenant compte des stratégies individuelles d'expression de l'hésitation apparait comme plus adaptée au vu de la variabilité inter-individuelle conséquente observée au-delà de la durée des voyelles.

3.4 Perception par des francophones des attitudes du chinois mandarin

Publication associée :

[ACTI38] Lu, Y., Aubergé, V., **Audibert, N.**, & Rilliard, A. (2014). Audiovisual Perception of Expressions of Mandarin Chinese social affects by French L2 Learners. *Proceedings of the 7th International Conference on Speech Prosody (Speech Prosody 2014)*, Dublin, Ireland, pp. 169-173.

En collaboration avec Yan Lu et ses encadrants Véronique Aubergé et Albert Rilliard dans le cadre de sa thèse, j'ai contribué à une étude portant sur la perception audiovisuelle par des apprenants francophones du chinois mandarin d'attitudes produites par des locuteurs natifs, selon une méthodologie similaire à certains des travaux menés dans le cadre de ma thèse mais avec l'objectif d'exploiter ces résultats pour des applications en didactique L2 des affects sociaux. En effets contrairement aux expressions émotionnelles qui partagent plus de caractéristiques communes entre langues et cultures, les attitudes prosodiques sont susceptibles d'être beaucoup plus spécifiques à une langue ou une culture et de donner lieu à des situations d'incompréhension, comme par exemple Shochi et al. (2006) l'ont montré entre le japonais et le français, notamment pour l'expression de la politesse. Ma contribution à cette étude s'étant limitée à des conseils méthodologiques pour la mise en place du protocole d'évaluation perceptive et l'analyse statistique, je me contenterai ici de résumer brièvement les choix effectués et les principaux résultats obtenus.

Les données évaluées ont été extraites d'un corpus dans lequel 19 modalités, attitudes, et styles de parole susceptibles de varier entre langues et cultures ont été produits sur des énoncés sémantiquement neutres d'une à neuf syllabes par une locutrice native du chinois mandarin enregistrée en chambre sourde en modalité audio et vidéo de face avec un cadrage permettant de voir à la fois le visage et le buste de la locutrice. Parmi ces productions, 11 expressions produites sur la phrase de quatre syllabes « 四天三夜 » (si4 tian1 san1 ye4, « quatre jours et trois nuits ») ont été sélectionnées en raison du taux de confusion plus élevé que pour les autres attitudes observé préalablement lors de l'évaluation par des auditeurs natifs et non-natifs en modalité audio seule, en dépit de l'identification par les natifs à un niveau significativement supérieur au hasard : déclaration, question, irritation, doute, mépris, déception, évidence, surprise neutre, politesse, autorité, et parole dirigée vers l'enfant.

Trois groupes de sujets ont été mis à contribution, dans une tâche de sélection en choix forcé parmi ces 11 expressions : un groupe de 30 sinophones natifs dont les résultats ont été utilisés comme référence, un groupe de 9 apprenants francophones du chinois de niveau débutant A1 ayant suivi moins de 100 heures d'enseignement, et un groupe de 10 apprenants francophones de niveau A2 ayant suivi moins de 200 heures d'enseignement, les deux groupes d'apprenants étant composés d'étudiants suivant des cours de mandarin dans la même structure universitaire. Dans chacun des groupes, la moitié des sujets, sélectionnés aléatoirement, évaluaient les stimuli en condition audio seule avant d'évaluer les stimuli en condition vidéo seule, tandis que l'autre moitié devait évaluer vidéo puis audio, la condition audiovisuelle étant présentée en dernier dans les deux groupes.

L'analyse des matrices de confusion dans les différents groupes de sujets et dans les différentes conditions a indiqué que la majorité des expressions étaient mieux reconnues en modalité audiovisuelle qu'en modalité audio ou vidéo dans l'ensemble des trois groupes. Les deux groupes d'apprenants, dont les performances ne différaient pas de façon notable, ont eu

plus de difficultés que les natifs à identifier certaines expressions, notamment la politesse en condition audio seule majoritairement confondue avec la déclaration, et la question en condition visuelle seule confondue avec la déclaration. Les différences entre conditions dans chaque groupe de sujets, ainsi que les différences limitées en condition audio-visuelle entre les sujets natifs et les apprenants francophones, ont suggéré que dans un contexte de communication face-à-face, les apprenants pourraient compenser leurs difficultés à identifier les corrélats acoustiques des affects sociaux en s'appuyant plus largement sur des indices visuels, plus largement partagés entre langues et culture.

3.5 Expression d'attitudes agressives et hostiles

En collaboration avec Charlotte Kouklia, en majeure partie dans le cadre de sa thèse (Kouklia, 2019) que j'ai codirigée avec Jacqueline Vaissière, nous avons mené une série d'études dans lesquelles nous avons cherché à caractériser l'expression en français d'attitudes agressives ou hostiles, tout d'abord à partir de données produites en laboratoire dans un contexte contrôlé, puis sur des extraits de parole politique produits dans un contexte écologique.

Si les caractéristiques phonétiques les plus saillantes de la colère chaude, généralement désignée implicitement lorsque le terme de colère est employé sans autres spécifications, sont assez largement décrites dans la littérature (voir par exemple Juslin & Laukka (2003) pour une revue) avec notamment une augmentation de l'intensité acoustique et du registre de fréquence fondamentale au-delà d'une simple conséquence de l'augmentation d'intensité, des descriptions divergentes peuvent être trouvées pour les autres affects liés à l'agressivité tels que la colère froide, associée à une réaction émotionnelle moins activée et donc exprimée dans la parole de façon plus variable (Banse & Scherer, 1996; Scherer, 2003). En français, les formes dites réprimées de colère ont été décrites en phonostylistique par une augmentation du débit de parole, des perturbations temporelles telles que des schémas rythmiques instables, un renforcement des accents secondaires et une modification du rapport entre la durée des voyelles et celle des consonnes (Léon, 1999; Fónagy, 1983). Ces caractéristiques tendent à dévier de la structure rythmique canonique du français, caractérisée par son isochronie et l'allongement des syllabes et des voyelles en fin de mot et en fin de groupes prosodiques.

3.5.1 Du corpus de laboratoire…

Publication associée :

[ACTI40] Kouklia, C., & **Audibert, N.** (2013). Perceptual, acoustic and electroglottographic correlates of 3 aggressive attitudes in French: a pilot study. *Proceedings of Interspeech 2013*, Lyon, France. pp. 1389-1393.

La première étude à laquelle j'ai contribué en m'impliquant dans l'analyse des données ainsi que dans l'interprétation des résultats s'est appuyée sur les données recueillies dans le cadre du mémoire de master de Charlotte Kouklia, antérieur à mon recrutement à la Sorbonne Nouvelle. Ces données consistent en la production par deux acteurs francophones semi-professionnels d'une attitude agressive d'ironie sarcastique et de deux formes d'expression de la colère incluant une forme contrôlée de colère froide, qui de par son caractère contrôlé peut être assimilée à une attitude au sens de Léon (1999), complétées par une expression neutre à

titre de comparaison. Ces expressions ont été recueillies dans une démarche classique de phonétique expérimentale, avec un enregistrement en chambre sourde, une variation systématique du contenu segmental autour d'un énoncé sémantiquement neutre et l'utilisation de mesures articulatoires, tout en s'appuyant sur l'utilisation d'un scénario afin de contextualiser les productions via la co-construction par l'acteur et l'expérimentatrice de dialogues dans une approche similaire à celle d'Enos & Hirschberg (2006). Les énoncés ont consisté en la phrase « CaCaCa, tu n'as pas dit CaCa, mais CaCaCa », avec C = {p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, ʁ, l}, le choix de la voyelle /a/ étant guidé par les résultats de la littérature qui indiquent que les voyelles ouvertes sont plus influencées par l'expression des émotions et attitudes (voir par exemple Yildrim et al. (2004)), et en un ensemble de huit phrases françaises sémantiquement neutres inspirées d'études précédentes et notamment de l'étude de Fónagy & Bérard (1972). Parmi ces huit phrases, seul l'énoncé « Il vient demain ? » a été inclus dans nos travaux communs.

L'analyse perceptive par 28 auditeurs francophones natifs a consisté en des jugements sur des échelles de Likert à sept points du degré d'agressivité, utilisé en remplacement de l'échelle de valence en raison de l'inclusion uniquement d'expressions négatives, ainsi que du degré de contrôle et de la dominance, considérées comme potentiellement indépendants l'un de l'autre dans ce contexte. Les résultats ont montré notamment que les expressions agressives étaient toutes perçues comme moins contrôlées que les expressions neutres, et qu'elles étaient par ailleurs perçues comme d'autant moins contrôlées qu'elles étaient perçues comme agressives et dominantes. La comparaison entre les jugements attribués à la colère chaude et ceux attribués à la colère froide et à l'ironie sarcastique, perçues comme moins agressives et dominantes, sont consistants avec la définition de la colère chaude comme une émotion très activée et incontrôlée (Scherer et al., 1991).

L'analyse acoustique a consisté en l'extraction des durées segmentales après une étape d'alignement forcé suivi d'une correction manuelle de la segmentation afin d'en dériver des mesures de débit de parole, des fréquences des trois premiers formants, et de la fréquence fondamentale et du quotient ouvert à partir du signal électroglottographique (EGG) recueilli de façon synchrone aux enregistrements acoustiques. Dans les cas minoritaires (8%) dans lesquels le signal EGG était trop faible pour permettre d'en extraire la fréquence fondamentale, les valeurs détectées par l'algorithme de Boersma (1993) à partir du signal acoustique ont été utilisées en remplacement, après inspection des valeurs extraites pour s'assurer de l'absence de sauts d'octave et de la consistance avec les valeurs mesurées à partir de l'EGG. En revanche ces portions ont été exclues de l'analyse des valeurs de quotient ouvert en raison de l'impossibilité d'identifier les pics positifs et négatifs de la dérivée du signal EGG.

La Figure 14 illustre les contours de fréquence fondamentale relevés sur les productions du logatome central /mama/ précédant une frontière de continuation majeure, produit par l'un des deux acteurs avec les quatre cibles expressives. On peut noter que les contours descendants abrupts observés dans ce cas sur les expressions de colère chaude et d'ironie sarcastique rejoignent les prédictions de Fónagy (1983) concernant les expressions agressives, ainsi que certains des postulats d'Ohala (1983a) liés au code fréquentiel. Des stratégies opposées d'expression de l'ironie sarcastique ont toutefois été observées entre le locuteur représenté sur cette figure qui l'exprime avec une fréquence fondamentale descendante en fin de groupe prosodique qui correspondrait plutôt à une expression de moquerie dans la taxonomie établie par Anolli et al. (2002), tandis que l'autre locuteur opte pour une forte

montée de la fréquence fondamentale qui correspondrait à une expression de « moquerie vive ».

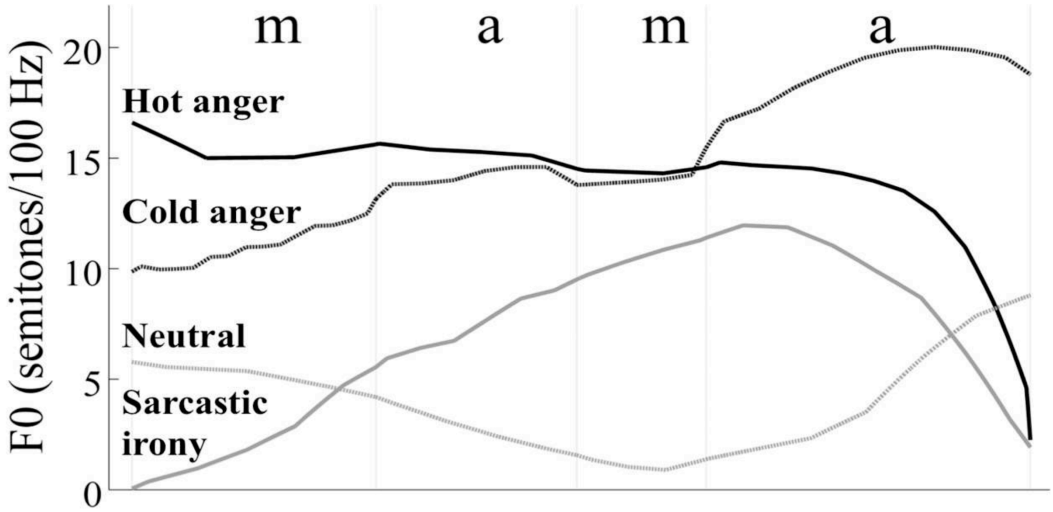


Figure 14 : Contours de fréquence fondamentale en demi-tons (le niveau 0 correspondant à la fréquence 100 Hz) produits par l'un des deux acteurs sur le logatome /mama/ au milieu de l'énoncé « mamama, tu n'as pas dit mama mais mamama », comparé entre expression de la colère chaude, de la colère froide, de l'ironie sarcastique et l'expression neutre. D'après Kouklia & Audibert (2013, [ACTI40]).

Pour les deux locuteurs les résultats ont montré un fort lien entre degré d'agressivité perçue et étendue de la fréquence fondamentale, tandis que les jugements de dominance étaient principalement liés aux mesures de F1 sur la voyelle /a/ dans les logatomes, la phrase n'ayant pas fait l'objet d'une analyse formantique en raison de la difficulté de mesure des formants sur les voyelles nasales.

A partir des productions des quatre expressions sur les logatomes, les mesures acoustiques relevées sur les différentes occurrences de la voyelle /a/ ont été comparées entre catégories d'affects et contexte consonantique, afin d'évaluer via la comparaison des tailles d'effet si les réalisations de cette voyelle sont plus fortement impactées par le type d'attitude produit ou par le contexte consonantique, les résultats de la littérature ne prenant généralement pas en compte la variation consonantique, notamment pour évaluer les liens entre affects et formants. Cette analyse a montré que la fréquence fondamentale et la position du premier formant dépendent plus de l'expression que du contexte consonantique, conformément aux résultats de la littérature qui mettent souvent en avant ces mesures comme corrélats acoustiques des émotions (voir par exemple Juslin & Laukka (2003) pour une large revue des liens entre émotions et mesures acoustiques). C'est également le cas du quotient ouvert qui reflète le degré de tension laryngée avec des différences faibles mais néanmoins supérieures au seuil de perception (Henrich et al., 2003), mais aussi des stratégies divergentes d'ajustement de la tension vocale entre les deux locuteurs. De façon plus surprenante, le deuxième formant, connu pour être fortement influencé par la coarticulation, ainsi que le troisième formant, se sont avérés influencés tout autant par l'attitude exprimée que par le contexte consonantique.

3.5.2 … à l'étude contrôlée de productions écologiques

Publication associée :

[ACTI31] Kouklia, C., & **Audibert, N.** (2017). Relationships Between Speech Timing and Perceived Hostility in a French Corpus of Political Debates. *Proceedings of Interspeech 2017*, Stockholm, Suède, pp. 899-903

Les travaux suivants que j'ai menés en collaboration avec Charlotte Kouklia sur l'expression d'attitudes hostiles se sont inscrits plus directement dans le cadre de sa thèse, et se sont appuyés sur les données de parole politique issues du corpus composé des sessions du conseil municipal de Montreuil, présenté brièvement en section 7.3.2, dans lequel les expressions plus ou moins fortement hostiles des locuteurs ont été complétées par une condition de contrôle de relecture. Ces productions ont été évaluées perceptivement dans plusieurs conditions afin d'évaluer le degré d'hostilité véhiculé par les modulations prosodiques et de qualité de voix indépendamment du contenu sémantique, et le mettre en rapport avec un ensemble de mesures extraites du signal acoustique.

En complément des principes méthodologiques adoptés dans ces travaux et évoqués en section 7.3.2, je reviens ici plus spécifiquement sur un autre aspect du travail réalisé dans cette thèse sur lequel j'ai collaboré plus étroitement avec la doctorante et qui a fait l'objet d'une publication commune, la méthodologie d'analyse ainsi définie ayant ensuite été étendue dans la thèse (Kouklia, 2019) à des mesures spectrales et dérivées de la fréquence fondamentale. Dans ce travail, nous avons cherché à caractériser les déviations temporelles qui caractérisent les expressions d'attitudes hostiles par rapport au patron temporel canonique en français standard, caractérisé par une structure rythmique isochronique et un allongement syllabique en fin de groupe ou de mot dépendant du niveau prosodique de la frontière droite.

Pour cela, nous nous sommes appuyés sur les évaluations perceptives du degré d'hostilité perçu réalisées auparavant en condition audio, c'est-à-dire en présentant aux auditeurs les 125 énoncés originaux (25 pour chacun des cinq locuteurs sélectionnés) extrait des sessions du conseil municipal de Montreuil, ainsi que sur l'évaluation par autre ensemble de sujets selon le même protocole des transcriptions orthographiques de ces énoncés. Cela nous a permis d'en dériver une mesure notée dHost, calculée comme la différence entre l'évaluation de l'hostilité perçue dans l'énoncé originale et celle attribuée à sa transcription. Cette mesure est supposée refléter le degré d'hostilité spécifiquement véhiculé par les variations prosodiques et de qualité de voix, indépendamment du poids du contenu sémantique dans l'attribution du degré d'hostilité.

Nous avons également utilisé comme condition de référence pour l'analyse des réalisations acoustiques de ces énoncés les relectures par chacun des cinq locuteurs de leurs propres énoncés transcrits au préalable, avec la consigne d'effectuer cette relecture à la manière d'une dictée afin de limiter les facteurs de variation incontrôlés.

L'ensemble de ces énoncés originaux ou relus, segmentés en phones au moyen d'un alignement forcé suivi d'une étape de correction manuelle, ont été segmentés en unités inter-pausales, en considérant le seuil de 140 ms. La plupart des études qui considèrent les pauses silencieuses dans la parole spontanée (Ferré, 2005) ou dans la parole politique (Béchet et al., 2013) ont retenu un seuil de 200 ms à partir duquel un silence est considéré comme une pause silencieuse, tandis que Duez (1991) a proposé l'utilisation d'un seuil variable défini en fonction

des durées segmentales pour tenir compte des différences de débit de parole. Toutefois, étant donnée la grande variabilité observée dans les données du conseil municipal de Montreuil en termes de variation locale du débit de parole, cette approche a été jugée difficilement applicable et le seuil de 140 ms, soit le seuil le plus court utilisé par Duez (1991), a été retenu.

Les métriques extraites sur chaque unité inter-pausale ainsi définie ont consisté d'une part en un comptage des voyelles, consonnes, syllabes, pauses silencieuses et pauses pleines, et en des mesures de moyenne et d'écart-type des durées des différentes catégories d'unités considérées ainsi que de la pause précédant et suivant l'unité inter-pausale, et du rapport V/C entre durée de la voyelle et durée de la rime au sein de la syllabe. Ces mesures ont été complétées par des mesures de débit de parole et de débit articulatoire ainsi que par les mesures rythmiques de variation de durée en unités consécutives PVI et nPVI (Grabe & Low, 2002). Enfin des mesures ont été spécifiquement développées pour capturer les déviations par rapport au patron d'accentuation canonique en français, dans lequel d'après Léon (1999), au-delà de l'allongement final lié à la frontière de groupe intonatif, l'allongement des syllabes finales de mots est supposé augmenter graduellement entre le début et la fin d'une unité inter-pausale. Ces mesures, calculées à la fois à partir des durées des voyelles et de celles des rimes, considèrent la première et/ou la dernière syllabe de chaque mot plurisyllabique. La mesure notée « rapport durée mot fin/début » capture l'allongement relatif de la dernière syllabe du mot comparativement à la première, et est donc supposé être supérieur à 1 dans la grande majorité des cas, mais potentiellement raccourci dans le cas de la présence d'un accent didactique augmentant la durée de la syllabe initiale. Une mesure de pente de durée obtenu par régression entre le début et la fin de de l'unité inter-pausale a également été considérée, mais n'a pas été retenue parmi les mesures sélectionnées en raison de son très faible pouvoir informatif.

Chacune des métriques ainsi extraites a été d'une part moyennée sur l'ensemble des unités inter-pausales de l'énoncé, et a d'autre part fait l'objet de mesures de différences entre unités inter-pausales consécutives afin d'évaluer la fluctuation entre unités avec un empan temporel intermédiaire entre la syllabe ou le mot et l'énoncé complet. Enfin, après appariement entre les énoncés originaux et les énoncés relus à la façon d'une dictée (au niveau de l'énoncé, les différences de structure prosodiques entre versions ne permettant pas toujours un appariement direct entre unités inter-pausales), les différences entre ces deux conditions ont été calculées. Cela nous a permis d'obtenir une estimation de la divergence de réalisation acoustique entre les énoncés originaux extraits des sessions du conseil municipal et la version relue considérée comme référence en termes de réalisation des patrons rythmiques canoniques du français standard. Après une étape de comparaison entre les deux versions, ce sont ces différences entre relecture et énoncé d'origine qui ont été confrontées aux jugements perceptifs.

Les mesures liées à un même phénomène temporel étant fortement intercorrélées (par exemple la durée des intervalles est fortement liée au nombre de segments ou syllabes qui les composent), nous avons choisi parmi des ensembles de variables similaires celle pour laquelle l'effet de la comparaison entre la condition originale et la condition de relecture était le plus fort au sens du d de Cohen utilisé comme mesure de taille d'effet. Ainsi nous avons obtenu un ensemble de huit variables sélectionnées, pour lesquelles les mesures de différence entre UIP consécutives ont également été retenues. Outre la durée des pauses, les principales différences entre les productions originales et relues concernaient les aspects rythmiques, avec notamment la variabilité du rapport V/C due aux différences entre les moyennes des

durées relatives de l'attaque et de la rime, la valeur moyenne de nPVI avec une variabilité accrue dans les énoncés originaux, la durée des unités inter-pausales, et la durée relevée en position finale d'IUP avec un allongement final réduit dans les énoncés originaux.

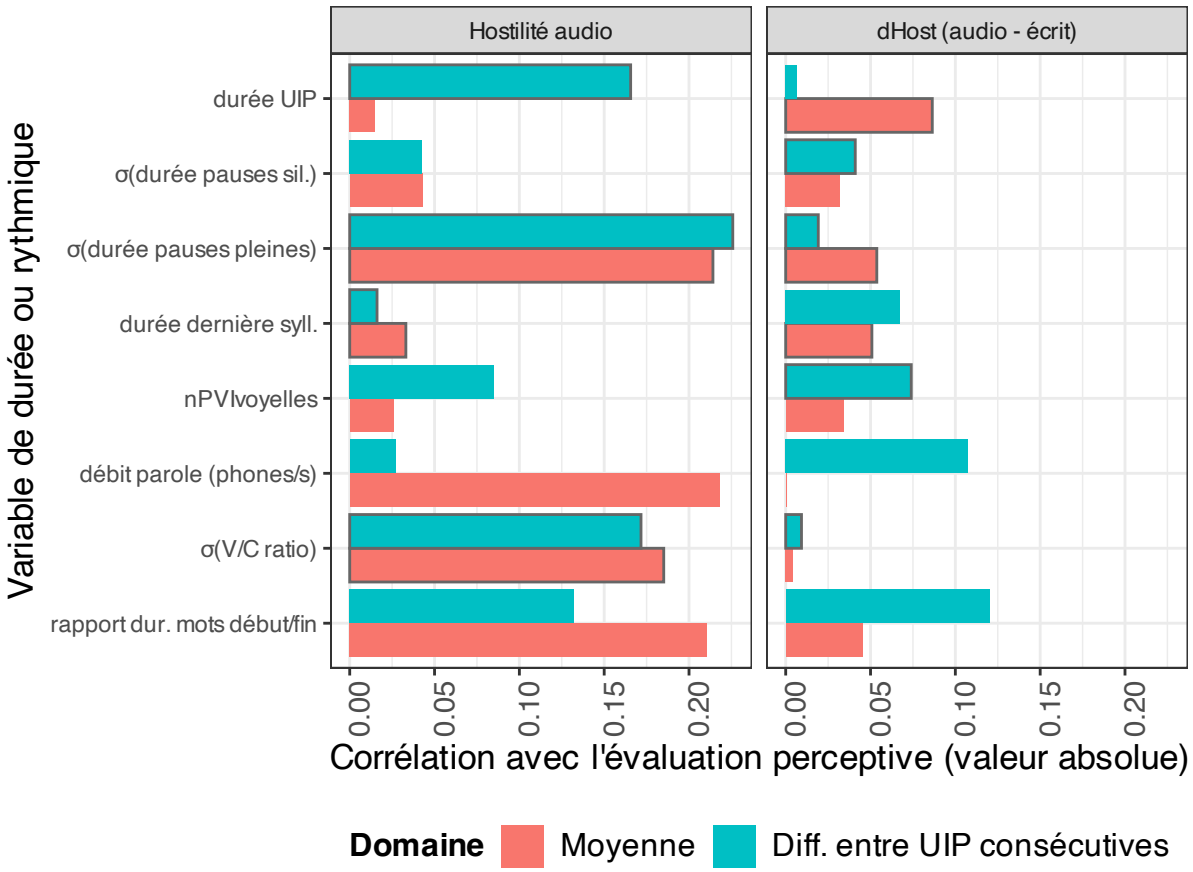


Figure 15 : Corrélations entre (1) mesures moyennes sur l'ensemble des unités inter-pausales (UIP) ou moyenne de la différence entre UIP consécutives extraites des énoncés originaux et soustraites des valeurs mesurées sur les énoncés relus pour chacune des huit variables temporelles sélectionnées et (2a, gauche) jugements d'hostilité sur ces énoncés originaux, ou (2b, droite) différence dHost de jugement d'hostilité entre la version originale de l'énoncé et sa transcription orthographique. Les corrélations négatives qui correspondent aux cas dans lesquels l'hostilité perçue diminue quand la valeur de la variable augmente sont signalées par un encadré gris. Adapté des données présentées dans Kouklia & Audibert (2017, [ACTI31]).

Les corrélations avec les évaluations du degré d'hostilité, présentées dans la Figure 15, indiquent qu'en considérant les mesures effectuées sur chaque unité inter-pausale (UIP), étiquetées « Moyenne » dans la figure, les énoncés évalués comme plus hostiles sont réalisés avec un rapport V/C plus stable, un débit de parole plus instable à l'intérieur de l'unité inter-pausale (UIP), une accentuation relative plus faible des syllabes finales au niveau du mot, des UIP plus courtes (en relation avec le nombre de pauses silencieuses insérées), et en ne considérant que les jugements sur les énoncés originaux et non la différence dHost, une accentuation plus courte à la fin des UIP. Bien que les corrélations avec la fluctuation entre UIP consécutives aient été globalement plus faibles, on peut noter en particulier que la perception différentielle de l'hostilité dHost augmente lorsque le rapport de durée entre début et fin de mots et le débit de parole augmentent. Sur ce dernier point, les différences de débit de parole

entre passages adjacents pourraient être liées à l'utilisation de pauses de focalisation, fréquentes dans le discours politique en français (Duez, 1991; Béchet et al., 2013).

Du point de vue des différences entre locuteurs, le résultat le plus notable est observé pour la variabilité du rapport V/C, avec la maire qui contrairement à ses quatre opposants politiques est perçue comme plus hostile lorsque ce rapport est plus instable d'une syllabe à l'autre.

Si les inversions de signe observées entre les corrélations d'une même mesure avec le degré d'hostilité obtenu en condition audio et le degré d'hostilité différentiel dHost peuvent suggérer dans certains cas une atténuation prosodique du niveau d'hostilité attribué à partir du contenu sémantique seul, les différences importantes entre les deux versions de l'évaluation du degré d'hostilité observées pour certaines mesures suggèrent aussi un masquage possible des expressions prosodiques de l'hostilité par le contenu sémantique au-delà de la redondance probable de l'information affective véhiculée par ces deux canaux (avec dans ces données, une corrélation de r = .76 entre l'hostilité estimée à partir de l'audio et celle estimée à partir des transcriptions orthographiques). Ainsi et bien que la méthode proposée pour tenter de caractériser les informations affectives véhiculées par la prosodie et la qualité de voix indépendamment du contenu sémantique ait ouvert des pistes prometteuses pour l'étude phonétique fine d'expressions d'affects produits dans des conditions écologiques, en l'absence d'une évaluation d'une version délexicalisée de ces énoncés on ne peut exclure que certaines des corrélations observées puisse ne refléter que des phénomènes non pertinents perceptivement.

Au-delà de ce constat qui limite l'interprétation possible des résultats en termes de contribution à l'expressivité perçue, le protocole proposé avec la condition de relecture dépendante du locuteur et des mesures exprimées relativement à cette condition de référence pourrait s'avérer particulièrement utile pour l'étude fine des styles de parole produits dans un contexte écologique, notamment pour permettre la quantification des variations prosodiques et de qualité de voix.

4 Variation vocalique

Résumé du chapitre 4

Dans ce chapitre je reviens sur mes travaux autour de la variation de la réalisation des voyelles par des locuteurs sains, en revenant tout d'abord sur les différents facteurs de variation documentés dans la littérature que j'ai pu prendre en compte, et j'étends une partie de ces travaux à la question de la variation inter-individuelle.

Une première étude sur grands corpus de l'effet du style de parole sur l'espace vocalique a conclu à une centralisation et une variation intra-catégorie plus importantes en parole conversationnelle qu'en parole lue ou journalistique à durée égale. Une nouvelle analyse de ces données indique que contrairement à la centralisation qui dépend surtout de la durée, la variation intra-catégorie et le recouvrement acoustique entre catégories dépendent autant voire plus du style de parole que de la durée, avec un lien entre durée et organisation de l'espace vocalique très variable entre locuteurs. Une étude contrôlée confirme un effet plus important du style sur la variabilité intra-catégorie que sur la dispersion et suggère un effet principalement lié à l'interactivité de la situation de production de parole. Les résultats d'une étude sur corpus de la distinction entre /e/ et /ɛ/ en fin de mot en français standard suggèrent que le processus de fusion entre ces voyelles serait plus avancé en parole conversationnelle que journalistique. Une nouvelle analyse par locuteur révèle qu'en parole journalistique également une partie des locuteurs tendent à fusionner ces deux catégories.

Une série de travaux sur la réalisation acoustique des voyelles dans les productions codifiées parlées et chantées adressées au nourrisson en allemand a confirmé les principaux résultats de la littérature sur la parole dirigée vers l'enfant et la distinction entre parole et chant, et a montré un effet de la distinction entre productions destinées à apaiser ou stimuler l'enfant sur la fréquence fondamentale et sa variabilité et dans une moindre mesure sur la durée, mais pas sur l'organisation du système vocalique. Une nouvelle analyse par locutrice révèle des stratégies individuelles divergentes d'adaptation entre productions apaisantes ou stimulants, et dans une moindre mesure entre chant et parole et entre productions adressées ou non à l'enfant.

Une étude acoustique et une étude articulatoire de l'effet de la position prosodique sur le contraste d'arrondissement en position initiale en français ont indiqué un renforcement du contraste en position prosodique forte, lié aux variations de l'aire aux lèvres.

Enfin une série d'études à partir de grands corpus sur la coarticulation voyelle-à-voyelle en français a montré notamment que les voyelles moyennes postérieures sont plus sensibles à l'harmonie vocalique et que les voyelles en position initiale absolue y résistent plus, avec un faible effet du style de parole et de la durée. Une nouvelle analyse individuelle sur les locuteurs les plus représentés confirme un effet plus important sur les voyelles postérieures pour la majorité des locuteurs, mais révèle également une grande variabilité inter-individuelle qui ne dépend ni du style de parole ni de différences entre hommes et femmes.

4.1 Des facteurs de variation multiples

De nombreux facteurs de variation sont susceptibles d'influencer la réalisation des sons de la parole et tout particulièrement des voyelles. Parmi les facteurs de variation intra-locuteur les plus largement documentés, on peut bien entendu mentionner l'effet du contexte segmental à travers la coarticulation, notamment l'effet connu de longue date du contexte consonantique sur la fréquence du second formant (voir par exemple Liberman et al. (1967))

qui a par la suite été formalisé à travers les équations du locus (Sussman et al., 1991). L'influence de la coarticulation voyelle-à-voyelle est également un phénomène documenté par de multiples études menées sur différentes langues, par exemple Recasens (1984) sur le catalan, Magen (1998) sur l'anglais ou Nguyen & Fagyal (2008) sur le français.

L'influence sur la réalisation des segments vocaliques du débit de parole et de la durée de ces segments a également été largement étudiée, les segments de durée réduite produits avec un débit élevé étant généralement associés à une hypo-articulation (voir par exemple Gay (1978) sur l'anglais américain, ou Nadeu (2014) sur l'espagnol et le catalan). La variation en fonction du style de parole, qui peut être liée à la fois à l'appartenance à un groupe idiolectal, à un rôle sociétal ou encore à l'adaptation à l'interlocuteur ou plus largement à une situation de communication particulière, a également fait l'objet d'un nombre important de travaux (voir par exemple Harmegnies & Poch-Olivé (1992) sur l'espagnol, ou encore Lancien & Côté (2018) sur le français laurentien).

Sans que cette liste soit exhaustive, on peut également mentionner d'autres facteurs de variation intra-locuteur identifiés dans la littérature, comme la densité du voisinage phonologique et la fréquence lexicale, dont le rôle dans la réalisation des voyelles a été étudié par exemple par Wright et al. (2004), ou encore l'accentuation (Fourakis, 1991; Nadeu, 2014; van Bergem, 1993) et plus généralement le renforcement prosodique de l'articulation au niveau segmental (Fougeron & Keating, 1997; Cho, 2005). En outre, il existe une abondante littérature scientifique traitant des interactions entre certains de ces facteurs de variations intra-locuteur.

La réalisation des voyelles est également variable entre locuteurs d'une même communauté linguistique, en raison notamment des caractéristiques anatomiques du locuteur dont dépendent les fréquences de résonance et qui expliquent en partie les différences de fréquences formantiques observées entre hommes et femmes (Fant, 1975; Diehl et al., 1996) et au-delà certains aspects de la variation inter-individuelle, mais aussi de différences de stratégies articulatoires.

Dans cette partie, je reviens sur certains de mes travaux dans lesquels je me suis penché sur le rôle de certains de ces facteurs de variation, considérés seuls ou en interaction. Dans la plupart de ces travaux, je me suis concentré sur l'organisation du système vocalique des locuteurs considéré de façon globale, plutôt que sur la réalisation d'exemplaires de voyelles considérés séparément. Par ailleurs, la majorité des travaux présentés dans cette section reposent sur des analyses à grande échelle de voyelles extraites de corpus de parole continue, dans une approche de phonétique de corpus.

4.2 Variation de l'espace vocalique en fonction du style de parole

Le terme de style de parole, largement utilisé dans la littérature en phonétique, est très vaste et recouvre des aspects variés, avec des définitions qui divergent entre auteurs. Si dans le cadre des études en sociophonétique ce terme est plus couramment utilisé pour désigner les variations idiolectales associées à un groupe socio-culturel, l'acception la plus courante en sciences de la parole reste de considérer le style de parole comme la manifestation de l'adaptation à une situation de communication particulière (voir par exemple Boula de Mareüil (2014) pour une réflexion sur ce qu'est un style de parole). Une telle définition laisse toutefois une marge d'interprétation importante. Cette notion d'adaptation à une situation de communication peut en effet correspondre à des phonogenres variés, pour reprendre la

définition de Simon et al. (2013) qui les distinguent des phonostyles plus spécifiques et pour certains associés à un rôle professionnel tels que ceux documentés dans le corpus C-Phonogenre (Goldman et al., 2014), qui incluent par exemple des questions aux gouvernement en session parlementaire, des productions liturgiques, le commentaire sportif ou encore le phonogenre caractéristique de la parole produite par les enseignants.

Bien que cette dimension ne couvre qu'une partie des variations pouvant être considérées comme relevant du style de parole, une approche plus répandue considère les styles de parole comme répartis sur un continuum allant de la parole dite claire (désignée comme « clear speech » dans la littérature anglophone), dans laquelle le locuteur fait en sorte que les indices acoustiques soient les plus informatifs possible afin de faciliter l'intelligibilité et la distinction entre segments, à la parole interactionnelle relâchée qui laisse plus de place à d'autres facteurs de variation qui ne sont pas directement liés à la communication du message linguistique, au détriment de la précision segmentale. Un tel continuum peut également être considéré comme une modulation entre productions segmentales hyper-articulées dans le cas de la parole claire, ou au contraire hypo-articulées dans le cas de la parole conversationnelle.

Si certains travaux ciblent directement des styles de parole supposés situés à l'extrémité la plus hyper-articulée de ce continuum en raison du but communicatif qui impose cette hyper-articulation, comme par exemple la parole adressée aux personnes malentendantes (Picheny et al., 1986), une approche plus courante de l'étude de la variation entre styles de parole repose sur la comparaison entre la parole conversationnelle ou une autre forme de parole considérée comme spontanée d'une part, et la parole lue d'autre part (voir par exemple Harmegnies & Poch-Olivé (1992) sur l'espagnol, ou Nakamura et al. (2008) sur le japonais). De plus on peut noter qu'à l'autre extrémité du continuum, la parole non-lue peut se subdiviser en de nombreuses sous-catégories en fonction du degré « d'attention » porté à la parole produite, de la tâche et de l'interlocuteur (Warner, 2012). Cette proposition rejoint également les conclusions de Scarborough & Zellou (2013), qui considèrent que la parole claire peut elle-même se subdiviser en catégories situées à différents niveaux du continuum en fonction de l'effort de clarté associé à la production, dont découlerait la réalisation acoustique. Les études ayant comparé le système vocalique entre styles situés à des niveaux intermédiaires de ce continuum entre parole claire et parole lue sont peu nombreuses. On peut toutefois mentionner l'étude d'Harmegnies & Poch-Olivé (1994) qui ont comparé les voyelles produites par un locuteur dans une tâche de lecture, des tâches de description d'image interactives ou non, un monologue non-scripté et une situation de dialogue spontané et ont conclu à une centralisation plus importante en parole spontanée tandis que les tâches de description d'image étaient caractérisées par un abaissement du premier formant.

Les caractéristiques acoustiques des voyelles associées dans la littérature à ce continuum entre parole claire et parole conversationnelle sont consistantes pour les différentes langues dans lesquelles elles ont été étudiées, avec une description de la parole conversationnelle (voir par exemple Meunier & Espesser (2011) pour le français, Bona (2014) pour le hongrois, ou Johnson (2004) pour l'anglais américain) caractérisée par une réduction à la fois de la durée des voyelles et de leur propriétés spectrales, les voyelles produites en parole conversationnelle étant décrites comme moins distinctes entre elles ou plus centralisées. A l'inverse la parole claire a été décrite comme caractérisée par une augmentation à la fois de la durée des voyelles (voir par exemple Ferguson & Kewley-Port (2007) pour l'anglais américain) et de la taille de l'espace vocalique (Picheny et al., 1986).

4.2.1 Interaction entre style de parole et durée segmentale

Publication associée :

[ACTI34] **Audibert, N.**, Fougeron, C., Gendrot, C., & Adda-Decker, M. (2015). Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS'15)*, Glasgow, Royaume-Uni, paper 0753 (actes en ligne).

4.2.1.1 Une variation vocalique liée au style de parole ou un simple effet de la durée ?

L'interprétation la plus courante du lien entre les changements temporels et spectraux en fonction du style de parole repose sur la notion de compromis biomécanique entre le temps d'articulation et la précision articulatoire. L'articulation incomplète des voyelles observée dans les segments brefs caractéristiques de la parole conversationnelle s'expliquerait ainsi par un manque de temps pour réaliser les ajustements dynamiques nécessaires à l'atteinte des cibles segmentales articulatoires (Lindblom, 1990; Moon & Lindblom, 1994), d'où la centralisation observée sur le plan acoustique. Plusieurs études menées à partir de l'analyse acoustique de grands corpus en français (Gendrot & Adda-Decker, 2005; Meunier & Espesser, 2011) ont conclu à une réduction progressive de l'espace vocalique avec la réduction de la durée des voyelles pour un même style de parole.

Cependant, on peut se demander si ce compromis spatiotemporel est le seul responsable de la réduction des voyelles observée dans la parole conversationnelle. En effet, plusieurs études menées sur le français ont mis en évidence des interactions entre le style de parole et la durée segmentale sur la réduction spectrale des voyelles. Ainsi, Rouas et al. (2010) ont montré à partir de mesures obtenues via une paramétrisation du signal acoustique par un ensemble de coefficients cepstraux que sur des segments de durée réduite, des différences de réduction acoustique pouvaient être observées entre parole lue et parole conversationnelle. De plus, la comparaison entre les résultats de l'étude de Gendrot & Adda-Decker (2005) sur la parole journalistique et celle de Meunier & Espesser (2011) pour des classes de durée comparables suggère également une variation dépendante du style à durée égale. En effet, si les deux études concluent à une réduction globale de l'espace vocalique entre les voyelles longues et les voyelles de durée plus réduite, le patron observé en parole conversationnelle (Meunier & Espesser, 2011) est celui d'une centralisation à la fois sur F1 et sur F2 qui concerne l'ensemble des catégories vocaliques, tandis qu'en parole journalistique cette réduction touche plus largement F1 en raison d'une réduction massive du premier formant sur les occurrences de la voyelle /a/.

Nous avons donc cherché à explorer la relation entre la réduction des voyelles en fonction de la durée et du style dans trois corpus de parole de grande taille en français standard, représentant chacun un style de parole différent : le corpus BREF de parole lue par des locuteurs non-professionnels (Lamel et al., 1991), le corpus ESTER (Galliano et al., 2006, 2009) de parole médiatisée extraite de journaux et émissions de débats diffusées à la radio et à la télévision, composé majoritairement de parole préparée produite par des journalistes professionnels, et le corpus NCCFr (Torreira et al., 2010) de parole conversationnelle produite lors d'échanges informels face-à-face entre amis. L'analyse de ces trois corpus s'est appuyée sur une segmentation obtenue à l'aide du système d'alignement forcé développé par le laboratoire LIMSI. Afin de rendre compte des différentes dimensions possibles de variation

d'un système vocalique, nous nous sommes appuyés sur un ensemble de métriques destinées à rendre compte non seulement du degré de centralisation du système vocalique (globalement et plus spécifiquement sur F1 et F2) mais aussi de la variabilité au sein de chaque catégorie de voyelle et de la neutralisation des contrastes entre catégories.

Je résume ci-dessous les choix méthodologiques effectués ainsi que les principaux résultats obtenus dans l'étude présentée en 2015, avant de proposer une réanalyse d'une partie de ces résultats à partir de mesures prises à l'échelle de l'exemplaire vocalique et de nouveaux développements axés sur la variation entre locuteurs.

4.2.1.2 Choix méthodologiques et principaux résultats de l'étude de 2015

Afin de quantifier les variations de l'espace vocalique à partir de valeurs de F1 et de F2 extraites automatiquement, nous nous sommes concentrés sur un sous-ensemble du système vocalique du français, en excluant les voyelles nasales sur lesquelles l'analyse formantique est particulièrement complexe, ainsi que les voyelles antérieures arrondies dont la caractérisation acoustique nécessite la prise en compte des valeurs du troisième formant. Le sous-ensemble considéré était donc composé des voyelles orales /i, e, ɛ, a, o, ɔ, u/. Les voyelles /e, ɛ/ et /o, ɔ/ ont été fusionnées en une seule classe, ci-après étiquetée /e/ et /o/, pour rendre compte de leur fréquente neutralisation (voir par exemple Fagyal et al. (2002)) et des variations régionales en français. En effet, bien que les locuteurs inclus dans l'étude soient tous supposés être locuteurs du français standard, les données disponibles pour les différents corpus ne permettent pas d'exclure l'influence d'autres variantes régionales sur les réalisations de certains de ces locuteurs. Les analyses reposent donc sur cinq catégories de voyelles orales : /i, e, a, o, u/. Les fréquences des deux premiers formants, extraites à l'aide de l'algorithme de Burg implémenté dans Praat (Boersma, 2001) et des valeurs par défaut suggérées respectivement pour les productions d'hommes et de femmes, ont été filtrées via l'application d'un crible dépendant de la catégorie vocalique afin d'éliminer les valeurs aberrantes selon la méthode de Gendrot & Adda-Decker (2005).

Seuls les locuteurs ayant produit un minimum de 15 exemplaires vocaliques dans chacune des cinq catégories (ce qui revient en pratique à filtrer en fonction du nombre d'occurrences de la catégorie /u/ la moins représentée) ont été inclus. Un total de plus d'un million de voyelles (1 143 941) produites par 74 locuteurs du corpus BREF, 61 locuteurs du corpus ESTER et 45 locuteurs du corpus NCCFr ont été incluses. Ces voyelles ont été subdivisées en trois classes de durée via l'application des mêmes seuils que dans l'étude de Gendrot & Adda-Decker (2005) pour les voyelles courtes et de durée moyenne, avec une durée maximale fixée respectivement à 50ms et 80ms. Contrairement à l'étude de Gendrot & Adda-Decker (2005) sur la parole journalistique dans laquelle la durée maximale des voyelles considérées comme longues avait été fixée à 110ms, les voyelles de durée allant jusqu'à 300ms ont été incluses dans cette catégorie pour tenir compte des allongements plus conséquents observés en parole lue et conversationnelle. On peut noter que si la distribution des voyelles entre classes de durée était très similaire entre les corpus ESTER et NCCFr avec environ 50% de voyelles courtes et 30% de voyelles moyennes, les voyelles courtes ne représentent que 22% des voyelles du corpus BREF contre 44% de moyennes et 34% de longues.

Six métriques parmi celles présentées en section 7.2.3 ont été calculées pour chaque locuteur et chaque classe de durée, dont trois relatives à la compression de l'espace vocalique et consistant en une projection sur une valeur unique pour l'ensemble des productions d'un locuteur dans une classe de durée donnée : l'aire du pentagone /a, e, i, u, o/ pVSA, le degré de

compression des valeurs du premier formant F1RR, et le degré de compression des valeurs du second formant F2RR. Afin de permettre une comparaison directe avec ces métriques pour lesquelles une valeur unique par locuteur et par condition est calculée, les valeurs des métriques DistCentroid (centralisation), VDispersion (variation intra-catégorie) et ContrastLoss (distinctivité entre catégories) également retenues et qui sont calculées pour chaque exemplaire ont été moyennées par catégorie de voyelle puis par locuteur. Les valeurs de ContrastLoss étaient toutefois faussées par une erreur de calcul qui a eu pour conséquence une surestimation de l'effet du style de parole sur cette métrique, avant de présenter les résultats obtenus à partir des valeurs corrigées je ne les reprends donc pas dans ce récapitulatif qui ne concerne que les cinq autres métriques sélectionnées.

Une distinction à trois niveaux a été relevée entre les voyelles courtes, de durée moyenne et longues pour les quatre métriques relatives à la centralisation du système vocalique pVSA, F1RR, F2RR et DistCentroid, la différence n'étant significative qu'entre les voyelles courtes d'une part et les voyelles moyennes et longues d'autre part pour la variabilité intra-catégorie V-Dispersion. Par ailleurs, à durée égale une distinction à deux niveaux a été relevée pour ces cinq métriques entre d'une part le corpus de parole conversationnelle NCCFr plus centralisé et associé à une moindre variabilité intra-catégorie, et le corpus de parole journalistique ESTER et celui de parole lue BREF d'autre part.

Une interprétation possible de cette différence entre la parole conversationnelle et les deux autres styles de parole pourrait être de la relier au caractère interactif ou non de la parole produite dans ces différents styles. Toutefois cette interprétation est à nuancer par le fait que le corpus ESTER inclut également des productions dans le cadre de débats.

4.2.1.3 Comparaison des effets du corpus et de la classe de durée à l'échelle de l'exemplaire

En raison du moyennage des valeurs de DistCentroid, VDispersion et ContrastLoss effectué préalablement à l'analyse statistique dans l'étude présentée en 2015, la variabilité intra-locuteur n'a pas été prise en compte dans l'analyse. De plus, comme évoqué précédemment les valeurs de ContrastLoss utilisées étaient biaisées par une erreur de calcul. Je propose donc ici une réanalyse de ces données en me concentrant sur les trois métriques calculées pour chaque exemplaire de voyelle, DistCentroid, VDispersion et ContrastLoss, dont je résume ci-dessous les principaux résultats. La Figure 16 illustre la distribution de ces métriques pour chacun des trois styles de parole et chacune des trois classes de durée avec les valeurs corrigées de ContrastLoss. Afin de préserver la lisibilité de la figure, j'ai fait le choix ici de représenter par des points individuels les valeurs moyennes par locuteur de chacune des métriques, pondérées par le nombre d'exemplaires dans chaque catégorie vocalique.

De même que dans l'étude de 2015, l'analyse statistique est effectuée au moyen d'un modèle de régression linéaire mixte pour chacune des métriques, avec le locuteur comme facteur aléatoire, et dans ce cas également la catégorie de voyelle comme facteur aléatoire afin de tenir compte du déséquilibre entre nombre d'exemplaires représentés dans les différentes catégories. Les deux facteurs aléatoires sont codés comme une ordonnée à l'origine aléatoire (*random intercept*) pour tenir compte des différences entre locuteurs ou entre catégories de voyelles de niveau moyen de centralisation, dispersion intra-catégorie ou recouvrement entre catégories. Les comparaisons post-hoc sont effectuées à partir des moyennes marginales estimées pour chacun des modèles mixtes, avec l'application d'une correction de Tukey pour les comparaisons par paire.

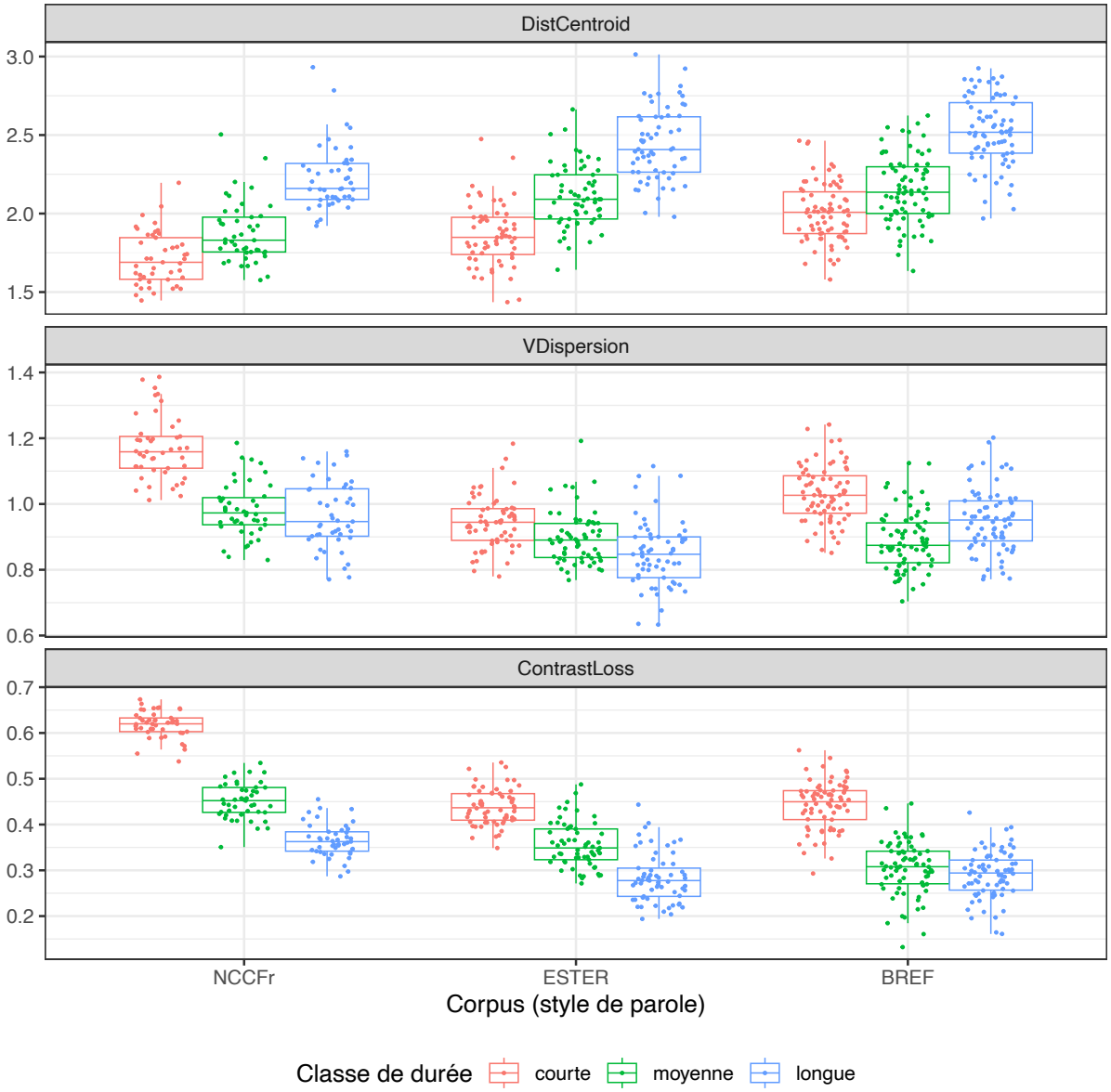


Figure 16 : Distribution des valeurs des métriques de l'espace vocalique DistCentroid, VDispersion et ContrastLoss, moyennées par locuteur après pondération entre catégories vocaliques, pour chacun des trois corpus représentant l'un des styles de parole comparés et chaque classe de durée. Les points individuels représentent la valeur moyenne pour chaque locuteur inclus dans le corpus, en complément des boîtes à moustache en trait fin qui n'incluent pas l'affichage des valeurs extrêmes pour éviter les confusions. Version corrigée et complétée de la figure présentée dans Audibert et al. (2015, [ACTI34]).

Le schéma général observé à partir des comparaisons statistiques correspond majoritairement aux prédictions, avec pour chacune des trois métriques un effet significatif du style de parole et de la classe de durée ainsi que de leur interaction. De plus, les comparaisons post-hoc par paire entre corpus et entre classes de durées des valeurs de DistCentroid révèlent un espace vocalique de plus en plus centralisé en allant de la lecture à la parole spontané, et d'autant plus centralisé que la voyelle est brève. Si pour les valeurs de VDispersion les différences entre paires de corpus et de classes de durées sont toutes significatives, la direction de ces différences est en partie contre-intuitive avec une variabilité intra-catégorie plus faible pour la parole journalistique que pour la parole lue, et légèrement plus élevée pour les voyelles

longues que pour les voyelles de durée moyenne. Enfin les comparaisons par paire des valeurs de ContrastLoss indiquent un degré de recouvrement entre catégories plus important pour la parole conversationnelle mais une absence de différence entre parole lue et journalistique, et un degré de recouvrement d'autant plus important que la voyelle est brève. Les comparaisons entre styles de parole au sein de chaque classe de durée confirment globalement les résultats obtenus précédemment à partir d'une valeur par locuteur et par classe de durée pour DistCentroid avec une distinction à trois niveaux uniquement pour les voyelles les plus courtes mais pas de différence entre ESTER et BREF pour les deux autres classes de durée.

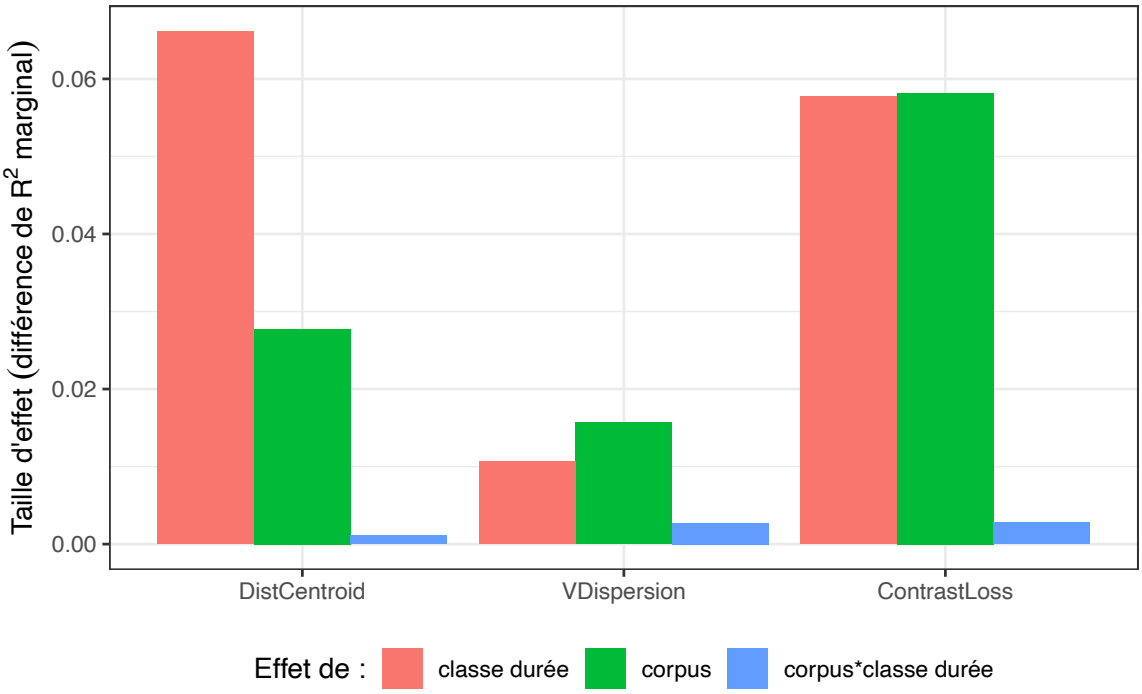


Figure 17 : Comparaison des tailles d'effet estimées comme la différence de valeur de R^2 marginal entre modèles incluant ou non l'effet évalué de la classe de durée, du corpus et de l'interaction entre corpus et classe de durée pour les métriques DistCentroid, VDispersion et ContrastLoss, calculées à partir des modèles linéaires mixtes à l'échelle de l'exemplaire.

En raison du nombre élevé de points de mesure, la significativité des effets principaux et de l'interaction est peu informative, puisque des différences modérées peuvent être significatives. Ce critère ne permet donc pas de hiérarchiser l'importance relative de l'effet du style de parole et de la durée sur l'organisation du système vocalique. Pour mieux rendre compte du poids relatif de ces deux facteurs et de leur interaction, nous avons donc estimé la taille de l'effet de chacun de ces facteurs pour chacune des trois métriques, en comparant la proportion de la variance expliquée par les effets fixés (R^2 marginal, estimé par la méthode de Nakagawa et al. (2017)) du modèle incluant ou non l'effet évalué. La Figure 17 illustre cette comparaison entre tailles d'effet.

On peut noter que pour les trois métriques, l'effet de l'interaction entre la classe de durée et le style de parole est beaucoup plus faible que celui de chacun des deux facteurs. La centralisation mesurée par DistCentroid dépend plus largement de la classe de durée que du style de parole, en revanche les effets de ces deux facteurs sur le degré de recouvrement entre catégories mesuré par ContrastLoss sont de magnitude comparable. On peut enfin noter que le style de parole a un effet légèrement plus important que la classe de durée sur la dispersion

intra-catégorie mesurée par VDispersion, avec toutefois des effets de magnitude réduite par rapport aux effets observés pour les deux autres métriques, ce qui amène à nuancer les observations relatives à la variation de la dispersion intra-catégorie dans ces données.

4.2.1.4 Un aperçu de la variation inter-locuteur

Dans les analyses réalisées, aussi bien pour l'article publié en 2015 que dans la nouvelle analyse présentée ci-dessus, la variation inter-locuteur au sein de chaque style de parole représenté par l'un des corpus a été neutralisée en la considérant comme un effet aléatoire. Toutefois, comme le suggère la Figure 16, les caractéristiques des différents locuteurs en termes de centralisation de l'espace vocalique, de variation intra-catégorie et de chevauchement entre catégories ne sont pas uniformes au sein de chaque style de parole et de chaque classe de durée. Quels que soient le corpus et la métrique, on peut ainsi noter un recouvrement conséquent entre classes de durée, ainsi qu'entre corpus pour une même classe de durée.

L'une des limites de l'analyse statistique réalisée, à la fois dans la version publiée de l'étude et dans la nouvelle analyse présentée, est que si le lien entre mesures correspondant à un même locuteur est pris en compte comme facteur aléatoire dans l'analyse, il ne l'est qu'en tant qu'ordonnée à l'origine aléatoire. Cela revient à prendre en compte les différences interindividuelles moyennes de degré de centralisation (respectivement de variabilité intra-catégorie et de recouvrement entre catégories) mais pas les éventuelles spécificités individuelles quant aux liens entre classes de durées. De tels modèles statistiques nécessiteraient l'utilisation d'une pente aléatoire, qui n'est toutefois pas applicable sur nos données en raison de la complexité des modèles correspondants dont la convergence ne peut être assurée. L'utilisation d'un modèle de régression bayésienne serait envisageable, mais consisterait en une neutralisation de la variation individuelle plutôt qu'en une analyse de cette variation.

Je propose ici une analyse descriptive de cette variation individuelle, illustrée par la Figure 18. Pour chacune des trois métriques la tendance majoritaire est à une évolution dans la même direction entre voyelles courtes et de durée moyennes et entre voyelles de durée moyennes et longues : croissante pour DistCentroid, décroissante pour VDispersion et ContrastLoss, avec toutefois une variabilité plus importante pour VDispersion qui pourrait expliquer en partie les effets plus faibles observés pour cette métrique.

L'analyse de l'évolution des valeurs des métriques DistCentroid et ContrastLoss entre les voyelles les plus brèves et celles les plus longues indique que l'ensemble des locuteurs du corpus de parole conversationnelle NCCFr suivent cette tendance majoritaire. Néanmoins, un certain nombre de contre-exemples peuvent être observés. Ainsi, même dans le cas de la dispersion/centralisation mesurée par DistCentroid pour laquelle le patron observé est le plus systématique, un locuteur du corpus ESTER et deux locuteurs du corpus BREF présentent des voyelles de durée moyennes plus centralisées que les voyelles courtes. Par ailleurs deux autres locuteurs du corpus ESTER présentent un degré de recouvrement entre catégorie mesuré par ContrastLoss plus important pour les voyelles de durée moyenne que pour les voyelles les plus brèves, et deux autres encore un degré de recouvrement plus important pour les voyelles longues que pour celles de durée moyenne, ce qui est également le cas de 29 des 74 locuteurs du corpus BREF.

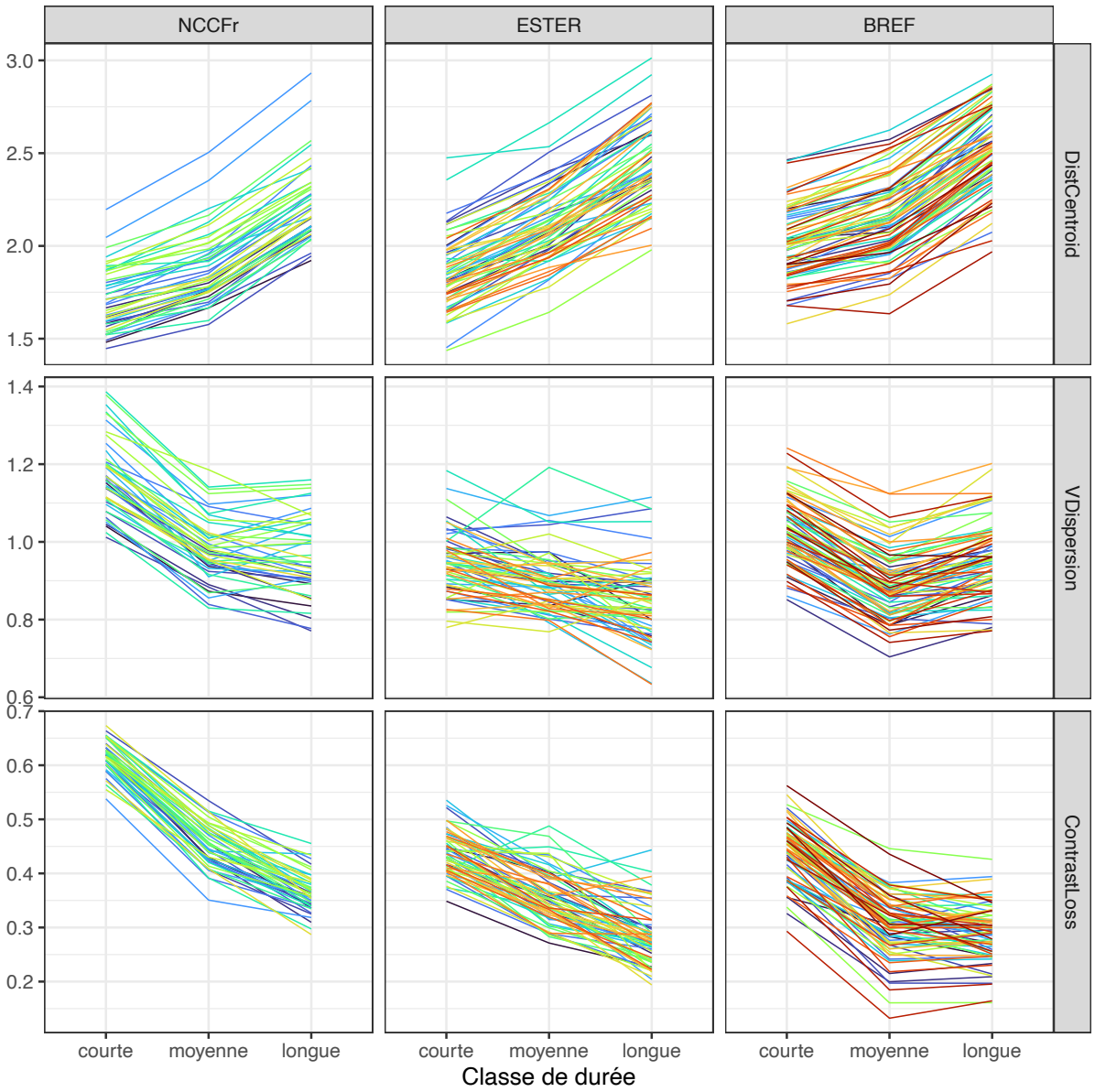


Figure 18 : Tracés individuels (une série de deux segments consécutifs par locuteur) des liens entre les trois classes de durée pour les valeurs des trois métriques DistCentroid, VDispersion et ContrastLoss moyennées par locuteur après pondération entre catégories vocaliques, dans chacun des trois corpus comparés.

On peut s'interroger sur les raisons de ces variations, qui pourraient être liées à l'influence du contexte segmental, particulièrement complexe à contrôler dans des corpus de parole continue de grande dimension. Bien que cette interprétation ne puisse être exclue, elle n'apparaît pas comme une explication évidente à la lumière d'une inspection qualitative des contextes segmentaux représentés dans les contre-exemples identifiés comparativement aux contextes représentés dans les voyelles produites par les locuteurs qui pour leur part suivent la tendance générale.

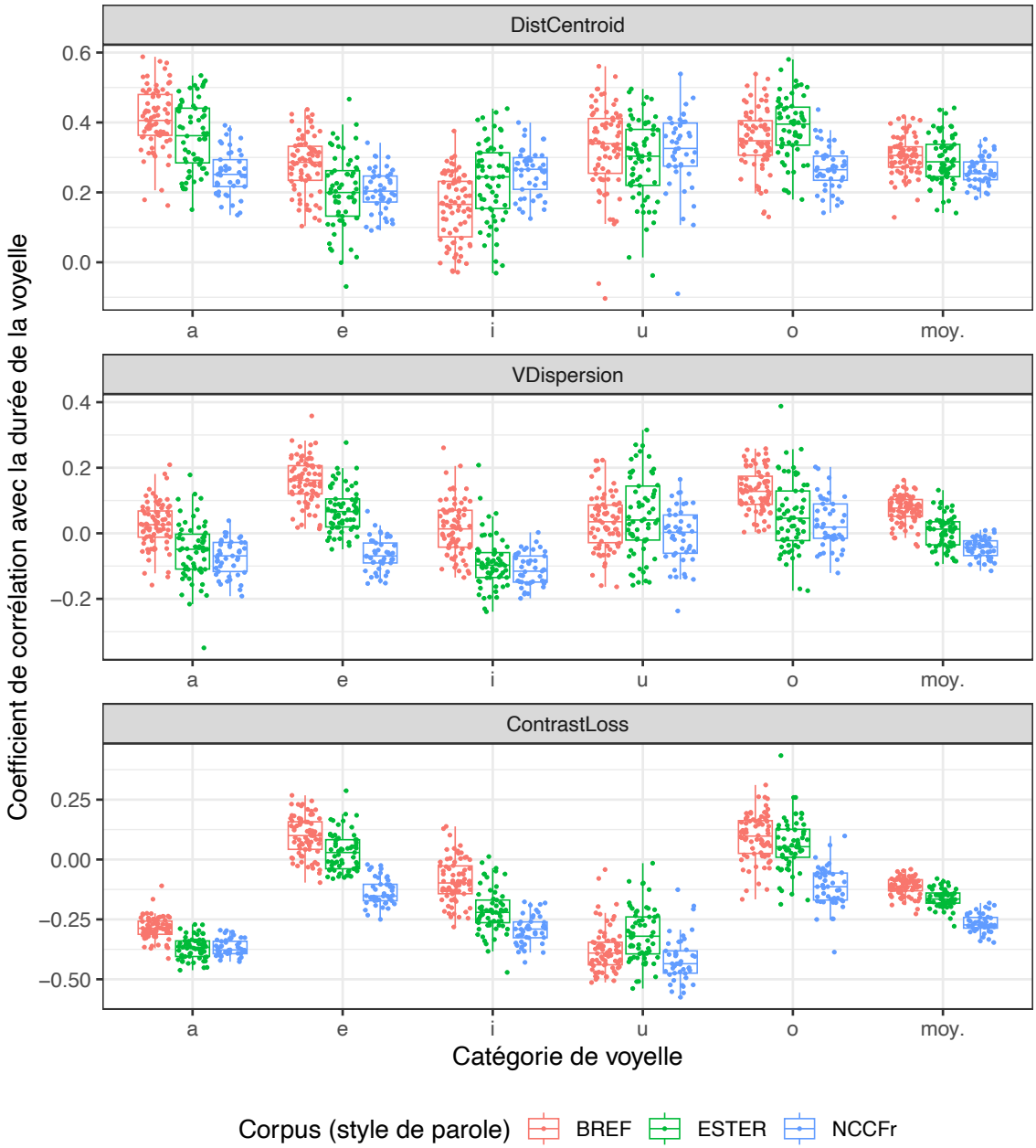


Figure 19 : Coefficient de corrélation entre valeurs prises par les trois métriques DistCentroid, VDispersion et ContrastLoss et la durée de la voyelle, calculé pour chaque locuteur et catégorie de voyelle et comparé entre corpus et catégories de voyelles. La distribution des corrélations dans chaque catégorie vocalique est complétée par la distribution de la moyenne par locuteur sur l'ensemble des catégories, étiquetée « moy. ».

Une autre analyse possible de la variation inter-locuteur repose sur la comparaison inter-individuelle du lien entre la durée des voyelles et les valeurs prises par nos trois métriques. Pour les besoins de cette analyse, les valeurs des métriques ont été recalculées en considérant l'ensemble des voyelles sélectionnées de chaque locuteur sans les subdiviser en classes de durée, afin d'obtenir le coefficient de corrélation entre la durée de la voyelle et les valeurs prises par les métriques pour chaque catégorie de voyelle. La Figure 19 illustre la distribution entre locuteurs des coefficients de corrélation entre durée et valeurs prises par les métriques dans les trois corpus analysés, pour chaque catégorie de voyelle ainsi qu'en considérant la corrélation moyenne sur l'ensemble des cinq catégories de voyelles pour chaque locuteur.

En considérant séparément les catégories de voyelles, la variabilité entre locuteurs est très conséquente, avec des locuteurs pour lesquels une corrélation inverse à la tendance majoritaire est observée. Pour la métrique DistCentroid, c'est ainsi le cas de quelques locuteurs des corpus BREF et ESTER pour les voyelles /i/ et /u/, et de deux locuteurs du corpus ESTER pour la catégorie fusionnée /e, ɛ/ : pour ces locuteurs et ces catégories de voyelle, la dispersion a ainsi légèrement tendance à augmenter lorsque les durées de voyelles sont plus réduites. Dans le cas de la métrique ContrastLoss, une corrélation inverse à la tendance majoritaire avec une distinctivité plus importante pour les voyelles les plus brèves est observée pour quelques locuteurs du corpus BREF sur la voyelle /i/ mais surtout pour la majorité des locuteurs des corpus BREF et ESTER sur les catégories fusionnées de voyelles /e, ɛ/ et /o, ɔ/. Le statut de ces catégories fusionnées ne semble pas pouvoir expliquer cette observation contre-intuitive, mais on peut s'interroger sur un éventuel effet lexical qui conduirait les réalisations de ces voyelles dans des mots fréquents souvent réalisés plus brefs à être représentés comme plus typiques par l'analyse linéaire discriminante à la base du calcul de cette métrique. Dans le cas de la métrique VDispersion, aucun lien clair entre durée segmentale et variation inter-catégorie n'émerge, avec des corrélations faibles et réparties entre corrélations positives et négatives.

Si on considère la corrélation moyenne par locuteur sur les cinq catégories de voyelles, la variabilité est réduite comparativement aux corrélations séparées par catégorie, avec pour DistCentroid et ContrastLoss des corrélations dont la direction suit la tendance majoritaire pour l'ensemble des locuteurs. La variabilité inter-locuteur reste toutefois conséquente, notamment pour les corrélations entre la durée segmentale et les valeurs de DistCentroid avec des corrélations allant de r=.13 à r=.44, sans différence notables entre les trois corpus hormis une variation inter-individuelle moindre pour le corpus NCCFr (qui pourrait toutefois être due en partie au nombre de locuteurs plus faible dans ce corpus que dans les deux autres). Ainsi, dans les trois styles de parole la tendance à plus centraliser les segments vocaliques plus courts est très variable entre locuteurs.

Si les corrélations avec la durée segmentale sont globalement plus modérées et moins variables pour les valeurs de la métrique ContrastLoss, elles apparaissent comme plus dépendantes du style avec des corrélations plus importantes en valeur absolue dans le corpus NCCFr que dans les deux autres. La tendance à moins bien préserver les contrastes entre voyelles sur les voyelles plus courtes est donc plus prononcée en parole conversationnelle qu'en parole lue ou en parole journalistique.

4.2.2 Effet du style de parole sur les variations intra-locuteur

Publication associée :

[ACTI29] Lancien, M., **Audibert, N.,** & Fougeron, F. (2018). Effet de la situation de parole sur la variabilité des voyelles en français. *Actes des 32èmes Journées d'Études sur la Parole*, Aix-en-Provence, France, pp. 338-346.

Si le recours à des corpus de parole continue de grande taille dans une approche telle que celle présentée ci-dessus permet une estimation plus pertinente de la variabilité inter-locuteurs à travers l'inclusion d'un nombre important de locuteurs pour chacun des styles de parole étudiés, on touche ici aux limites d'une approche dans laquelle les différents styles de parole sont représentés par différents locuteurs. En complément d'une telle approche, une

approche expérimentaliste plus classique reste nécessaire afin d'appréhender plus directement l'effet du style de parole sur la réalisation des voyelles en contrôlant les variations inter-locuteurs.

Le travail présenté dans cette section est issu du mémoire de master de Mélanie Lancien que Cécile Fougeron et moi avons coencadré lors de l'année universitaire 2016-2017. L'objectif était de documenter plus finement la variation vocalique entre styles de parole au-delà de la distinction entre parole dite « claire » et parole de laboratoire, en comparant directement les productions des mêmes locuteurs dans différents styles de parole afin de contrôler ce facteur de variation. Une condition de jeu interactif a été incluse afin d'évaluer l'incidence de l'interaction sur la modification de l'espace vocalique, en comparaison de trois tâches de lecture : une condition de lecture sans consignes spécifiques, étiquetée « lecture normale », une condition de lecture rapide afin d'évaluer l'impact de l'augmentation de débit, et une condition de lecture adressée à une personne malentendante comparable à l'une des conditions de parole claire documentée dans la littérature.

Huit locutrices francophones natives ont été enregistrées dans quatre conditions de production. De façon similaire aux critères retenus pour la constitution du corpus NCCFr (Torreira et al., 2010), les locutrices ont été sélectionnées par binômes entretenant un lien amical. De plus, afin de s'assurer de leur implication dans la tâche de jeu, leur habitude de participation à des jeux interactifs et leur compétitivité dans ce type de jeu ont été également été retenues comme critères de sélection. Le jeu interactif proposé reposait sur un ensemble de cartes incluant un mot à faire deviner à l'autre joueuse en un temps limité, à l'aide d'autres mots présentés comme des indices rapportant un nombre plus élevé de points afin de susciter leur utilisation dans le cadre du jeu. Ces mots cibles bisyllabiques, également insérés dans les tâches de lecture, étaient sélectionnés en fonction de la voyelle présente en syllabe finale et de leur contexte consonantique, les consonnes coronales étant privilégiées autant que possible. Ils étaient conçus pour que l'ensemble des voyelles orales du français soient présentes en syllabe finale de mot avec une distribution homogène entre voyelles.

En raison de fluctuations dans le déroulement du jeu, certains mots-cibles n'ont pas toujours été produits. Afin d'assurer l'inclusion d'un minimum de cinq exemplaires de chacune des voyelles analysées dans chacune des conditions de production, les productions de l'une des locutrices ont été exclues, de même que les voyelles /ø/ et /e/ pour lesquelles le nombre d'exemplaires exploitables était insuffisant. Les analyses ont donc été effectuées sur un total de 1702 exemplaires des voyelles orales /i, y, ɛ, œ, a, ɔ, o, u/ produits par sept locutrices.

Une mesure de centralisation dans l'espace acoustique des deux premiers formants, et une mesure de variabilité intra-catégorie dans l'espace des trois premiers formants ont été extraites après conversion en Bark des positions formantiques, pour chaque exemplaire de voyelle après regroupement par locutrice et par condition de production. Bien que nous ne les ayons pas désignés par ces noms dans l'article, on peut noter que ces deux mesures calculées comme des distances aux centroïdes correspondent respectivement aux métriques DistCentroid et VDispersion présentées en section 7.2.3.2. Nous avons fait ici le choix pour les mesures de centralisation de nous limiter aux deux premiers formants en raison du manque d'informations sur l'interprétation possible d'une centralisation au-delà du plan F1*F2, toutefois VDispersion prend également en compte les mesures de F3.

La comparaison statistique des deux métriques ainsi que de la durée segmentale entre styles de parole et catégorie de voyelle, effectuée à l'aide l'un modèle linéaire mixte dans

lequel la locutrice et le contexte consonantique étaient pris en compte comme facteurs aléatoire, a montré un fort effet du style de parole sur la durée, et un effet plus modéré sur les caractéristiques spectrales des voyelles, avec un effet plus important sur la centralisation que sur la dispersion intra-catégorie. La Figure 20, dans laquelle les voyelles antérieures arrondies ne sont pas représentées, illustre la projection dans le plan acoustique F1*F2 des espaces vocaliques moyens dans les quatre styles de parole. La lecture rapide est caractérisée par une compression de F1 et F2 et la lecture pour une personne malentendante au contraire par une expansion de F1 et F2, tandis que les espaces vocaliques en condition de jeu et de lecture normale sont très similaires. On peut noter également que les réalisations de /i/ sont moins impactées par les changements de style que celles des autres voyelles.

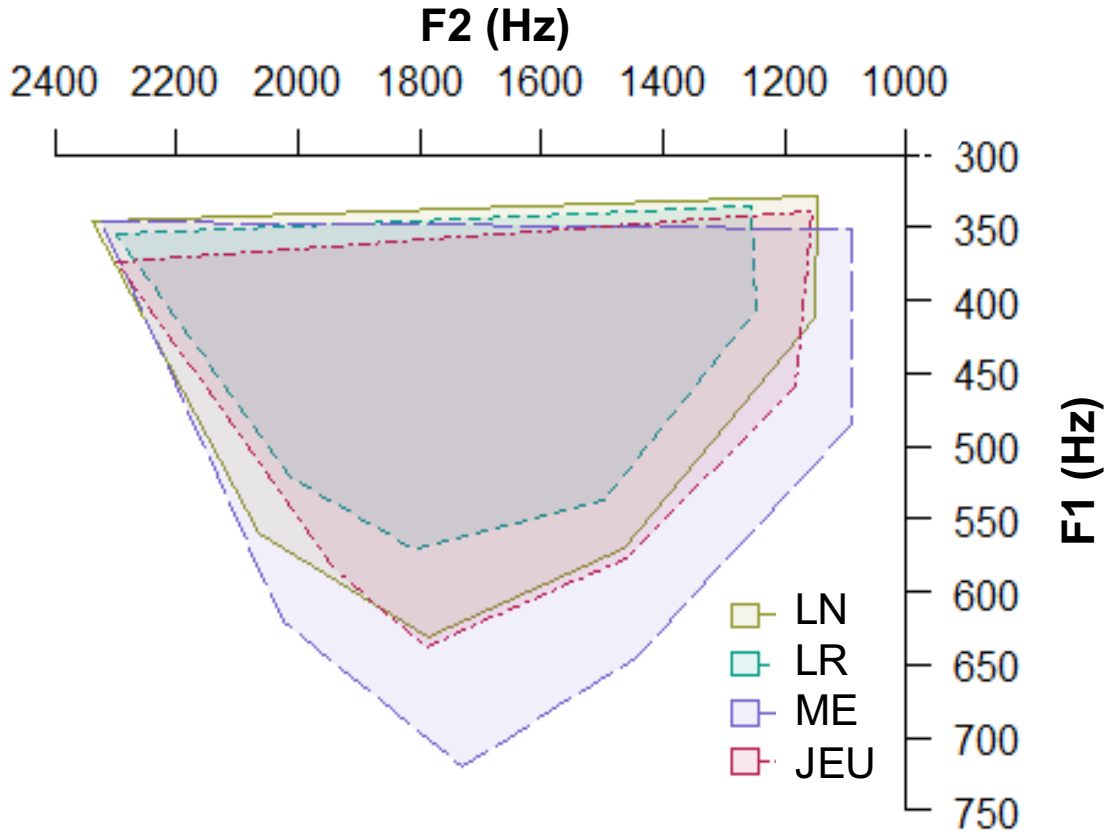


Figure 20 : Espaces vocaliques moyens dans l'espace acoustique des deux premiers formants pour les sept locutrices incluses dans l'étude, comparés entre les quatre styles de parole : LN = lecture normale ; LR = lecture rapide ; ME = lecture pour une personne malentendante ; JEU = condition de jeu interactif. Les sommets des polygones représentent les centroïdes des voyelles /i, ɛ, a, ɔ, o, u/. Adapté de Lancien et al. (2018, [ACTI29]).

Une autre observation intéressante est que bien que l'espace vocalique dans le jeu interactif soit très proche de celui observé en lecture normale, cette condition de jeu est par ailleurs caractérisée par des segments vocaliques aussi brefs que ceux produits en condition de lecture rapide. La comparaison de ces deux styles constitue donc une autre illustration de la relative indépendance de la durée segmentale et du degré de centralisation du système vocalique, confirmée par les faibles corrélations entre ces deux dimensions (ces corrélations étant encore plus faibles que celles relevée dans l'étude sur grand corpus présentée précédemment, de r=-.1 pour la condition de jeu à r=-.25 pour la lecture pour une personne malentendante).

Si la variabilité intra-catégorie, également faiblement liée à la durée, est globalement plus importante en condition de jeu et plus faible en condition de lecture pour une personne malentendante, d'importantes différences inter-individuelles entre les sept locutrices peuvent être observées comme illustré par la Figure 21, avec notamment deux locutrices qui présentent des voyelles plus variables au sein d'une même catégorie en condition de lecture rapide qu'en condition de jeu. L'utilisation de la centralisation/dispersion dans les différents styles est plus consistante entre locutrices, avec toutefois une locutrice qui centralise autant ses voyelles en condition de jeu qu'en condition de lecture rapide, et une locutrice qui joue principalement sur l'augmentation de la durée en condition de lecture pour une personne malentendante avec une expansion plus limitée de l'espace vocalique, et surtout une importante variabilité intra-catégorie. Cette dernière observation va en sens inverse de l'interprétation la plus évidente du lien entre variabilité intra-catégorie et degré de clarté de la parole produite, maximale dans les productions destinées à une personne malentendante. En effet, on peut supposer qu'à taille d'espace vocalique égale, une variabilité intra-catégorie réduite induit une parole plus claire en limitant les chevauchements entre catégories. Une autre interprétation possible serait que l'augmentation de la variabilité puisse être liée chez certaines locutrices à un degré plus important de coarticulation, qui peut contribuer à l'amélioration de l'intelligibilité (Scarborough & Zellou, 2013).

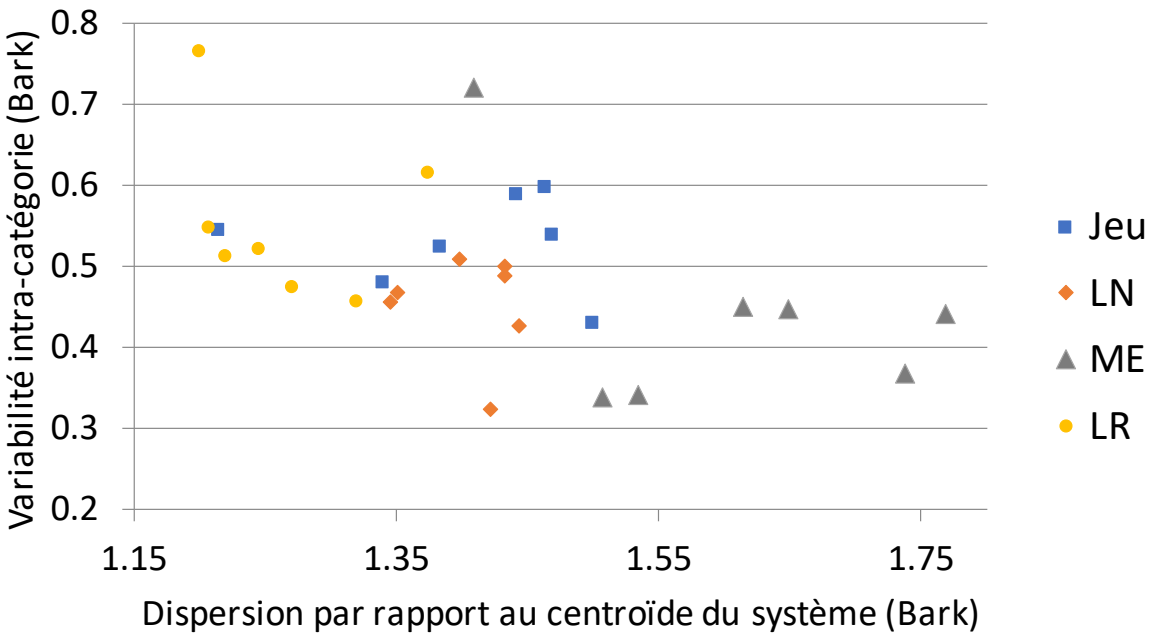


Figure 21 : Distribution des valeurs moyennes de dispersion/centralisation (DistCentroid) et de variabilité intra-catégorie (VDispersion) pour les sept locutrices incluses dans l'étude et les quatre styles de parole. Figure non-publiée adaptée d'une version intermédiaire de Lancien et al. (2018, [ACTI29]).

4.2.3 Style de parole et neutralisation de contrastes vocaliques

Publication associée :

[ACL4] Gendrot, C., & **Audibert, N.** (2019). La distinction /e/ vs /ɛ/ en français standard est-elle maintenue en finale de mot ? Étude sur des corpus de parole journalistique et de parole spontanée. *Langue française*, 203(3), 53-65.

4.2.3.1 Méthodologie et principaux résultats

En collaboration avec Cédric Gendrot, nous avons étudié plus spécifiquement le cas du contraste entre /e/ et /ɛ/ en français standard afin d'évaluer à partir de deux grands corpus dans quel mesure ces deux voyelles moyennes sont dans un processus de fusion, comme cela a été suggéré dès le milieu des années 1980 pour le français parisien populaire (Landick, 1995). Nous nous sommes pour cela appuyés sur les données du corpus NCCFr de parole spontanée (Torreira et al., 2010) et sur celles du corpus ESTER de parole journalistique (Galliano et al., 2009), en faisant l'hypothèse que si ce processus de fusion était avéré, il devrait être plus marqué en parole spontané que dans la parole journalistique associée à un degré d'articulation plus important. Par ailleurs nous avons fait le choix de nous concentrer sur les voyelles produites en position finale de mot, supposées plus propices au maintien de l'opposition. Un total de 110 mots grammaticaux fréquents, représentant 21% de l'ensemble des occurrences de /e/ et /ɛ/, ont été exclus de l'analyse car susceptibles d'être fortement réduits (voir par exemple Vasilescu et al. (2012)) et donc peu représentatifs du phénomène ciblé. Par ailleurs nous avons également pris en compte deux cas particuliers : la distinction entre les mots « et » et « est », et celle entre les infinitifs en « _er » et les flexions en « _ai* ». Un seuil de 15 exemplaires par locuteur et par catégorie dans les différentes conditions comparées a été retenu.

Suite à l'extraction des mesures formantiques et à leur filtrage selon la méthode de Gendrot & Adda-Decker (2005), le degré de fusion entre /e/ et /ɛ/ a été évalué via la comparaison avec la voyelle /a/. Cette comparaison a été effectuée à partir de la projection des voyelles dans l'espace acoustique de F1 et F2, et via la mesure de distances euclidiennes dans le plan F1*F2 après conversion en Bark (Traunmüller, 1990) des fréquences formantiques, entre /ɛ/ et /a/ d'une part et entre /e/ et /ɛ/ pour chacune des trois conditions d'autre part.

Comme illustré par les ellipses de dispersion moyennes représentées sur la Figure 22 pour les trois catégories vocaliques dans le corpus NCCFr, le recouvrement est plus important entre /e/ et /ɛ/ qu'entre /ɛ/ et /a/. De plus, les distances mesurées entre /e/ et /ɛ/ en position finale et dans les deux cas particuliers étaient en moyenne au moins deux fois plus réduites que celles entre /ɛ/ et /a/ en position finale. Ces observations tendaient donc à confirmer l'hypothèse d'une fusion en cours entre /e/ et /ɛ/.

Dans l'ensemble des conditions à l'exception de la distinction entre « et » et « est » pour les femmes, les distances acoustiques mesurées étaient plus faibles dans le corpus NCCFr que dans le corpus ESTER. Si la réduction des distances entre /e/ et /ɛ/ pourrait être interprétée comme une confirmation d'un processus de fusion plus avancé en parole spontanée qu'en parole journalistique, cette interprétation est à nuancer par la distance plus réduite en parole spontanée également en /ɛ/ et /a/, qui irait plutôt dans le sens d'une réduction plus générale en parole spontanée.

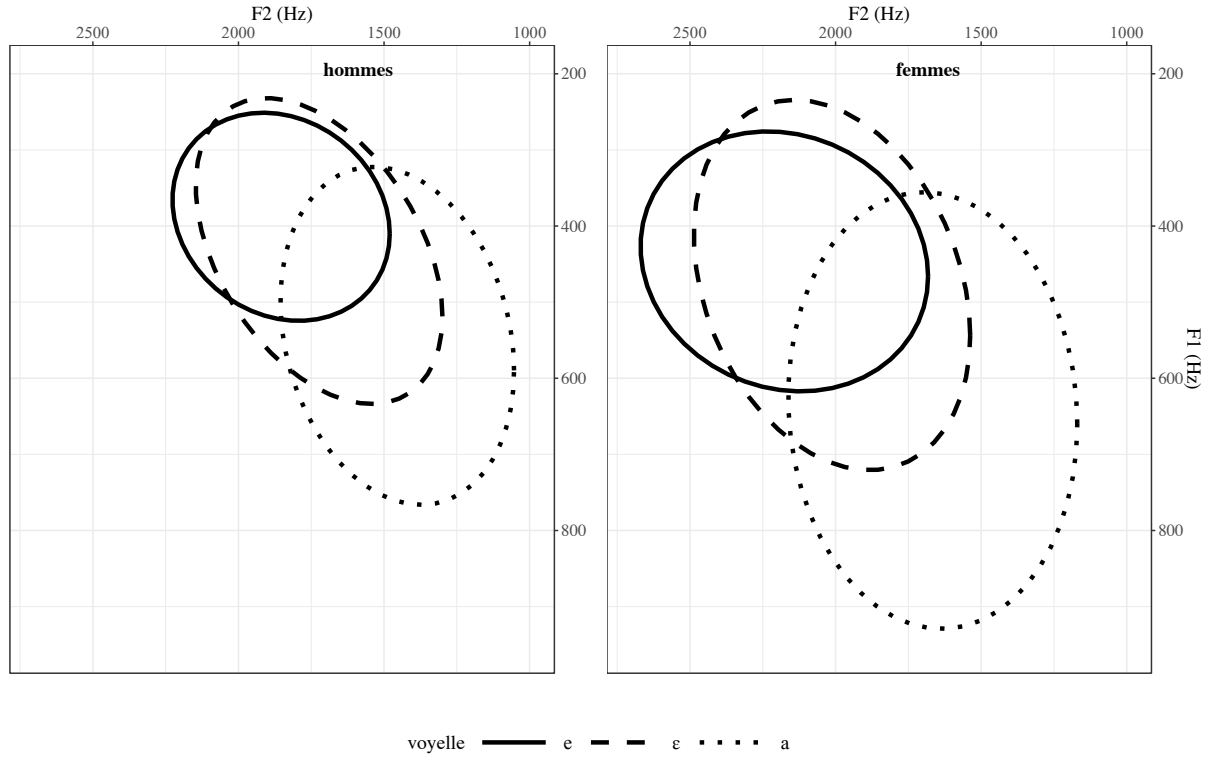


Figure 22 : Ellipses de dispersion à 95% moyennes observées pour les voyelles /e/, /ɛ/ et /a/ produites en position finale par les hommes (gauche) et les femmes (droite) du corpus NCCFr. D'après Gendrot & Audibert (2019, [ACL4]).

4.2.3.2 Variation entre locuteurs et lien entre distances acoustiques

Cette étude s'est appuyée sur les distances entre catégories vocaliques moyennes pour chaque locuteur, moyennées ensuite entre locuteurs. Bien que cela permette de donner le même poids à chaque locuteur dans chacun des deux styles, le moyennage entre locuteurs peut avoir pour effet de masquer la variation individuelle. Je propose donc une nouvelle visualisation de la variation entre locuteurs de ces distances (Figure 23). Comme l'illustre cette figure, au-delà des tendances majoritaires déjà observées initialement, on peut remarquer pour le corpus de parole journalistique ESTER une distribution bimodale de la plupart des distances entre /e/ et /ɛ/ tandis que ces distributions sont majoritairement unimodales pour le corpus de parole spontanée NCCFr. Cette observation suggère que même en parole journalistique, une partie des locuteurs se trouveraient déjà à un stade plus avancé de ce processus de fusion entre /e/ et /ɛ/, avec une neutralisation plus massive du contraste entre infinitifs en « _er » et flexions en « _ai* » que celle observée pour les autres contrastes en /e/ et /ɛ/. Cette observation soulève la question d'un éventuel effet générationnel au-delà de celui du style de parole, les locuteurs du corpus ESTER étant majoritairement moins jeunes que ceux de NCCFr.

Il est également possible d'approfondir certaines des analyses réalisées pour l'étude publiée en 2019. Les corrélations entre la distance entre /a/ et /ɛ/ en position finale considérée comme référence et celle entre /e/ et /ɛ/ en position finale indiquent qu'à l'exception des femmes du corpus NCCFr pour lesquelles cette corrélation est beaucoup plus faible (r = -.10), les locuteurs pour lesquels l'écart entre /a/ et /ɛ/ est le plus important ont tendance à montrer une distinction réduite entre /e/ et /ɛ/, avec des corrélations comprises entre r = -.53 et

r = -.63. Ceci suggère que chez les locuteurs qui neutralisent le plus le contraste entre /e/ et /ɛ/, cette neutralisation se manifeste majoritairement par une fermeture plus importante des réalisations de /ɛ/, d'où l'augmentation de la distance entre /a/ et /ɛ/. Autrement dit, les locuteurs chez qui ce contraste est le plus neutralisé auraient plus tendance à produire la voyelle /ɛ/ comme un /e/ plutôt qu'à converger vers une réalisation intermédiaire.

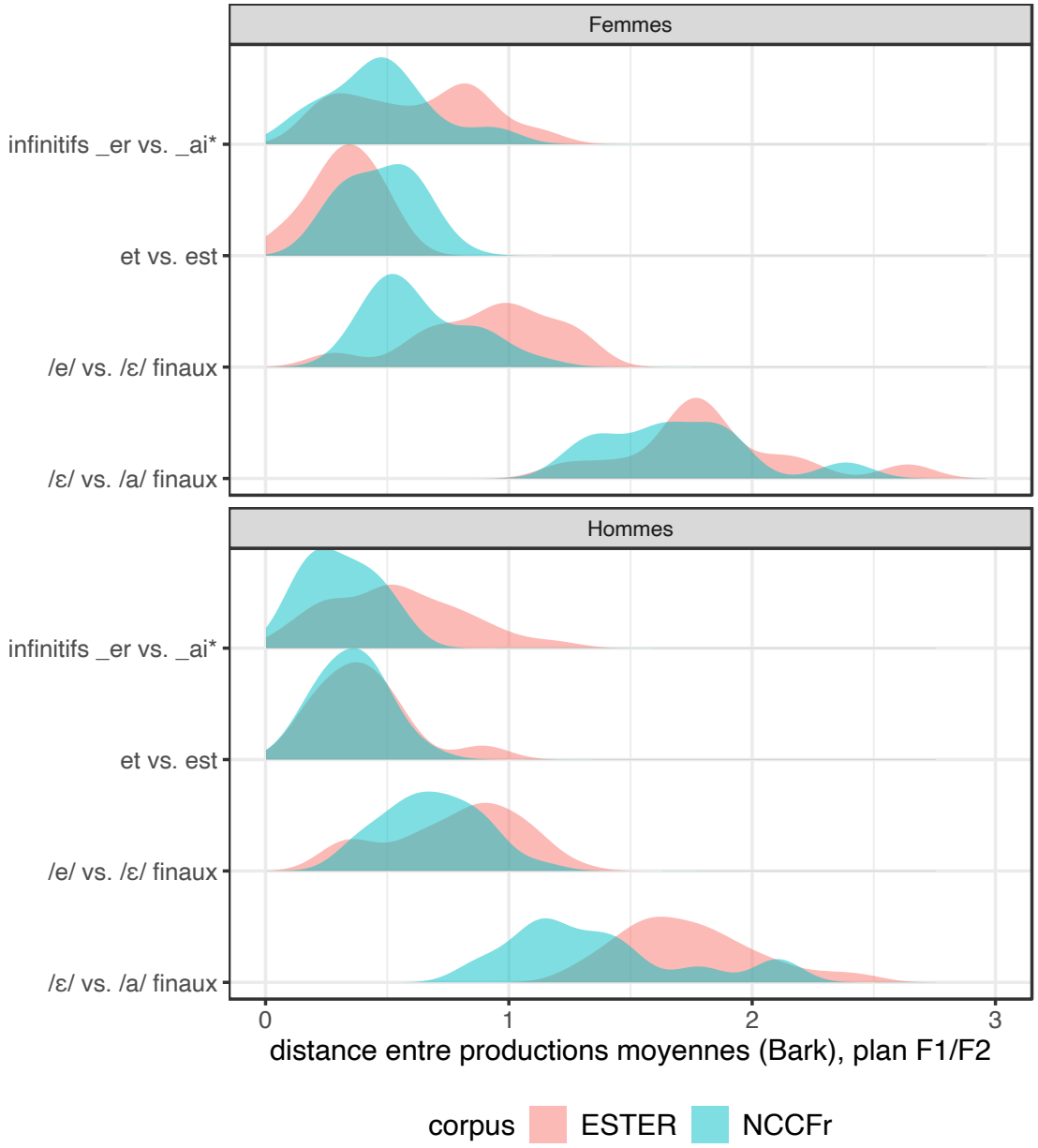


Figure 23 : Distribution des distances entre catégories vocaliques calculées pour chaque locuteur des corpus ESTER et NCCFr, séparément pour les femmes (haut) et les hommes (bas).

Enfin, le calcul par locuteur du rapport entre les différentes distances /e/ vs. /ɛ/ et la distance /a/ vs. /ɛ/ en position finale fournit une mesure normalisée du degré de réduction du contraste entre /e/ et /ɛ/. La comparaison entre corpus de ces rapports de distances indique que si pour les femmes le maintien du contraste entre /e/ vs. /ɛ/ en parole journalistique est plus courant quoique minoritaire, pour les hommes il est tout aussi neutralisé en parole journalistique qu'en parole spontanée. De plus, en parole journalistique le contraste entre « et » et « est » apparaît comme autant voire plus neutralisé qu'en parole spontanée.

4.3 Le cas de la parole et du chant codifiés adressés au nourrisson

Publications associées :

[ACL3] Falk, S., & **Audibert, N.** (2021). Acoustic signatures of communicative dimensions in codified mother-infant interactions. *The Journal of the Acoustical Society of America*, 150(6), 4429-4437.

[ACTI27] **Audibert, N.**, & Falk, S. (2018). Vowel space and f0 characteristics of infant-directed singing and speech. *Proceedings of the 9th International Conference on Speech Prosody*, Poznan, Pologne, pp. 153-157.

4.3.1 Parole et chant adressés à l'enfant et but communicatif

Dans le cadre d'une collaboration avec Simone Falk, amorcée lorsqu'elle était membre du Laboratoire de Phonétique et Phonologie, nous avons analysé les productions codifiées de chant et de parole de mères allemandes de jeunes enfants d'environ six mois. L'utilisation de ces productions codifiées, notamment les chants ou comptines, est d'un usage courant dans les sociétés occidentales dans le cadre des interactions avec les enfants préverbaux et leur rôle dans le développement cognitif des nourrissons est reconnu (voir par exemple Cirelli et al. (2020)). Par ailleurs, ces productions codifiées sont associées à des situations de communication spécifiques (Trehub & Gudmundsdottir, 2015).

En dépit de leur importance dans la communication adressée à l'enfant, les caractéristiques acoustiques spécifiques des productions codifiées ont peu été étudiées auparavant. Dans le cas du chant on peut cependant mentionner l'étude de Bergeson & Trehub (2002) qui ont montré que du fait de la répétition fréquente de ces productions codifiées, elles se caractérisent par une grande stabilité de la hauteur mélodique mais aussi du tempo, y compris dans le cas de productions espacées dans le temps. Kalashnikova et al. (2017) ont suggéré que dans la parole adressée à l'enfant, la présence de patrons caractéristiques (en l'occurrence le renforcement de la distinction acoustique entre voyelles due à un raccourcissement du conduit vocal via la montée du larynx) pourrait être destinée en priorité à la régulation affective et à la transmission de buts communicatifs. Dans cette étude nous avons donc cherché à identifier des formes acoustiques caractéristiques de productions codifiées adressées aux jeunes enfants, en nous concentrant sur les productions vocaliques et en comparant directement trois dimensions communicatives étudiées séparément auparavant : le caractère dirigé vers le jeune enfant ou non, la nature parlée ou chantée de la production, et le type de stimulation visant à faire jouer l'enfant ou à l'amuser, ou au contraire à le calmer ou l'endormir.

Parmi les caractéristiques acoustiques associées à ces dimensions dans la littérature, la spécificité de la parole adressée à l'enfant a été la plus largement étudiée, l'élévation de la fréquence fondamentale et de sa variabilité comparativement à la parole adressée à l'adulte en étant les caractéristiques la plus saillantes (voir par exemple Narayan et al. (2016)), le ralentissement du débit de parole mesuré dans certaines études pouvant être également interprété comme lié avant tout à la longueur des énoncés (A. Martin et al., 2016). De plus, le degré de coordination temporelle des productions adressées à l'enfant a été évalué par Falk & Kello (2017) comme plus important que celui mesuré dans les mêmes productions adressées à l'adulte. Les résultats de la littérature concernant la réalisation des voyelles sont moins clairs. En effet, si de nombreux auteurs concluent à des différences accrues entre les voyelles

périphériques /a, i, u/ dans la parole adressée à l'enfant (voir par exemple Kuhl et al. (1997)), ce qui a conduit à faire l'hypothèse d'une hyperarticulation destinée à renforcer les contrastes phonologiques présentés à l'enfant, cette hypothèse a été remise en question par A. Martin et al. (2015) à partir de l'observation d'une variabilité intra-catégorie plus importante dans la parole adressée à l'enfant.

Les études portant sur les différences entre voix parlée et voix chantée dans les types de chants destinés aux jeunes enfants sont peu nombreuses. Le chant destiné à l'enfant a été décrit comme caractérisé par un rythme plus lent et une variabilité de la fréquence fondamentale moindre que dans la parole dirigée vers l'enfant par Tsang et al. (2017), qui ont également montré une préférence d'enfants de 6 à 10 mois pour ce type de production chantée comparativement à la parole. Par ailleurs l'une des caractéristiques du chant classique comparativement à la parole est la compression de l'espace vocalique accompagnée d'une moindre variabilité de la réalisation des cibles vocaliques (Bradley, 2018), toutefois la question de la transposition de cette caractéristique au chant adressé aux jeunes enfants reste ouverte.

Dans différentes langues et cultures, les interactions stimulantes associées aux situations de jeu ont été associées à des contours mélodiques montants majoritaires, contrairement aux productions destinées à calmer l'enfant (Falk, 2011). Par ailleurs dans le chant, les berceuses sont associées à des rythmes lents et une hauteur mélodique plus basse, au contraire des comptines (Trainor et al., 1997), et les enfants montrent une préférence pour ces caractéristiques (Conrad et al., 2011). Toutefois, cette dimension n'a pas été associée dans la littérature à des différences d'articulation des voyelles.

4.3.2 Données analysées et principaux résultats

Quinze mères germanophones natives âgées en moyenne de 31,8 ans ont été enregistrées dans des conditions contrôlées, toutes étant habituées à chanter quotidiennement pour leur enfant. Les enfants, âgés de six mois au moment des enregistrements, étaient présents lors des enregistrements et les productions de leur mère leur étaient directement adressées. Quatre productions codifiées, toutes issues du répertoire traditionnel allemand d'histoires et de chansons pour enfants et bien connues des mères, ont été utilisées pour représenter les interactions stimulantes ou apaisantes en chant et parole : deux productions chantées (une comptine et une berceuse) et deux productions parlées (une histoire rimée amusante et une histoire lue apaisante). En complément des productions en présence de l'enfant à proximité immédiate de la mère, chacune des quatre productions codifiées était produite en son absence à destination de l'expérimentatrice, lors de sessions d'enregistrement à domicile dans un environnement calme. Dans la condition en présence de l'enfant, l'effet stimulant de la comptine et de l'histoire rimée a été validé par l'observation des mouvements et vocalisations de l'enfant, significativement plus nombreux que ceux observés pendant les productions de la berceuse et de l'histoire lue.

Les paroles des productions codifiées étaient légèrement adaptées en modifiant notamment les noms de personnages pour les remplacer par les cibles /biːba/, /baːbu/ et /buːbɪ/ afin d'obtenir un nombre équilibré d'occurrences de chaque catégorie de voyelle en position accentuée (associée à un allongement vocalique) ou non. Dans les analyses réalisées, les voyelles accentuées et non-accentuées ont été fusionnées notée /a/, /i/ ou /u/, les timbres /iː/ et /ɪ/ étant regroupés pour former une seule catégorie.

Suite à une correction manuelle de l'alignement des voyelles cibles et à un filtrage des voyelles inexploitables en raison d'une superposition avec les vocalisations de l'enfant, les valeurs des deux premiers formants extraites au milieu de chaque voyelle avec Praat ont été filtrées selon la méthode de Gendrot & Adda-Decker (2005), les valeurs du crible étant adaptées aux voyelles de l'allemand en prenant comme référence les fréquences mesurées par Pätzold & Simpson (1997). En raison du nombre élevé d'erreurs de détection dans les productions de l'une des quinze mères, cette locutrice a été exclue de l'analyse. La fréquence fondamentale a également été extraite avec Praat au milieu de chaque voyelle selon une méthode inspirée de De Looze (2010) et présentée en section 7.2.2, complétée par une inspection visuelle afin d'éliminer les erreurs résiduelles de détection correspondant à des sauts d'octave. Suite à ces mesures et au filtrage effectué, 14 519 voyelles au total ont été incluses dans l'analyse.

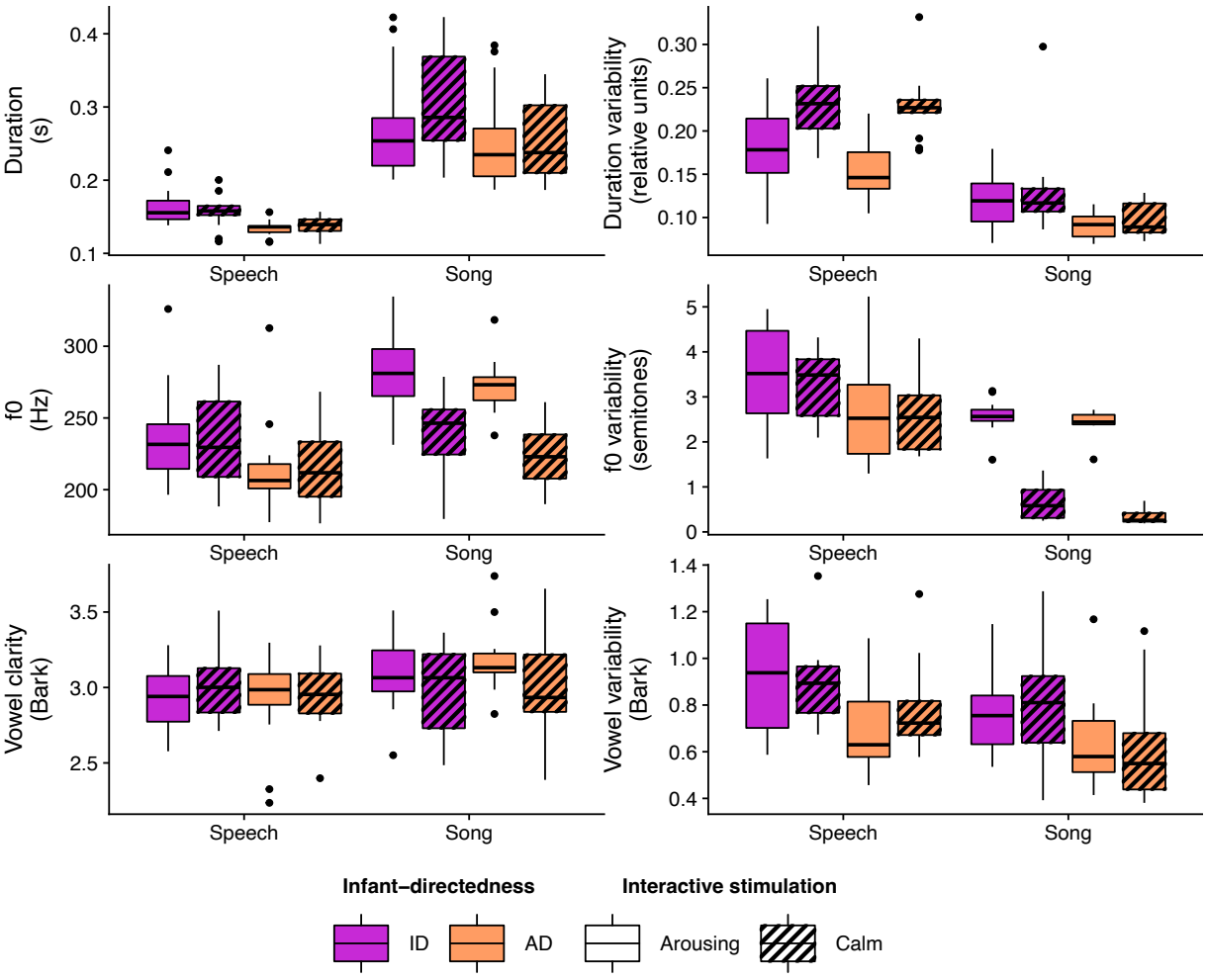


Figure 24 : Distribution comparée entre conditions de production des six mesures acoustiques retenues : durée moyenne par répétition et variabilité intra-répétition, f0 moyenne en Hertz et variabilité, mesure de centralisation DistCentroid (ici étiquetée 'Vowel clarity') et mesure de variabilité intra-catégorie VDispersion (ici étiquetée 'Vowel variability'). Les valeurs de ces mesures sont comparées entre parole et chant, entre productions dirigées vers l'enfant (ID) ou vers l'adulte (AD), et entre productions destinées à faire jouer l'enfant ou à l'amuser (Arousing) ou à le calmer pour l'endormir (Calm). D'après Falk & Audibert (2021, [ACL3]).

Les mesures de durée, de fréquence fondamentale et de fréquence des deux premiers formants ont été regroupées par locutrice et par type de production, au nombre de huit pour chaque locutrice entre considérant les quatre interactions codifiées et la présence ou absence de l'enfant. Dans chacune de ces sous-catégories, la déviation exprimée comme un rapport à la durée moyenne ainsi que la déviation en demi-tons à la fréquence fondamentale moyenne a été calculée pour chaque exemplaire de voyelle. De plus, les métriques DistCentroid (étiquetée 'Vowel clarity' dans l'étude publiée en 2021) et VDispersion (étiquetée 'Vowel variability') présentées en section 7.2.3.2 ont été calculées après conversion en Bark des fréquences des formants. La Figure 24 présente la distribution de ces six mesures, comparée entre parole et chant, entre productions adressées à l'enfant ou à l'adulte et entre but communicatif stimulant ou apaisant.

Afin d'estimer la taille relative de l'effet du type de production (chantée ou parlée), de la présence ou non de l'enfant et du but communicatif de la production codifiée, ces mesures ont été converties en z-scores au moyen d'une transformation centrée-réduite, puis modélisées par une régression bayésienne appliquée à chacune des six mesures, suite à une première analyse via un modèle de régression linéaire mixte ayant montré ses limites en ne permettant pas une prise en compte conjointe de l'ensemble des facteurs.

La comparaison entre facteurs de variation et entre variables de la magnitude des effets a montré que la distinction entre productions stimulantes et apaisantes avait un fort effet sur la fréquence fondamentale et sa variabilité, le niveau de fréquence fondamentale dépendant plus de cette dimension que du type de production et de la présence ou non de l'enfant. Cette observation est consistante avec les résultats de Tsang & Conrad (2010) dont les travaux ont montré que les enfants de six à sept mois présentent une préférence pour une fréquence fondamentale élevée dans une situation de jeu, et au contraire basse dans les berceuses. Un léger effet du but communicatif sur la durée, plus importante dans les productions stimulantes, a par ailleurs été observé. En revanche, aucun effet du but communicatif n'a été observé sur la réalisation spectrale des voyelles.

Cette analyse a montré un effet du type de production chantée ou parlée sur l'ensemble des variables à l'exception de la centralisation mesurée par la métrique DistCentroid, cet effet étant le plus massif sur la durée et sur la variabilité de la fréquence fondamentale. Si l'effet relevé sur la variabilité spectrale intra-catégorie mesurée par VDispersion était consistant avec les observations de Bradley (2018), avec une variabilité réduite dans le chant, la magnitude de l'effet était sensiblement plus faible que celle observée sur les autres mesures.

Enfin, bien qu'un effet de la présence de l'enfant ait été observé sur l'ensemble des mesures à l'exception de la centralisation de l'espace vocalique, avec des valeurs plus élevées en présence de l'enfant, la magnitude de cet effet était modérée sur les mesures liées à la durée et la fréquence fondamentale. Outre le caractère codifié des productions étudiées ici, l'absence d'effet sur la centralisation, contrairement à celui observé dans d'autres études (voir par exemple Kalashnikova & Burnham (2018)), pourrait s'expliquer par le jeune âge des enfants inclus dans l'étude. En effet, l'âge critique de l'enfant auquel cet effet est supposé maximal en lien avec l'acquisition lexicale est plutôt considéré comme proche d'un an (voir par exemple Bernstein Ratner (1984)). On peut toutefois noter qu'en accord avec les observations de A. Martin et al. (2015), la variabilité intra-catégorie de la réalisation des voyelles était plus importante dans les productions adressées à l'enfant, avec un effet de la présence de l'enfant plus important que celui des autres facteurs. Le rôle de cette variabilité intra-catégorie dans la

parole adressée à l'enfant reste débattu, certains auteurs tels que McMurray et al. (2013) la considérant comme destinée à renforcer l'acquisition du système phonologique de l'enfant tandis que d'autres (voir par exemple Kalashnikova et al. (2017)) estiment que cette variabilité découlerait de mécanismes de production employés pour servir d'autres buts communicatifs qu'un accroissement de l'intelligibilité.

4.3.3 Variabilité entre locutrices

Dans l'analyse bayésienne effectuée, la variation entre les 14 locutrices incluses dans l'étude a été considérée comme un facteur aléatoire afin de faire émerger des patrons de variation supposés généraux. Toutefois ces locutrices pourraient avoir eu recours à des stratégies individuelles variables pour distinguer acoustiquement les productions adressées ou non à leur enfant, le chant vs. la parole et les situations stimulantes vs. apaisantes. Afin de mieux identifier ces éventuels patrons individuels de variation, je propose une réanalyse séparée de ces données afin de comparer les tailles d'effets associées aux différents facteurs pour chacune des locutrices. Cette nouvelle analyse, dont les résultats sont récapitulés dans la Figure 25, s'appuie sur un modèle de régression bayésienne distinct pour chaque mesure et chaque locutrice. De même que dans la première analyse, les valeurs des mesures sont préalablement standardisées via une conversion en z-scores pour permettre une comparaison directe des tailles d'effets entre mesures, locutrices et facteurs. Contrairement à la première analyse dans laquelle le choix avait été fait de présenter les tailles d'effet en valeur absolue pour mettre l'accent sur leur magnitude, le signe correspondant à la direction de l'effet est ici conservé pour permettre une interprétation plus directe des différences, d'où une correspondance visuelle avec la figure publiée qui n'est que partielle.

Comme illustré par la Figure 25, dans la majorité des cas la direction de l'effet (c'est-à-dire le signe positif ou négatif de l'estimation médiane représentée par un point, en ne considérant pas les quelques cas considérés comme une absence d'effet dans lesquels l'intervalle crédible à 95% inclut la valeur 0) est la même pour l'ensemble des locutrices.

On peut néanmoins noter des contre-exemples, notamment pour l'effet du but communicatif sur la durée. Ainsi, pour trois des 14 locutrices, les durées sont allongées lorsque le but communicatif est orienté vers la stimulation ou le jeu plutôt que vers l'apaisement de l'enfant, tandis qu'aucun effet n'est observé pour deux locutrices et que les neuf autres allongent au contraire leurs voyelles en condition d'apaisement. On peut noter que l'une de ces trois locutrices est par ailleurs celle qui augmente le plus fortement son registre de fréquence fondamentale dans les productions stimulantes, toutefois cette stratégie n'est pas généralisable aux autres locutrices.

Bien que minoritaire, l'un de ces contre-exemples est particulièrement contre-intuitif, avec une locutrice dont la fréquence fondamentale est plus élevée dans les productions adressées à l'adulte que dans celles adressées à son enfant. La direction inattendue de cet effet s'explique en partie par la fréquence fondamentale plus basse de deux demi-tons adoptée par la locutrice pour chanter la berceuse à son enfant, ce qui est cohérent avec la représentation majoritairement associée aux berceuses (Trainor et al., 1997). Cependant, de façon plus surprenante, la fréquence fondamentale de cette locutrice est également légèrement plus élevée dans les trois autres productions codifiées lorsque celles-ci sont adressées à l'adulte que lorsqu'elles sont adressées à l'enfant.

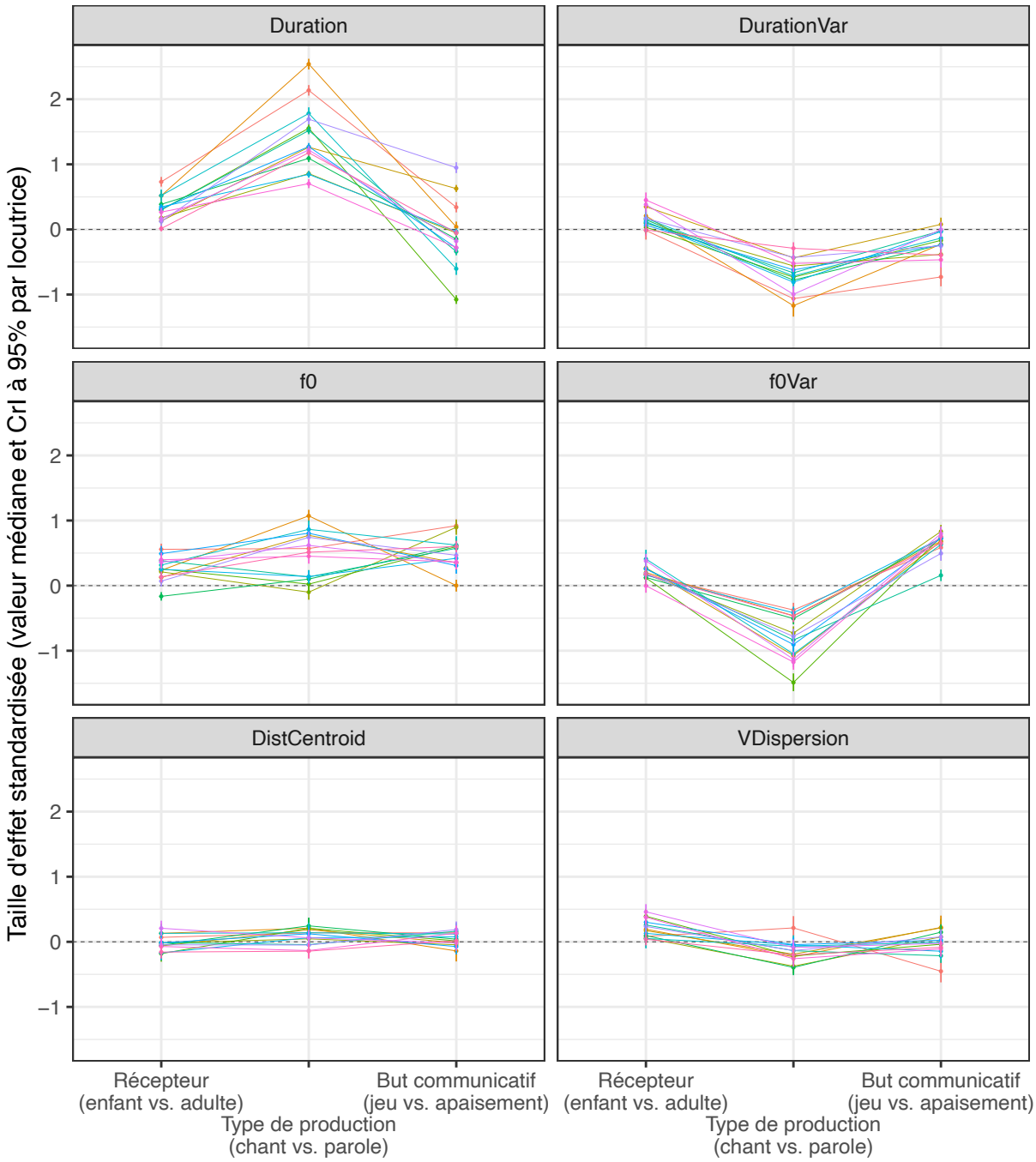


Figure 25 : Tailles d'effet standardisées de la production adressée à son enfant plutôt qu'à un adulte, du chant plutôt que de la parole, et d'une production destinée au jeu plutôt qu'à calmer l'enfant, comparées entre les 14 locutrices. Chaque couleur correspond à l'une des locutrices. Les points représentent la valeur médiane de la taille d'effet estimée par le modèle bayésien appliqué à la locutrice correspondante, les segments verticaux autour des points représentent les limites de l'intervalle crédible à 95% de l'effet. Les segments colorés permettent de visualiser la pente de la différence entre effets. La valeur 0 (ligne grise horizontale pointillée) correspond à une absence d'effet, tandis que les valeurs positives indiquent des valeurs plus élevées de la variable pour la première des deux modalités comparées (par exemple pour l'effet du « Récepteur », des valeurs plus élevées pour des productions adressées à l'enfant qu'à l'adulte), et inversement.

De telles divergences individuelles dans la direction de l'effet peuvent également être notées pour la mesure VDispersion de variabilité intra-catégorie, avec toutefois des effets de magnitude modérée. Tandis que neuf locutrices montrent une variabilité de réalisation spectrale des voyelles plus faible dans le chant que dans la parole et qu'aucun effet n'est observé pour quatre d'entre elles, une locutrice montre le comportement inverse avec une variabilité accrue dans le chant, en raison à la fois d'une variabilité moindre pour le récit de l'histoire rimée et d'une variabilité supérieure à celle des autres locutrices pour la berceuse. Par ailleurs si la majorité des locutrices ont montré une absence d'effet du but communicatif, ce qui correspond aux observations sur le modèle toutes locutrices confondues, trois augmentent la variabilité intra-catégorie de la réalisation des voyelles dans les productions stimulantes, tandis que trois autres le font dans les productions apaisantes.

Des observations similaires peuvent être faites pour la mesure de centralisation DistCentroid sur laquelle aucun effet global n'avait été relevé dans l'étude initiale, là aussi avec des effets non-nuls quoique de faible magnitude pour une minorité de locutrices et une direction d'effet variable entre locutrices. Ainsi, trois locutrices centralisent plus leurs voyelles dans les productions adressées à l'enfant qu'à l'adulte, tandis qu'une seule montre le schéma inverse plus conforme aux résultats observés dans certaines études telles que celle de Kalashnikova & Burnham (2018). Quatre locutrices produisent des voyelles plus périphériques dans le chant que dans la parole, tandis que deux locutrices produisent à l'inverse des voyelles plus centralisées dans le chant. Enfin, deux locutrices produisent des voyelles plus périphériques dans les productions stimulantes que dans celles apaisantes. Cette variabilité individuelle des effets observés sur les mesures relatives à l'espace vocalique suggère que l'absence d'effet du but communicatif sur la variabilité intra-catégorie ainsi que l'absence d'effet des différents facteurs sur la centralisation dans l'analyse initiale pourraient être dues en réalité à des divergences de stratégies individuelles plutôt qu'à une constance de la réalisation des voyelles entre productions codifiées.

Un autre aspect de cette variation individuelle concerne les différences observées entre tailles d'effet des différents facteurs, qui se manifestent dans la Figure 25 sous forme de croisements entre les segments qui relient les tailles d'effet des différents facteurs pour une même locutrice. Ces différences individuelles concernent majoritairement la magnitude relative de la taille d'effet des différents facteurs, avec par exemple pour les effets sur la durée une variabilité interindividuelle moins importante pour l'effet de la présence de l'enfant que pour les effets du type de production et du but communicatif. Dans une moindre mesure, on observe également des différences individuelles sur la hiérarchie de ces tailles d'effet. Ainsi, pour deux locutrices l'effet du but communicatif sur la durée est supérieur à celui de la présence de l'enfant, et pour une autre l'effet du but communicatif sur les variations de durée est supérieur à celui du type de production.

Comme l'illustre la Figure 25, ces différences individuelles de hiérarchie entre tailles d'effets sont plus marquées pour les effets sur la fréquence fondamentale, avec notamment la moitié des locutrices pour lesquelles l'effet du but communicatif est supérieur à celui du type de production, l'autre moitié présentant le schéma inverse. Ces observations confortent l'hypothèse de stratégies individuelles variables dans l'exploitation des différents indices acoustiques permettant de transmettre à l'enfant différents buts communicatifs véhiculés par des productions parlées ou chantées.

4.4 Variation en fonction de la position prosodique

Publications associées :

[ACTI41] Georgeton, L., & **Audibert, N.** (2013). Is protrusion of French rounded vowels affected by prosodic positions? *Proceedings of Interspeech 2013*, Lyon, France. pp. 3547-3551.

[ACTI47] Georgeton, L., **Audibert, N.,** & Fougeron, C. (2011). Rounding and height contrasts at the beginning of different prosodic constituents in French. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong-Kong, Chine, pp. 739-742.

En collaboration avec Laurianne Georgeton et Cécile Fougeron, j'ai contribué à l'étude de l'influence de la position prosodique sur la réalisation des contrastes vocaliques en français réalisée dans le cadre de sa thèse. Ma contribution à ces travaux a été principalement méthodologique afin d'extraire des mesures d'articulation labiale obtenus à partir d'un dispositif de capture optique de mouvement et de les évaluer (voir section 7.1.4 pour la présentation de ces mesures). Dans une moindre mesure, j'ai également été associé à travers ma contribution méthodologique à quelques études plus directement liées à la thématique scientifique de la thèse de Laurianne Georgeton (2014), sur l'influence de la position prosodique initiale sur l'articulation des voyelles et plus spécifiquement le geste d'arrondissement.

Les études auxquelles j'ai participé ont été menées à partir d'un corpus contrôlé composé de phrases conçues pour faire apparaître différentes voyelles en position initiale de groupe intonatif, en position initiale de groupe accentuel ou en position initiale de mot, tout en maintenant constants les contextes segmentaux gauche et droit. Ce corpus a ainsi permis d'évaluer l'influence du type de frontière prosodique sur la réalisation du contraste d'arrondissement via la comparaison entre paires de voyelles de même aperture arrondie ou non-arrondie ou entre voyelles non-arrondies de différents niveaux d'aperture. L'effet de la position prosodique a été évalué au niveau acoustique (Georgeton et al., 2011 [ACTI47]), et via l'étude articulatoire du geste d'arrondissement en combinant plan sagittal et coronal (Georgeton & Audibert, 2013 [ACTI41]) à partir de données vidéo et de capture de mouvements. Je reviens ici brièvement sur l'étude acoustique, et sur la seconde étude articulatoire.

L'étude acoustique, menée à partir de 1 025 exemplaires des voyelles /i, y, e, ø, a/ produites par quatre locutrices, s'est appuyée pour le contraste d'arrondissement sur des mesures de fréquence du deuxième et du troisième formant, abaissés par l'arrondissement, ainsi que de l'écart entre deuxième et troisième formant F3-F2, supposé réduit par l'arrondissement dans le cas de la paire de voyelles /i, y/ (voir par exemple Vaissière (2008)). Ces mesures sont illustrées par la Figure 26. La comparaison entre voyelles et conditions a montré que le contraste entre les voyelles non-arrondies /i/ et /e/ et les voyelles arrondies correspondantes /y/ et /ø/ était renforcé, de façon d'autant plus importante que la frontière prosodique correspondante était d'un niveau hiérarchique élevé. Tandis que les corrélats acoustiques de l'absence d'arrondissement ont été renforcés de façon consistante entre locutrices, les corrélats acoustiques de l'arrondissement ne l'ont été que par deux des quatre locutrices. Parmi les deux interprétations théoriques concurrentes, le renforcement observé en position initiale de groupe intonatif via les effets sur F2, F3 et F3-F2 a été jugé comme plus en accord avec l'hypothèse d'hyperarticulation locale (K. De Jong, 1995) qu'avec l'hypothèse d'expansion

de la sonorité (Beckman et al., 1992) qui postule une augmentation de la sonorité des segments en position prosodique forte. Les contrastes d'aperture ont quant à eux été évalués via la comparaison des mesures de fréquence du premier formant, montrant une augmentation du F1 de /a/ en position initiale de groupe intonatif compatible avec les deux interprétations théoriques, tandis que l'effet de la position prosodique sur les autres voyelles non-arrondies était beaucoup moins systématique et très variable entre locutrices.

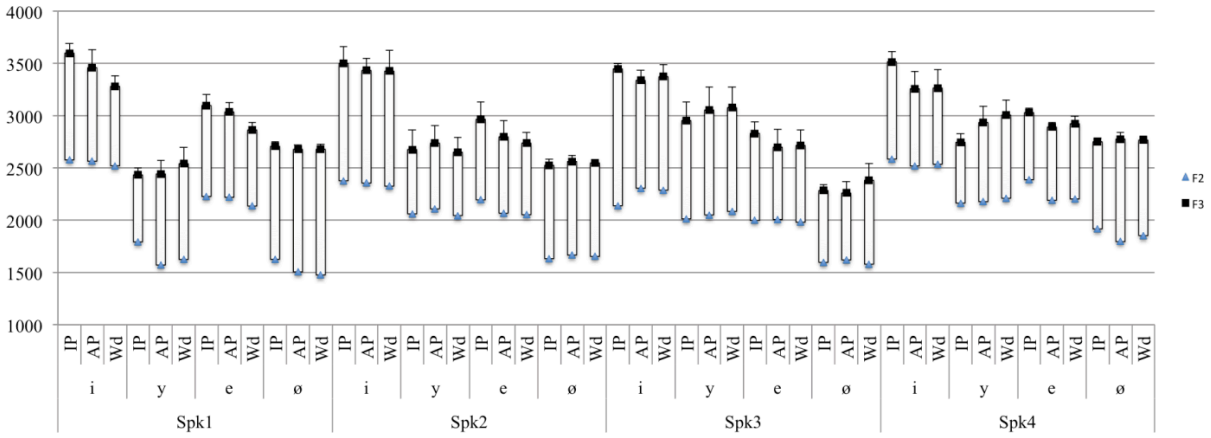


Figure 26 : Positions de F2 et F3 relevées sur les voyelles /i, y, e, ø/ en position initiale de groupe intonatif (IP), de groupe accentuel (AP) et de mot (Wd), pour chacune des quatre locutrices. D'après Georgeton et al. (2011, [ACTI47]).

L'étude articulatoire, menée à partir de 1 368 exemplaires des voyelles /i, e, ɛ, a, y, ø, œ, u, o, ɔ/ produites par trois locutrices, s'est appuyé sur l'analyse des données vidéos prises de face pour les mesures d'aire aux lèvres, et sur les données obtenues par capture optique de mouvement pour les mesures de protrusion. Ces mesures ont été combinées pour l'analyse en considérant comme mesure articulatoire du degré d'arrondissement le rapport entre aire aux lèvres et protrusion (Stevens & House, 1955), ainsi que le rapport inverse introduit par Fant (1971) pour permettre une interprétation plus directe en lien avec la modélisation acoustique des sons de la parole. Cette étude fait suite à une première étude essentiellement méthodologique (Georgeton & Audibert, 2012 [ACTI43], cf. section 7.1.4.3) menée exclusivement dans le plan coronal, qui a également montré une aire aux lèvres et des distances horizontales et verticales supérieures en position initiale de groupe intonatif non seulement pour la voyelle arrondie /y/ mais également pour /i/ et /a/. Après une vérification de la capacité de la mesure de protrusion et des deux mesures combinant aire aux lèvres et protrusion à rendre compte de l'opposition d'arrondissement, les comparaisons se sont concentrées sur les différences entre positions prosodiques pour les voyelles arrondies. Ces comparaisons ont montré un effet du type de frontière prosodique sur le rapport entre protrusion et aire aux lèvres, avec des valeurs réduites interprétables comme un plus grand arrondissement en position prosodique forte, pour deux des trois locutrices. En revanche, contrairement aux résultats de l'études acoustique et à ceux de la littérature, les différences observées ne reflètent pas la hiérarchie attendue entre niveaux prosodiques avec une absence de distinction entre position initiale de groupe intonatif et de groupe accentuel. De plus la protrusion considérée seule n'est pas affectée par la position prosodique. Le renforcement du contraste d'arrondissement mesuré dans l'étude acoustique semble donc attribuable principalement aux variations de l'aire aux lèvres observées par ailleurs sur les mêmes données par Georgeton & Fougeron (2014).

4.5 Coarticulation voyelle-à-voyelle

Publications associées :

[ACTI32] Turco, G., Fougeron, C., & **Audibert, N.** (2016). The Effects of Prosody on French V-to-V Coarticulation: A Corpus-Based Study. *Proceedings of Interspeech 2016*, San Francisco, Etats-Unis, pp. 998-1001.

[ACTI33] Turco, G., Fougeron, C., & **Audibert, N.** (2016). Que nous apprennent les gros corpus sur l'harmonie vocalique en français ? *Actes des 31èmes Journées d'Études sur la Parole*, Paris, pp. 571-579.

4.5.1 L'harmonie vocalique en français

En collaboration avec Giuseppina Turco et Cécile Fougeron, nous avons procédé à une étude de corpus de la coarticulation voyelle-à-voyelle en français, plus couramment appelée harmonie vocalique, à partir des données des corpus ESTER (Galliano et al., 2009) et NCCFr (Torreira et al., 2010). Ce phénomène est généralement décrit en français comme un processus par lequel l'aperture des voyelles moyennes en syllabe non-finale peut, de façon optionnelle, être affectée par l'aperture de la syllabe finale suivante (voir par exemple Tranel (1987)). Toutefois, les conditions dans lesquelles l'harmonie vocalique peut ou non survenir ne sont pas clairement définies dans la littérature, comme relevé par Fagyal et al. (2003) qui ont été les premiers à proposer une étude acoustique de la réalisation de l'harmonie vocalique à partir d'un ensemble de paires de mots insérés dans une phrase porteuse. L'objectif de notre travail, dans lequel je suis intervenu principalement au niveau méthodologique afin de mettre en œuvre le codage des données, l'extraction des mesures acoustiques et leur analyse statistique, a été de mieux caractériser l'apparition de ce phénomène en français à partir de l'analyse acoustique de grands corpus de parole continue qui comprennent une variété lexicale et un nombre de locuteurs importants. Nous avons ainsi cherché à déterminer quelles sont les voyelles qui subissent et/ou déclenchent le plus l'harmonie vocalique, ainsi que le poids des différents facteurs mentionnés dans la littérature comme susceptibles d'influencer l'harmonie vocalique. Suite aux premiers travaux sur l'anglais (Cho, 2004) et sur l'italien (Gili Fivela et al., 2011) qui ont conclu à une résistance à la coarticulation voyelle-à-voyelle à travers une frontière prosodique, le degré de résistance étant dépendant du niveau de cette frontière dans la hiérarchie prosodique, nous avons également cherché à déterminer dans quelle mesure l'harmonie vocalique en français pouvait être influencée par le renforcement prosodique en position initiale de constituant.

4.5.2 Principaux résultats

Une première étude s'est concentrée sur les contextes potentiels d'harmonie vocalique à l'intérieur d'un même mot, en considérant tous les mots de deux syllabes ou plus susceptibles de faire l'objet d'une harmonie vocalique et présentant un nombre suffisant d'occurrences. L'ensemble de mots retenus a été ceux ayant la voyelle V_2 en syllabe finale de mot et incluse dans l'ensemble /i, e, ɛ, a, y, u, o, ɑ̃/, et la voyelle V_1 de la syllabe précédente dans l'ensemble /e, ɛ, o, ɔ/. Les voyelles arrondies /ø, œ/ susceptibles d'être confondues avec la réalisation du schwa en français ont été exclues de l'analyse. Les cas dans lesquels ces deux voyelles sont séparées par un schwa non-réalisé ont été inclus, conformément aux propositions de Dell (1973) qui considère que l'harmonie vocalique peut opérer après la délétion du schwa, mais

ont été codés de façon distincte afin de pouvoir les prendre en compte dans l'analyse. Un total de 33 000 mots a ainsi été inclus dans l'analyse. Afin de tenir compte de la variabilité de la réalisation des voyelles moyennes en français et tout particulièrement de celle des voyelles /e, ɛ/, l'aperture de ces voyelles en syllabe finale V_2 a été catégorisée en fonction des étiquettes attribuées par le système d'alignement avec variantes de prononciation (Adda-Decker & Lamel, 2000), combinées avec des critères orthographiques et avec la prononciation canonique extraite de la base Lexique (New et al., 2007). Les voyelles en syllabe finale V2 ont ainsi été catégorisées en deux grandes classes d'aperture étiquetées en tant que voyelles hautes ou basses.

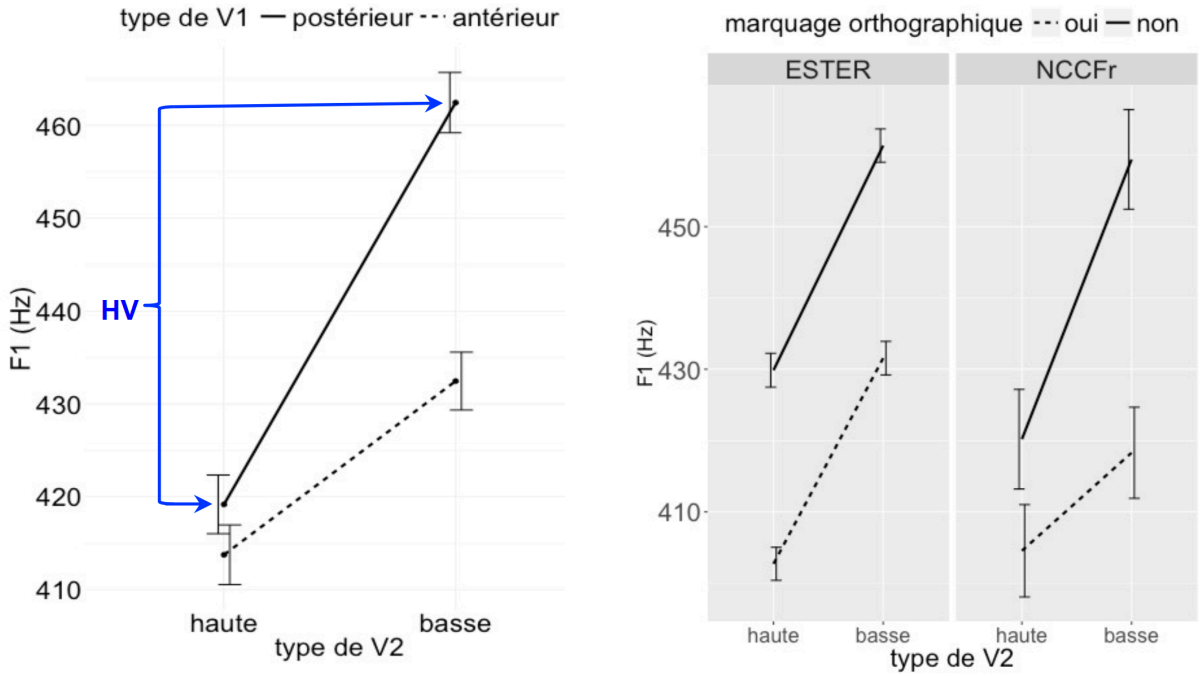


Figure 27 : Fréquence du premier formant de la première voyelle V_1 en fonction de l'aperture de de la seconde voyelle V_2, comparée entre V_1 antérieures et postérieures (gauche), ou entre V_1 marquées ou non orthographiquement en faveur d'une prononciation mi-fermée et entre corpus représentant les deux styles de parole comparés (droite). Dans les deux parties de la figure les valeurs de F1 représentées sont les valeurs moyennes prédites par le modèle linéaire mixte et les barres d'erreur représentent l'erreur-type. L'accolade bleue illustre l'écart entre valeurs de F1 induit par la différence entre une seconde voyelle V_2 haute ou basse, interprétable comme le degré d'harmonie vocalique. D'après Turco et al. (2016, [ACTI33]).

Comme illustré par la Figure 27, l'évaluation statistique de l'effet sur la fréquence du premier formant de la voyelle V_1 de la nature de V_1 et de V_2, du style de parole et de la présence ou non d'une graphie favorisant une prononciation mi-fermée de la première voyelle V_1 a montré une harmonie vocalique plus importante sur les voyelles postérieures que sur les antérieures (partie gauche de la figure). Comme illustré par la partie droite de la Figure 27, si le marquage orthographique favorisant la prononciation d'une voyelle mi-fermée plutôt que mi-ouverte (graphie é pour [e] et au/eau pour [o]) se traduit bien comme attendu par des valeurs de F1 de la voyelle V_1 plus basses, l'absence de marquage orthographique ne se traduit par une harmonie vocalique plus importante que dans le corpus de parole spontanée NCCFr, conformément aux descriptions antérieures (voir par exemple Tranel (1987)) qui font état d'une harmonie vocalique plus importante dans les styles moins formels. En revanche pour les

mots marqués orthographiquement, l'harmonie vocalique plus faible en parole spontanée que journalistique va à l'encontre du postulat de Taft & Hambly (1985) d'une modulation plus forte des représentations phonologiques par le marquage orthographique dans la parole plus formelle.

Par ailleurs des questions plus spécifiques ont également été évaluées à partir de ces données. La comparaison entre voyelles V_1 et V_2 séparées par une consonne labiale ou linguale a confirmé que la coarticulation voyelle-à-voyelle était favorisée par la présence d'une consonne labiale moins résistante à la coarticulation qu'une consonne linguale comme cela avait déjà été observé pour d'autres langues comme par exemple l'anglais américain (Fowler & Brancazio, 2000). La comparaison des cas dans lesquels la forme canonique du mot inclut un schwa non réalisé dans nos données aux cas dans lesquelles cette forme canonique n'inclut qu'une séquence de deux consonnes ou plus entre V_1 et V_2 a suggéré que l'harmonie vocalique n'était pas bloquée par la présence d'un schwa sous-jacent conformément aux postulats de Dell (1973), mais que son amplitude serait réduite par le schwa non-réalisé.

Dans une seconde étude à partir des mêmes données, nous nous sommes concentrés sur l'évaluation de l'effet de la position prosodique sur l'harmonie vocalique, en comparant les cas dans lesquels la première voyelle V_1 est en position initiale absolue de groupe intonatif à ceux dans lesquels cette voyelle V_1 est à l'intérieur du mot. Par ailleurs la durée de la première voyelle a été prise en compte, en distinguant les voyelles de 50 ms ou moins considérées comme courtes de celles de plus de 50 ms. Comme illustré par la Figure 28, un fort effet de la position prosodique a été relevé, avec une résistance à la coarticulation plus importante de la première voyelle V_1 lorsque celle-ci est en position initiale absolue que lorsqu'elle est en position médiale de mot. En revanche, cette analyse n'a montré aucun effet de la durée sur l'amplitude de l'harmonie vocalique.

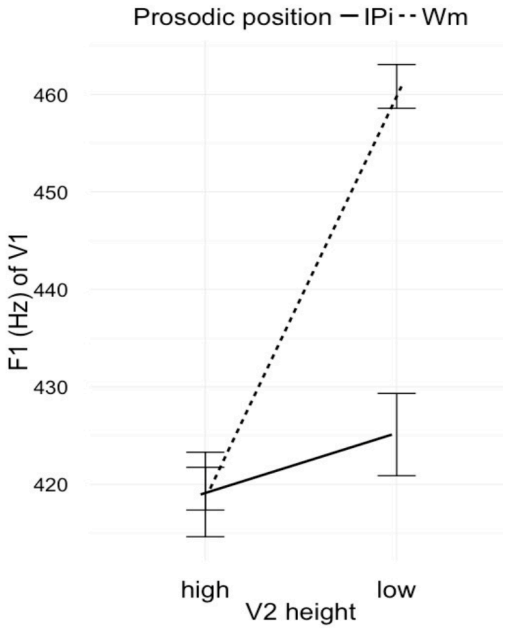


Figure 28 : Fréquence du premier formant de la première voyelle V_1 en fonction de l'aperture de de la seconde voyelle V_2, comparées entre V_1 en position initiale absolue de groupe intonatif (IPi) ou en position médiale de mot (Wm). Les valeurs de F1 représentées sont les valeurs moyennes prédites par le modèle linéaire mixte et les barres d'erreur représentent l'erreur-type. D'après Turco et al. (2016, [ACTI32]).

4.5.3 Variation entre locuteurs

En complément de ces deux études dans lesquelles le locuteur a été considéré comme un facteur de variation aléatoire dans la modélisation statistique afin de dégager des patrons de variation interprétables comme communs à un ensemble de locuteurs, je propose une nouvelle analyse de ces données en me concentrant sur la variation entre locuteurs. Cette nouvelle analyse nécessite de se limiter aux locuteurs du corpus ESTER les plus représentés dans les différents sous-ensembles à comparer. En effet, si les données prises en compte dans les études que nous avons publié sur l'harmonie vocalique en français incluent un total de 671 locuteurs du corpus ESTER ayant produit des mots susceptibles de donner lieu à une harmonie vocalique, 89% de ces locuteurs ont produit moins de 50 mots exploitables pour cette analyse, avec une répartition inégale entre catégories comparées. Si ce constat ne remet pas en question la validité des analyses initiales du fait de la capacité des modèles linéaires mixtes à modéliser efficacement des jeux de données déséquilibrés (comme souligné par exemple par Gelman & Hill (2006)), elle limite les possibilités en termes de comparaisons entre locuteurs pour lesquelles il est indispensable que chaque locuteur soit suffisamment représenté.

Du fait de la sous-représentation dans les données du corpus ESTER des mots avec la première voyelle V_1 marquée orthographiquement et/ou en position prosodique initiale, je me contente donc d'effectuer cette comparaison entre locuteurs de l'amplitude de l'harmonie vocalique en comparant voyelles V_1 antérieures et postérieures. Les données de chaque locuteur étant incluses dans un corpus distinct, je fais le choix de distinguer les locuteurs des deux corpus afin de pouvoir discuter la variation inter-locuteurs à la lumière des différences entre styles de parole. Afin d'assurer la représentativité de chaque locuteur inclus et d'équilibrer le nombre de locuteurs pris en compte dans chacun des deux corpus, j'ai retenu un seuil de 25 mots minimum produits dans chacune des quatre conditions (V_2 haute ou basse, V_1 antérieure ou postérieure) pour inclure les locuteurs. Cette comparaison porte donc sur 23 locuteurs du corpus NCCFr (11 femmes et 12 hommes), et 26 locuteurs du corpus ESTER (5 femmes et 21 hommes). Pour la majorité de ces 49 locuteurs, les voyelles V_1 /e, ɛ/ antérieures sont plus représentées que les voyelles V_1 /o, ɔ/ postérieures, en accord avec les fréquences d'occurrence relevées par Adda-Decker (2006) sur différents styles de parole. Dans chacune des quatre conditions, la tendance centrale de la fréquence du premier formant est estimée par la valeur médiane de F1.

La Figure 29 illustre l'harmonie vocalique observée pour chaque locuteur des deux corpus, les hommes étant représenté par un tracé différent de celui de des femmes, en séparant pour chaque locuteur le cas des voyelles V_1 antérieures et postérieures. Comme attendu, le niveau général de F1 est plus élevé pour les femmes que pour les hommes dans le corpus NCCFr en raison des différences inter-sexe de fréquences formantiques bien documentées dans la littérature (voir par exemple Fant (1975) parmi de nombreux autres auteurs), la sous-représentation des femmes parmi les locuteurs retenus du corpus ESTER ne permettant pas de comparaisons entre hommes et femmes dans ce corpus. On peut observer pour la majorité des locuteurs une pente plus importante pour les voyelles V_1 postérieures que pour les antérieures, conformément aux résultats obtenus auparavant tous locuteurs confondus. Toutefois, dans les quatre sous-ensembles représentés par des panneaux distincts de la Figure 29, on peut également noter une importante variabilité inter-individuelle dans les pentes entre voyelle V_2 haute et basse qui reflète une amplitude de l'harmonie vocalique variable entre locuteurs. Si cette pente reste positive pour l'ensemble des locuteurs dans le cas des voyelles

V_1 postérieures, avec des fréquences de F1 qui évoluent dans la même direction que celle correspondant à l'aperture de la voyelle V_2, de façon plus étonnante on observe également dans le cas des voyelles V_1 antérieures et tout particulièrement pour le corpus de parole conversationnelle NCCFr des pentes négatives, c'est-à-dire des locuteurs pour lesquels l'effet de la voyelle en syllabe finale V_2 va en sens inverse de celui attendu dans le cas d'une harmonie vocalique, avec un F1 de V_1 plus élevé pour les voyelles V_2 hautes et inversement.

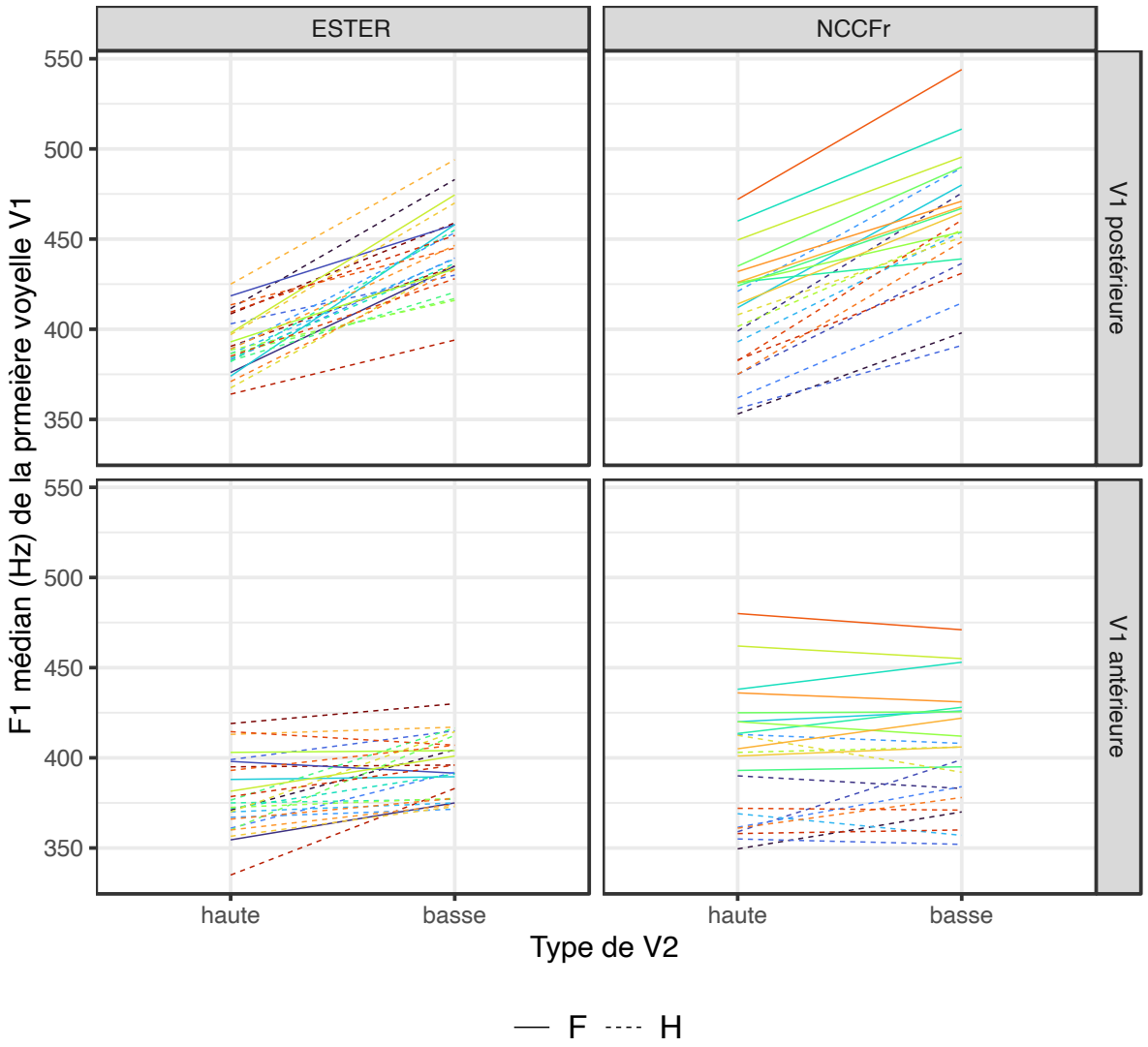


Figure 29 : Harmonie vocalique représentée comme la différence de fréquence du premier formant de la première voyelle V_1 en fonction de l'aperture de de la seconde voyelle V_2, pour les 26 locuteurs du corpus ESTER et les 23 locuteurs du corpus NCCFr inclus dans la comparaison, comparée entre V_1 postérieure (partie haute de la figure) ou antérieure (partie basse). Au sein de chacun des deux corpus, chaque locuteur est représenté par une ligne de couleur différente. Les femmes sont représentées par les lignes en trait plein, et les hommes par les lignes pointillées.

Une autre façon possible d'étudier les différences individuelles d'harmonie vocalique est de calculer pour chaque locuteur et chaque classe d'antériorité/postériorité de la voyelle V_1 l'amplitude de l'harmonie vocalique comme la différence entre la fréquence médiane du premier formant relevée sur la voyelle V_1 lorsque la voyelle V_2 est une voyelle basse et celle relevée lorsque V_2 est une voyelle haute. Si elle a pour conséquence de perdre le lien avec les mesures brutes de F1 sur la voyelle V_1, cette projection permet aussi d'interpréter plus

directement les différences individuelles d'harmonie vocalique, et éventuellement de les mettre en relation avec d'autres mesures comme la durée segmentale. La durée médiane de la voyelle V_1 pour chaque locuteur n'explique que très partiellement les différences individuelles observées, avec une amplitude de l'harmonie vocalique plus importante pour les voyelles plus courtes mais des corrélations modérées, tout particulièrement sur les voyelles V_1 postérieures pour lesquelles l'harmonie vocalique est globalement plus importante (corrélation de Spearman ρ=-0,33 pour les voyelles V_1 antérieures, ρ=-0,17 pour les voyelles V_1 postérieures).

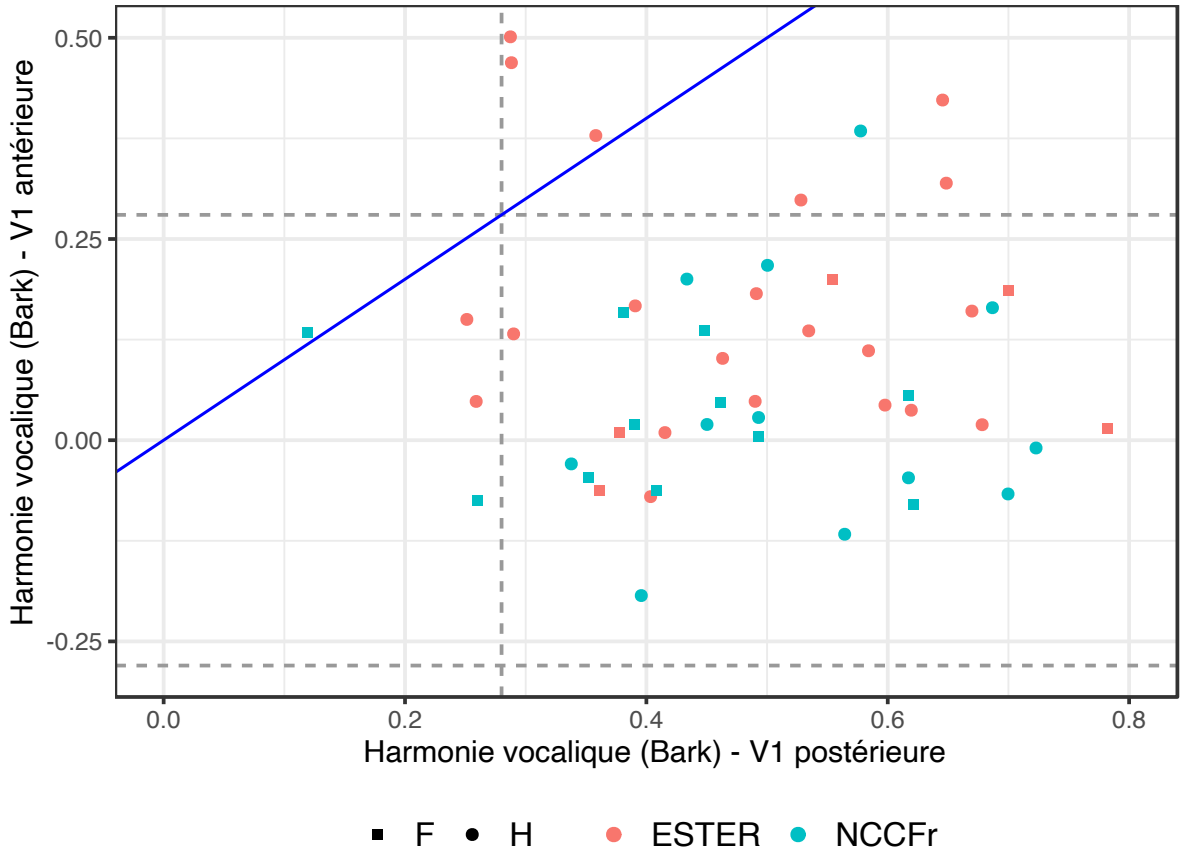


Figure 30 : Amplitude de l'harmonie vocalique estimée pour chaque locuteur et chaque classe d'antériorité/postériorité de la première voyelle V_1 comme la différence en Bark entre la valeur médiane de F1 de la voyelle V_1 associée à une V_2 basse, et celle associée à une V_2 haute. Chaque locuteur est représenté par un point dans l'espace à deux dimensions, formé par l'amplitude de l'harmonie vocalique subie par les voyelles V_1 postérieures (abscisses) et l'harmonie vocalique subie par les voyelles V_1 antérieures (ordonnées). Les locuteurs de chacun des deux corpus sont représentés par des couleurs différentes, et les hommes et femmes par des formes de point différentes. La ligne diagonale bleue représente la droite d'ordonnée à l'origine 0 et de pente 1 qui sépare les locuteurs pour lesquels l'harmonie vocalique est plus forte sur les voyelles V_1 postérieures (à droite de la diagonale) de ceux pour lesquels elle est plus forte sur les voyelles V_1 antérieures (à gauche de la diagonale). Les lignes pointillées grises horizontales et verticales correspondent au seuil de perception de 0,28 Bark (Kewley-Port & Zheng, 1999).

La Figure 30 illustre les mesures individuelles d'amplitude de l'harmonie vocalique pour une voyelle V_1 postérieure ou antérieure, représentées dans un espace à deux dimensions avec des points de couleur différente en fonction du corpus et de forme différente pour hommes et femmes. L'amplitude de l'harmonie vocalique est ici calculée après conversion en Bark

(Traunmüller, 1990) des valeurs de F1 pour mieux tenir compte des différences de niveau de F1 entre hommes et femmes observées précédemment. Comme on peut le voir, conformément aux résultats de l'analyse tous locuteurs confondus, l'harmonie vocalique est bien plus importante sur les voyelles postérieures pour la grande majorité des locuteurs, avec toutefois trois locuteurs du corpus ESTER sur 26 et une locutrice du corpus NCCFr pour lesquels elle est plus importante sur les voyelles antérieures.

Hormis la tendance modérée vers une harmonie vocalique plus importante sur les voyelles antérieures de la part des locuteurs du corpus ESTER comparativement au corpus NCCFr, on observe essentiellement une grande variabilité inter-individuelle qui ne semble pouvoir s'expliquer ni par le style de parole journalistique ou conversationnel, ni par les différences entre hommes et femmes. Par ailleurs cette visualisation confirme que les valeurs négatives d'amplitude de l'harmonie vocalique (soit ce qui pourrait correspondre à une « divergence vocalique » en lieu et place de l'harmonie vocalique, avec des valeurs de F1 qui varient dans la direction inverse à celle attendue) observées sur les voyelles V_1 antérieures concernent majoritairement les locuteurs du corpus NCCFr avec un total de dix locuteurs de ce corpus sur 23 dans ce cas, pour seulement deux locuteurs du corpus ESTER, de façon équilibrée entre hommes et femmes.

Cette observation peut amener à s'interroger sur les éventuels facteurs non contrôlés susceptibles d'influencer la catégorisation et/ou les mesures, qui constituent la principale limite des approches sur grands corpus. Dans le cadre d'une extraction automatisée à grande échelle de mesures formantiques, on ne peut bien entendu pas exclure l'hypothèse selon laquelle des erreurs de détection viendraient biaiser certaines analyses. Si les valeurs de F1 sont peu susceptibles d'être biaisées par un niveau d'énergie insuffisant comme c'est plus fréquemment le cas pour les formants supérieurs, elles peuvent plus fréquemment l'être du fait de la proximité avec des valeurs élevées de f0 (voir par exemple Kent & Vorperian (2018)). Nos données semblent toutefois peu exposées à ce type de biais, avec pour l'ensemble des locuteurs des écarts entre f0 et F1 supérieurs à 100Hz dans la grande majorité des cas. Des écarts inférieurs à 100Hz sont observés sur plus de 9% des voyelles V_1 pour une seule locutrice du corpus NCCFr (13% des voyelles de cette locutrice), pour laquelle cette proportion plus élevée n'explique pas l'amplitude de l'harmonie vocalique.

Des facteurs lexicaux pourraient par ailleurs influencer les mesures d'amplitude de l'harmonie vocalique et expliquer en partie ces observations contre-intuitives. Ainsi, bien qu'il ne soit pas le seul dans ce cas et que cela soit encore plus marqué pour d'autres locuteurs du corpus NCCFr, le locuteur pour lequel l'amplitude estimée de l'harmonie vocalique sur les voyelles antérieure est la plus fortement négative présente une proportion de 44% des voyelles V_1 antérieures suivies d'une voyelle V_2 considérée comme basse qui correspondent au mot « était » ou à l'un de ses homophones. Toutefois, ces observations sont à nuancer par l'amplitude de l'harmonie vocalique observée sur les voyelles V_1 antérieures. En effet, dans la grande majorité des cas l'écart mesuré en Bark est inférieur au seuil de perception (*Just Noticeable Difference*) de 0,28 Bark (Kewley-Port & Zheng, 1999) matérialisé par la ligne grise pointillée dans la Figure 30.

5 Voix et parole atypique : phonétique clinique et vieillissement

Résumé du chapitre 5

Dans ce chapitre je reviens sur mes travaux en phonétique clinique sur les dysarthries d'une part, et sur les dysphonies et voix de substitution d'autre part, ainsi que sur l'effet de l'âge sur la voix et la parole.

Deux études se sont concentrées sur la caractérisation par diverses métriques des distorsions de l'espace vocalique observées dans différents types de dysarthries et ont conclu à la nécessité d'une approche multidimensionnelle non limitée à l'aire de l'espace vocalique pour rendre compte de ces distorsions, variables en fonction de la pathologie considérée. L'évaluation comparative de méthodes d'estimation de l'intelligibilité de la parole dysarthrique a mis en évidence un biais de la cotation experte courante et suggéré la possibilité d'obtenir une approximation acceptable via une méthode automatique.

Dans le cadre de la thèse d'A. Pettirossi, une série d'études sur les voix de femmes modérément dysphoniques a conclu à l'attribution de traits de personnalité plus négatifs aux voix les plus altérées, à un décalage entre jugements naïfs et experts du trouble vocal avec une plus grande tolérance des naïfs à la raucité et au souffle, à une moindre capacité des dysphoniques à adapter leur voix au bruit environnemental, et à une incidence de la dysphonie des locutrices sur le traitement du contraste de voisement par des enfants de 6 à 10 ans.

Une étude centrée sur la voix d'expressions de colère et de tristesse ainsi que d'expressions neutres par des patients atteints de dysphonies sévères en raison d'une paralysie laryngée unilatérale suggère que les patients parviennent moins bien que les témoins à moduler leur fréquence fondamentale et leur qualité de voix pour produire des expressions émotionnelles, notamment de colère. Une étude acoustique et perceptive longitudinale de voix de substitution associées à diverses techniques chirurgicales de laryngectomies a conclu à une amélioration au cours du temps de la voix des patients, mais aussi à une altération du timbre des voyelles suite à la chirurgie.

L'analyse acoustique de l'organisation du système vocalique de locuteurs francophones âgés de 20 à 93 ans en lien avec l'âge a révélé une évolution distincte entre hommes et femmes chez les sujets âgés, ainsi qu'une différence accrue entre voyelles orales et nasales avec l'âge. Dans le cadre de la thèse de L. Wohmann-Bruzzo, une étude de l'évolution avec l'âge de l'anticipation d'arrondissement dans la production de la consonne /s/ a montré une diminution de la coarticulation anticipatoire chez les locuteurs âgés indépendamment du ralentissement du débit.

Enfin, les résultats d'une étude de la perception par des auditeurs jeunes et âgés de l'âge de locuteurs de différentes variantes régionales du français à partir de la lecture d'une phrase suggèrent que l'aptitude à estimer l'âge d'un locuteur est plus fine lorsqu'il s'agit de pairs en termes de variété régionale et de génération. Une étude complémentaire à partir de productions de locuteurs dysarthriques et témoins a conclu à une surestimation de l'âge des patients dépendante de la sévérité de la dysarthrie.

5.1 Dysarthries

5.1.1 Dysarthrie et espace vocalique

Publications et communications associées :

[ACTI42] **Audibert, N.**, & Fougeron, C. (2012). Distorsions de l'espace vocalique : quelles mesures ? Application à la dysarthrie. *Actes des 19èmes Journées d'Études sur la Parole (JEP 2012)*, Grenoble, France, pp. 217-224.

[ACTI46] Fougeron, C., & **Audibert, N.** (2011). Testing various metrics for the description of vowel distortion in dysarthria. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong-Kong, Chine, pp. 687-690.

En complément de travaux plus métrologiques sur la mesure de l'espace vocalique et de ses variations, présentés en section 7.2.3, en collaboration avec Cécile Fougeron nous nous sommes penchés sur l'application d'une telle méthodologie à la caractérisation des distorsions induites par la dysarthrie sur l'espace vocalique en comparaison de sujets sains d'âge comparable. Le terme de dysarthrie désigne un ensemble de troubles moteurs de la parole d'origine neurologique qui ont pour conséquence une altération de la capacité des patients à contrôler les mouvements articulatoires, ce qui peut entraîner des déficits à n'importe quel niveau du système de production de la parole. L'évaluation de la dysarthrie est souvent subjective, et a été cadrée notamment à travers la grille d'évaluation BECD (Auzou & Rolland-Monnoury, 2006). Pour la reproductibilité et la comparabilité des diagnostics et l'évaluation de la sévérité des troubles et de leur évolution dans le temps, des mesures quantitatives fiables dérivées de la parole produite par les patients sont toutefois nécessaires, comme souligné par exemple par Kent et al. (1999). Les études en la matière cherchent généralement à déterminer si la métrique retenue pour caractériser l'espace vocalique permet de distinguer témoins et patients dysarthriques et/ou à établir un lien entre une métrique et une dimension perceptive telle que l'intelligibilité (voir par exemple Weismer et al. (2001)). Cependant la question des mesures les plus appropriées pour rendre compte des distorsions observées dans divers types de dysarthries au-delà des mesures classiques d'aire de l'espace vocalique n'est pas résolue, et le choix peut s'avérer complexe parmi le foisonnement de mesures proposées dans la littérature sur l'évaluation acoustique de la parole dysarthrique.

Dans deux études successives, nous avons donc évalué la capacité d'un total de douze métriques (certaines n'étant incluses que dans l'une des deux études), directement issues de la littérature ou adaptées aux voyelles orales du français et incluses parmi les mesures présentées en section 7.2.3, à discriminer entre patients dysarthriques et locuteurs témoins et à rendre compte de jugements perceptifs globaux d'intelligibilité et de sévérité de la dysarthrie établis par un ensemble de juges. Une première étude s'est concentrée sur les productions de 16 femmes et 11 hommes atteints de sclérose latérale amyotrophique (SLA), avec une dysarthrie mixte de type paralytique qui se caractérise par une amplitude réduite des mouvements qui sont également lents et instables, ainsi que du même nombre de témoins appariés en sexe et en âge. Dans cette première étude nous avons considéré les métriques suivantes : aire du triangle tVSA défini à partir des voyelles /a, i, u/, aire pVSA du pentagone étendu aux archiphonèmes E (/e, ɛ/) et O (/o, ɔ/), mesures de centralisation FCR et cFCR, distance globale au centroïde, mesures de compression spécifiques à un formant F1RR et

F2RR, taux de recouvrement entre paires de voyelles et taux de recouvrement cumulé tOverlap.

Dans la seconde étude nous avons également pris en compte un groupe de 8 femmes et 22 hommes atteints de la maladie de Parkinson, avec une dysarthrie hypokinétique caractérisée par la réduction des mouvements articulatoire et la perte de modulation prosodique (Darley et al., 1969), et un groupe de 8 femmes et 14 hommes atteints d'un syndrome cérébelleux pur (désignés ci-après comme patients cérébelleux), avec une dysarthrie ataxique qui se traduit par un déficit de coordination spatiale et temporelle qui affecte l'exécution des mouvements articulatoires, l'articulation des voyelles étant généralement considérée comme plus préservée dans ce type de dysarthrie que dans les autres dysarthries (Kent et al., 1979). Dans cette seconde étude, suite à nos observations initiales nous n'avons pas conservé l'aire du triangle tVSA ni la mesure de centralisation FCR, jugées moins informatives sur le français que respectivement l'aire du pentagone pVSA et la mesure de centralisation adaptée au français cFCR, les mesures de recouvrement séparées entre paires de voyelles peu exploitables en pratique, ni la mesure de distance globale au centroïde considérée comme redondante avec la mesure CMinter ajoutée dans la seconde étude au même titre que CMintra et l'indice ϕ qui combine ces deux mesures (Huet & Harmegnies, 2000). En outre, nous avons fait évoluer le mode de calcul du taux de recouvrement entre paires de voyelles afin de tenir compte de la dispersion en deux dimensions des exemplaires vocaliques, pour aboutir à la mesure décrite en section 7.2.3. Les productions des patients et des témoins dans cette tâche de lecture ont été évaluées perceptivement par 10 juges sur différentes dimensions dont l'intelligibilité, sur une échelle de Likert à quatre points dans laquelle le niveau 0 désigne une absence d'altération, et le niveau 3 une altération sévère.

L'ensemble des métriques ont été calculées dans le plan F1*F2, sans considérer les mesures de F3 à la fois à des fins de comparabilité avec la littérature portant majoritairement sur l'anglais, et dans l'optique d'une automatisation de ces mesures en raison de la moindre robustesse de l'extraction automatique de F3 comparativement aux deux premiers formants. Pour cette raison également, nous n'avons pas considéré les voyelles antérieures arrondies pour la caractérisation desquelles la prise en compte du troisième formant est essentielle. Les métriques ont été calculées à partir de l'enregistrement de la lecture d'un texte d'environ 200 mots par chaque locuteur, sur une sélection de 10 à 12 occurrences des cinq catégories de voyelles /i, E, a, O, u/ (dans lesquelles E et O correspondent respectivement aux archiphonèmes /e, ɛ/ et /o, ɔ/) afin que le contexte consonantique soit aussi contrôlé que possible, pour un total de 5 746 voyelles segmentées manuellement.

Comme l'illustre la Figure 31, les patients atteints de SLA et les patients parkinsoniens se distinguent du groupe de locuteurs témoins par une réduction significative de l'aire de l'espace vocalique, qui se traduit par des valeurs réduites de la métrique pVSA, avec une taille d'effet plus importante pour les patients parkinsoniens, ainsi que par des valeurs réduites de CMinter pour le groupe de patients parkinsoniens et des valeurs plus élevées de cFCR pour le groupe de patients SLA. Les patients atteints de SLA se distinguent également du groupe témoin par une dynamique réduite de F2, capturée par la métrique F2RR, qui suggère une mobilité rédute sur l'axe antéro-postérieur (et/ou dans une moindre mesure des variations de l'articulation labiale).

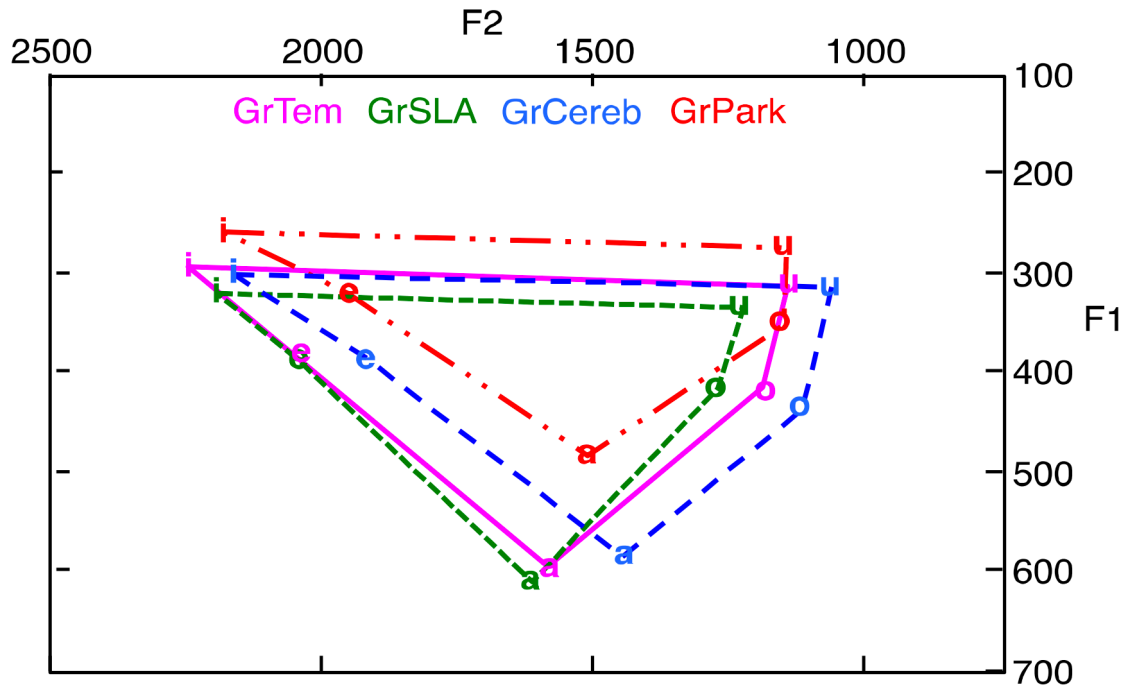


Figure 31 : Représentation dans le plan F1*F2 en Hertz des positions des centroïdes des cinq catégories vocaliques prises en compte (les archiphonèmes /e, ɛ/ et /o, ɔ/ étant notés respectivement e et o sur la figure) reliés par des segments pour visualiser l'aire du pentagone vocalique capturé par la métrique pVSA, pour les quatre groupes de locuteurs comparés. GrTem : témoins sains ; GrSLA : patients atteints de sclérose latérale amyotrophique ; GrCereb : patients atteints de syndrome cérébelleux pur ; GrPark : patients parkinsoniens. D'après Audibert & Fougeron (2012, [ACTI42]).

Le seul examen des espaces vocaliques moyens définis par les centroïdes de chaque catégorie vocalique et le recours à des métriques ne considérant que la centralisation pour quantifier les variations suggèrent une préservation de l'espace vocalique des patients cérébelleux comparativement au groupe témoins, avec un décalage observé sur les valeurs de F2 susceptible de s'expliquer par les spécificités individuelles des patients inclus dans ce groupes comparativement aux autres groupes. Cependant une telle conclusion serait réductrice puisqu'une altération du système vocalique des patients cérébelleux a bien été mise en évidence par des valeurs des métriques CMintra et tOverlap significativement supérieures à celles du groupe témoin, ce qui correspond respectivement à une variabilité accrue de la réalisation des voyelles au sein d'une même catégorie et à un chevauchement plus important entre catégories vocaliques. Un recouvrement significativement plus important entre catégories vocaliques que dans le groupe témoin a également été observé dans les deux autres groupes de patients avec une taille d'effet comparable. De plus les patients parkinsoniens présentent également une variabilité intra-catégorie supérieure à celle des témoins, avec toutefois une taille d'effet moindre que celle observé pour les patients cérébelleux. Ainsi le groupe de patients parkinsoniens est le seul dans nos données qui se distingue significativement des locuteurs témoins sur les trois dimensions principales de la variation de l'espace vocalique, avec à la fois un espace vocalique réduit, une variabilité intra-catégorie accrue et un taux de recouvrement entre catégories plus élevé. Du fait de la réduction de l'espace vocalique combinée à l'augmentation de la variabilité intra-catégorie chez les patients parkinsoniens, ce groupe de patients présente également des valeurs significativement plus

faibles de l'indice ϕ destiné à capturer à l'aide d'une mesure unique ces deux dimensions de la variation de l'espace vocalique.

Bien que la variabilité intra-catégorie et le taux de recouvrement entre catégories vocaliques soient susceptibles de varier de façon partiellement indépendante, la question d'une possible interdépendance plus forte entre ces deux dimensions dans la parole dysarthrique peut se poser. Comme illustré par la Figure 32, bien que le chevauchement entre catégories de voyelles tende à être plus important chez les locuteurs qui présentent une plus grande variabilité intra-catégorie, hormis pour le groupe des patients atteints de SLA pour lesquels la corrélation est plus importante (r = .7), le lien entre ces deux dimensions est faible dans les productions des patients dysarthriques. On observe par ailleurs un effet plancher sur les valeurs de tOverlap, avec pour de nombreux locuteurs (majoritairement mais pas exclusivement dans le groupe témoin) des valeurs nulles de chevauchement cumulé entre catégories. Cet effet plancher peut s'expliquer par le mode de calcul retenu ici, fondé sur les ellipses de dispersion et n'incluant donc pas l'intégralité des exemplaires, sans doute renforcé par le choix de ne prendre en compte que cinq catégories vocaliques dans cette analyse des productions de parole dysarthrique. Cette observation a été l'une de nos motivations pour le développement ultérieur d'une métrique alternative afin de capturer cette dimension de la variation vocalique, présentée en section 7.2.3.2.

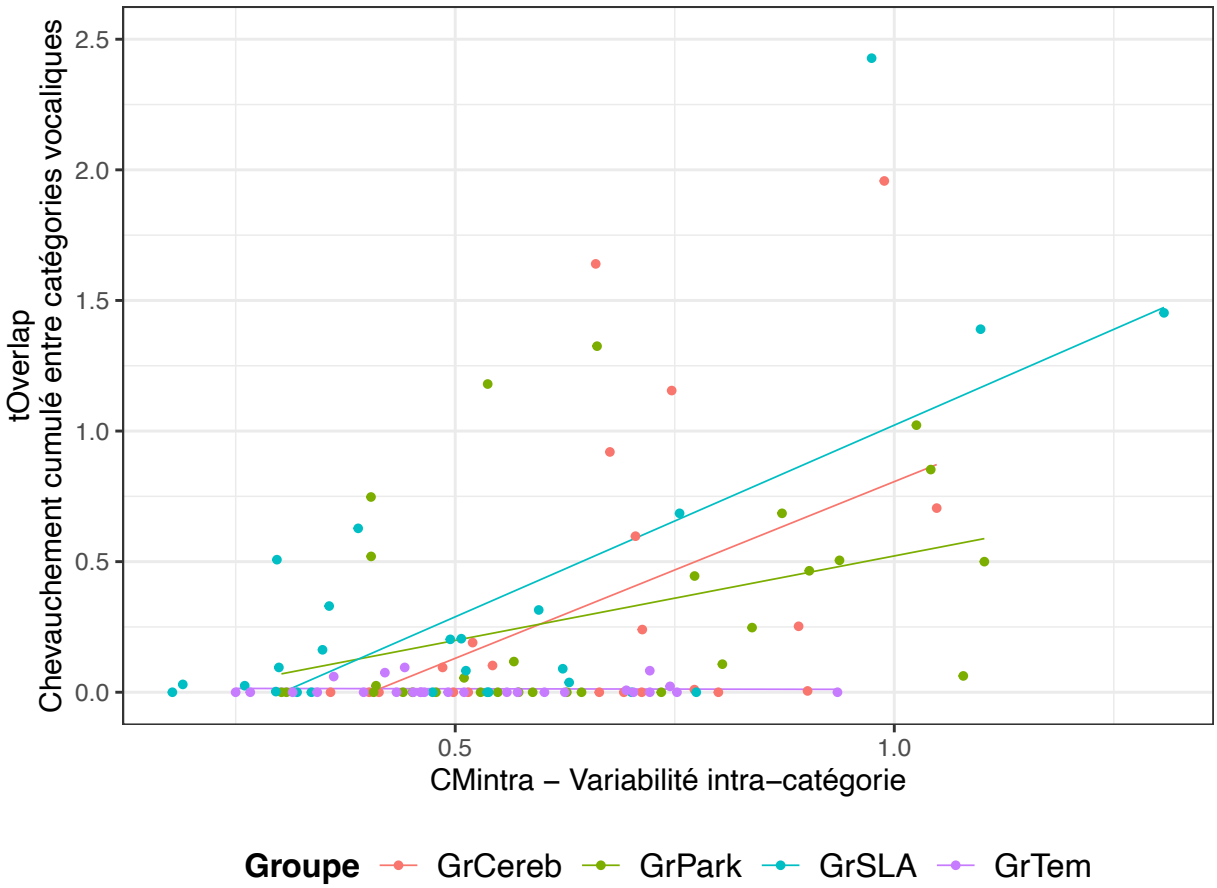


Figure 32 : Représentation sous forme de nuage de points et de droites de régression linéaire du lien entre la variabilité intra-catégorie capturée par la métrique CMintra, et le chevauchement cumulé entre catégories de voyelles capturé par la métrique tOverlap, pour les quatre groupes de locuteurs. Adapté de Audibert & Fougeron (2012, [ACTI42]).

Enfin, nous avons mis en relation les différentes métriques extraites de nos données dans chacun des groupes de locuteurs, et les scores perceptifs d'intelligibilité et de sévérité de la dysarthrie, moyennés entre les 10 juges ayant participé à l'évaluation perceptive. Les liens les plus importants, illustrés par la Figure 33, ont été observés entre l'aire du pentagone pVSA et le degré d'altération de l'intelligibilité estimé par les juges, avec pour le groupe de patients parkinsoniens et le groupe de patients atteints de SLA une intelligibilité jugée plus altérée chez les patients dont l'espace vocalique est le plus réduit (avec des corrélations de respectivement r = -.6 et r = -.7). En revanche ce lien est très faible chez les patients cérébelleux (r = -.3), avec notamment certains patients jugés parmi les moins intelligibles au sein de ce groupe en dépit d'un espace vocalique plus étendu que celui de la majorité des locuteurs témoins. Nous avons observé également un lien entre le chevauchement cumulé entre catégories tOverlap et l'altération de l'intelligibilité dans le groupe de patients parkinsoniens (r = .7), mais pas dans les autres groupes avec des corrélations très faibles voire nulles. En raison du fort lien entre jugements perceptifs d'altération de l'intelligibilité et de degré de sévérité de la dysarthrie (r = .8 pour les patients cérébelleux, r = .9 pour les patients parkinsonien et SLA), les liens entre degré de sévérité et métriques sont très proches de ceux observés entre altération de l'intelligibilité et métriques.

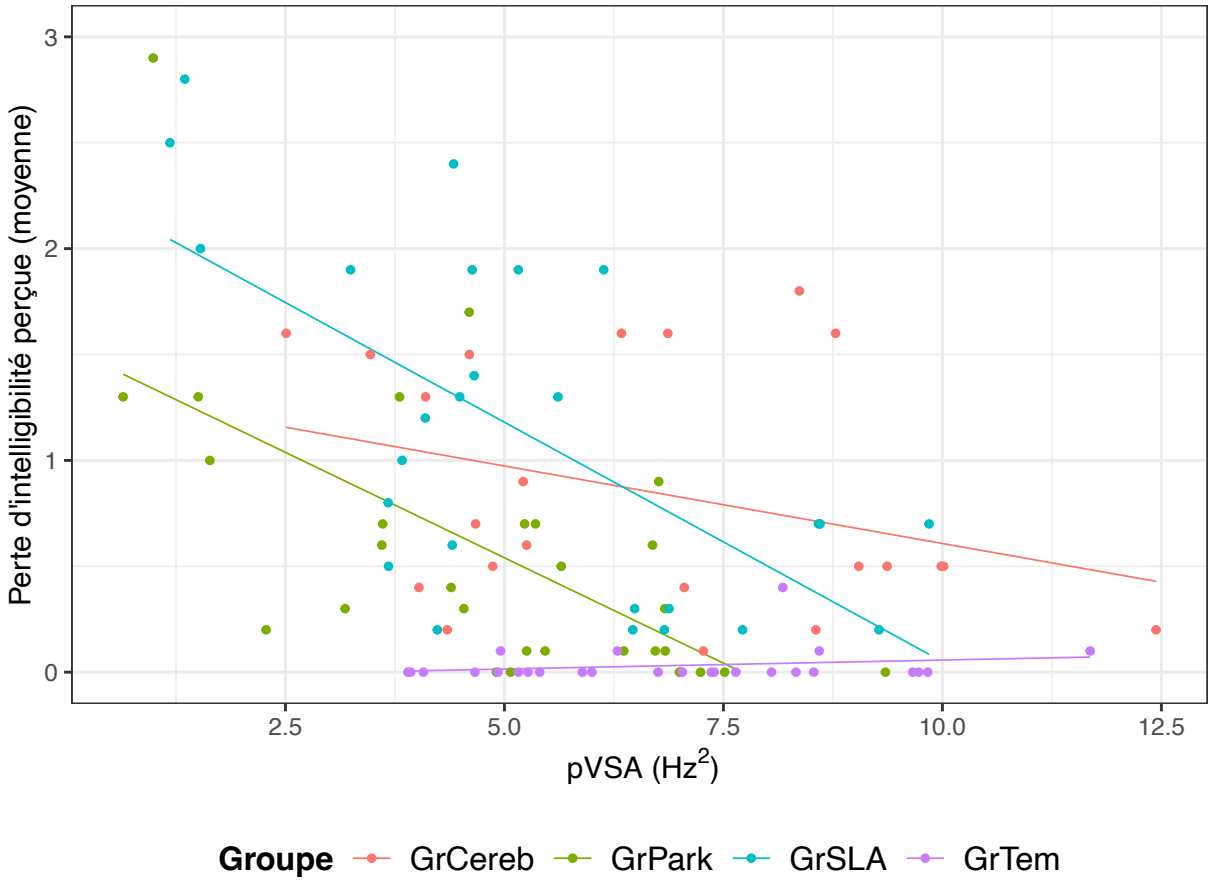


Figure 33 : Représentation sous forme de nuage de points et de droites de régression linéaire du lien entre l'aire du pentagone vocalique pVSA et le degré d'intelligibilité moyen évalué à partir de la grille BECD (Auzou & Rolland-Monnoury, 2006), pour les quatre groupes de locuteurs. Adapté de Audibert & Fougeron (2012, [ACTI42]).

Les résultats obtenus dans ces études, et notamment dans la seconde qui a pris en considération un panel plus varié de dysarthries, soulignent donc la nécessité de prendre en

compte les multiples dimensions de la variation de l'espace vocalique sans se limiter aux mesures classiques d'aire de l'espace vocalique ou autres mesures de dispersion qui, toutes informatives qu'elles soient, ne permettent pas de rendre compte de l'altération de l'espace vocalique observée dans différents types de dysarthries. Cette approche trop restrictive pourrait ainsi expliquer les résultats contrastés dans la littérature sur l'espace vocalique des patients parkinsoniens, décrit comme altéré mais sans que les approches quantitatives permettent de bien rendre compte de cette altération (voir par exemple Weismer et al. (Weismer et al., 2001)). En outre, les analyses réalisées sur nos données ont permis de mieux caractériser l'altération de l'espace vocalique associée à la dysarthrie cérébelleuse.

Les métriques prises en compte, qui consistent en une valeur unique par locuteur, limitent les possibilités d'analyse de la variation individuelle. On peut toutefois noter à partir de l'inspection des données représentées sur la Figure 32 et sur la Figure 33 une importante variation individuelle dans les valeurs prises par les trois métriques représentées. Si certaines plages de valeurs, correspondant à un espace vocalique très réduit, à une variabilité intra-catégorie ou encore à un chevauchement entre catégorie élevés, apparaissent comme spécifiques aux patients dysarthriques, on peut également noter un recouvrement important entre mesures relevées chez les patients et les témoins, qui correspondent pourtant dans certains cas à des patients jugés comme fortement dysarthriques. Aussi utiles soient-elles pour obtenir une description interprétable des altérations du système vocalique associées aux différents types de dysarthrie, il n'est donc pas surprenant que leur pouvoir de discrimination entre patients dysarthriques et témoins ou entre types de dysarthrie reste modéré, et à ce titre elles ne peuvent être utilisées directement comme outil diagnostic.

5.1.2 Intelligibilité de la parole dysarthrique

Publications et communications associées :

[ACTI17] Fougeron, C., **Audibert, N.**, Kodrasi, I., Janbakhshi, P., Pernon, M., Lévêque, N., Borel, S., Laganaro, M., Bourlard, H., & Assal, F. (2022) Comparison of 5 methods for the evaluation of intelligibility in mild to moderate French dysarthric speech. *Proceedings of Interspeech* 2022. Incheon, Korea, pp. 2188-2192.

En collaboration avec Cécile Fougeron ainsi qu'Ina Kodrasi et Parvaneh Janbakhshi de l'IDIAP (Martigny, Suisse) nous avons procédé à partir de données recueillies à l'aide du protocole MonPaGe (Lévêque et al., 2016; Fougeron et al., 2018) à une évaluation comparative de diverses méthodes d'estimation de l'intelligibilité de la parole dysarthrique, fondées sur des jugements perceptifs plus ou moins contrôlés ou sur des mesures automatiques. En effet, si l'évaluation de l'intelligibilité est un élément incontournable de l'évaluation clinique des patients qui permet d'évaluer l'impact du trouble de la parole sur la qualité de vie du patient à travers l'estimation de l'altération de la communication, les méthodes employées pour mesurer le degré d'intelligibilité des productions d'un patient sont multiples et suscitent des débats méthodologiques. Diverses études ont été consacrées à la comparaison de différentes méthodes d'évaluation de l'intelligibilité de la parole (Yorkston & Beukelman, 1978; Kent et al., 1989), les avantages et inconvénient des différentes méthodes découlant généralement des choix méthodologiques à l'origine de la méthode adoptée. Au-delà des scores globaux fréquemment employés dans la pratique clinique et qui englobent à la fois une estimation de l'intelligibilité, de la compréhensibilité et de la sévérité générale du trouble de la parole, des protocoles plus standardisés ont été proposés, dans la plupart des cas pour l'évaluation de

méthodes de remédiation orthophonique. Ces protocoles reposent généralement sur la transcription d'énoncés audio ou des tests de reconnaissance dans lesquels l'intelligibilité est estimée en fonction du nombre d'éléments reconnus. Toutefois la nature des éléments à identifier est très variable, et implique des interprétations diverses de ce qu'est l'intelligibilité, depuis le décodage acoustico-phonétique du signal de parole dans la reconnaissance de non-mots (voir par exemple Ghio et al. (2021)) à la compréhensibilité mettant en œuvre des informations contextuelles pour la reconnaissance de mots dans la parole continue (De Bodt et al., 2002), en passant par la reconnaissance de mots isolés dans laquelle la fréquence lexicale et le voisinage phonologique interviennent également.

Dans la plupart des méthodes, les scores d'intelligibilité sont obtenus via l'évaluation par un ou plusieurs juges des productions des patients, ce qui implique inévitablement une part de subjectivité, que l'évaluation soit établie à partir de transcriptions ou via des choix forcés en fonction du protocole retenu. Plusieurs méthodes automatiques ont été proposées afin de dépasser cette subjectivité et de permettre une évaluation à la fois objective, rapide et économique, qui permette un suivi de l'évolution au cours du temps de l'intelligibilité d'un patient avec des critères reproductibles. Ces méthodes automatiques peuvent être subdivisées en deux grandes catégories, celles fondées sur les mesures extraites du signal acoustique, et celles fondées sur la reconnaissance automatique de la parole (Janbakhshi et al., 2019). Dans les méthodes à partir de mesures extraites du signal acoustique, l'évaluation repose sur le postulat que des caractéristiques acoustiques présentes dans le signal sont susceptibles d'indexer des caractéristiques de parole révélatrices du degré d'intelligibilité, qu'il s'agisse de mesures plus directement interprétables comme par exemple des caractéristiques spectrales (Hummel et al., 2011) ou de représentations du signal apprises par des réseaux de neurones (Maisonneuve et al., 2024). Dans le cas des méthodes qui se fondent sur la reconnaissance automatique de la parole, les systèmes sont entraînés sur de grandes bases de données de productions de locuteurs sains et appliqués aux productions des patients, le taux d'erreurs de reconnaissance (généralement le taux d'erreur mot WER) étant utilisé comme estimation de l'intelligibilité du patient (Schuster et al., 2005; Maier et al., 2009). En dépit des importants progrès techniques récents, les méthodes automatiques restent sujettes à un certain nombre de biais potentiels, notamment liés aux éventuels décalages entre données d'apprentissage et données de test, et les performances de celles qui s'appuient sur des mesures acoustiques extraites du signal sont tributaires de l'équilibre entre catégories dans les données utilisées (voir par exemple Janbakhshi et al. (2019)).

L'étude que nous avons menée a porté sur l'évaluation comparée de cinq méthodes sur les productions de 32 locuteurs francophones avec une dysarthrie légère à modérée et de 17 locuteurs témoins âgés sans dysarthrie attestée. Les locuteurs dysarthriques étaient atteints de diverses pathologies, chacun des sous-groupes comptant huit locuteurs : ataxie de Friedreich, maladie de Parkinson, sclérose latérale amyotrophique, et maladie de Wilson. Le niveau de sévérité a été indexé par les scores de la BECD (Auzou & Rolland-Monnoury, 2006) évalués par un clinicien expert sur une échelle à 20 points dans laquelle le niveau 0 correspond à une absence de dysarthrie. Tous les locuteurs ont été enregistrés avec le protocole MonPaGe, conçu pour l'évaluation quantitative de la parole de patients atteints de troubles moteurs de la parole selon plusieurs dimensions dont celle de l'intelligibilité.

Parmi les cinq méthodes évaluées, trois reposaient sur l'évaluation par des auditeurs. La méthode désignée comme « face à face » correspond au test d'intelligibilité standard du protocole MonPaGe, sous la forme d'une tâche interactive entre l'expérimentateur et le

participant afin de reproduire une situation de communication. Le participant est invité à demander à l'expérimentateur de placer 15 mots sur une grille 5x5 de formes et de couleurs, l'expérimentateur devant transcrire les productions du participant. Ces mots sont sélectionnés aléatoirement parmi un ensemble de 437 mots imagés préselectionnés en fonction de leurs caractéristiques phonologiques et de celles des concurrents lexicaux, les combinaisons de forme et de couleur à utiliser étant également sélectionnées aléatoirement. Le locuteur a pour instruction de toujours formuler ses demandes à l'expérimentateur avec la même phrase porteuse apprise au préalable : « Placez le mot [mot-cible] sur la [couleur] [forme]. », l'expérimentateur ayant pour consigne de toujours écrire quelque chose même lorsque les productions du participant sont inintelligibles afin de ne pas le décourager ni d'induire une hyperarticulation forcée des productions suivantes. Pour chaque mot produit le score de 1 est attribué a posteriori lorsque le mot transcrit correspond à la cible, et le score de 0,5 est attribué lorsque deux interprétations dont une correcte ont été transcrite (possibilité laissée à l'expérimentateur en cas d'hésitation), d'où un score d'intelligibilité par locuteur sur 15 pour les 15 stimuli évalués.

La méthode « multi-juges » correspond aux résultats d'un test en ligne dans lequel des auditeurs naïfs francophones devaient transcrire les productions des participants, avec dans ce cas également la possibilité de donner deux réponses en cas d'hésitation. Chaque participant a été évalué par un minimum de 15 juges parmi les 75 ayant participé. Bien que le panel de juges ait inclus également des auditeurs plus expérimentés dont quelques orthophonistes chevronnées, leurs performances ont été homogènes avec celles des auditeurs naïfs et nous avons donc décidé de fusionner leurs réponses pour l'analyse ultérieure des résultats.

La méthode « experte », plus proche des pratiques cliniques courantes, a consisté en la cotation par une orthophoniste expérimentée du niveau général d'intelligibilité sur une échelle à cinq points correspondant à l'item « intelligibilité » de la grille d'évaluation BECD (Auzou & Rolland-Monnoury, 2006), à partir de la lecture d'un texte court et une production plus spontanée dans une tâche de description d'images. Cette cotation a ensuite été recodée afin d'exprimer l'intelligibilité ainsi cotée sur une échelle directement comparable à celle des deux autres méthodes humaines, le niveau 0 de la BECD étant considéré comme correspondant à un taux d'intelligibilité de 100% et le niveau 4 associé aux altérations les plus importantes comme correspondant à une intelligibilité nulle.

Deux méthodes automatiques ont également été incluses. La méthode exploitant des caractéristiques acoustiques, désignée comme « feature-based », combine plusieurs mesures extractibles automatiquement à partir de signaux non-étiquetés. Ces mesures sont supposées capturer une partie des modulations prosodiques et des variations de qualité de voix, et présentées dans la littérature comme associées à l'intelligibilité de la parole pathologique et notamment de la dysarthrie : le rapport entre modulation de l'énergie spectrale dans les fréquences de modulation inférieures à 4kHz et celles supérieures à 4kHz (Paja & Falk, 2012), le taux de voisement, l'étendue et le coefficient d'aplatissement (kurtosis) de la fréquence fondamentale, ainsi que les valeurs moyennes de jitter et de shimmer (Fang et al., 2017). Ces mesures ont été extraites pour chaque locuteur sur l'énoncé complet incluant la phrase porteuse.

La méthode « ASR » utilisant la reconnaissance de parole s'est appuyée sur un système entrainé avec Kaldi (Povey et al., 2011) sur le corpus SpeechDat (Hoge et al., 1997) complété

par d'autres conversations téléphoniques, sans adaptation spécifique aux données de MonPaGe puisque l'objectif était de comparer les performances de la reconnaissance entre participants et non d'optimiser ces performances. Deux mesures de performance sont prises en compte : le taux d'erreur mot WER estimé sur l'énoncé complet, et la précision de reconnaissance du mot-cible TWA ne considérant que la performance de la reconnaissance automatique sur ce mot-cible. Enfin une méthode automatique composite a été considérée en combinant au moyen d'une régression linéaire régularisée (qui permet d'améliorer les performances de prédiction de la régression) les scores issus de chacun des paramètres considérés dans la méthode « feature-based » et les scores issus de la méthode « ASR ».

Une première constatation tirée de l'analyse des résultats a été que, du fait de la sévérité limitée des dysarthries considérées, certains patients étaient considérés comme parfaitement intelligibles par au moins l'une des méthodes évaluées. A l'inverse, certains des témoins âgés n'étaient pas considérés par l'ensemble des méthodes comme parfaitement intelligibles. La comparaison entre méthodes reposant sur un ou plusieurs jugements humains a révélé une consistance plus importante entre la méthode « face-à-face » et la méthode « multi-juges » qu'entre la méthode experte et chacune de ces deux méthodes, avec une corrélation entre jugements attribués par la méthode experte et la méthode multi-juge qui n'est que de ρ=.58. Les jugements se sont montrés particulièrement divergents entre méthodes d'évaluation de l'intelligibilité dans le cas des 30 locuteurs considérés comme n'étant pas parfaitement intelligibles d'après les résultats de la méthode « face-à-face », ce qui peut s'expliquer par le lien fort entre la sévérité de la dysarthrie et la cotation de l'intelligibilité dans les jugements experts (ρ=-.87) alors que ce lien est sensiblement moins important pour les méthodes « face-à-face » (ρ=-.58) et « multi-juges » (ρ=-.54).

Tous groupes de locuteurs confondus, la méthode « face-à-face » tend à aboutir à une évaluation plus élevée du taux d'intelligibilité, qui pourrait s'expliquer par l'exploitation par les évaluateurs d'indices visuels, qui incluent bien entendu les corrélats visuels de l'articulation segmentale mais aussi les expressions faciales et autres gestes manuels (L. Hunter et al., 1991; Keintz et al., 2007), ainsi que d'indices contextuels en complément des indices acoustiques lors de la transcription des productions des participants. La méthode experte tend quant à elle à aboutir à des jugements plus tranchés dans un sens ou l'autre, en partie en raison du mode d'évaluation de la grille BECD avec l'utilisation d'une échelle à cinq points qui favorise le recours aux valeurs extrêmes de l'échelle. Les divergences entre méthodes ont une incidence sur la proportion de locuteurs considérés comme parfaitement intelligibles, avec 25 des 49 locuteurs considérés comme tels à partir de la méthode « face-à-face » et jusqu'à 27 locuteurs avec la méthode experte, mais seulement 3 avec la méthode « multi-juges ».

Comme l'illustre la Figure 34, des divergences entre méthodes ont également été relevées dans les jugements d'intelligibilité attribués aux différentes populations dysarthriques. Si le décalage entre évaluations « face-à-face » et « multi-juges » est relativement consistant entre groupes de locuteurs, avec toutefois un décalage moindre dans le cas des patients atteints d'ataxie de Friedreich, on peut remarquer en particulier que les jugements experts attribuent une intelligibilité plus faible en comparaison des autres méthodes dans les groupes qui comprennent les patients les plus sévèrement dysarthriques (ataxie de Friedreich et maladie de Wilson, avec un grade de dysarthrie moyen respectif de 2,9 et 2,5 sur 4 contre 2,1 pour les patients Parkinsoniens ou atteints de sclérose latérale amyotrophique).

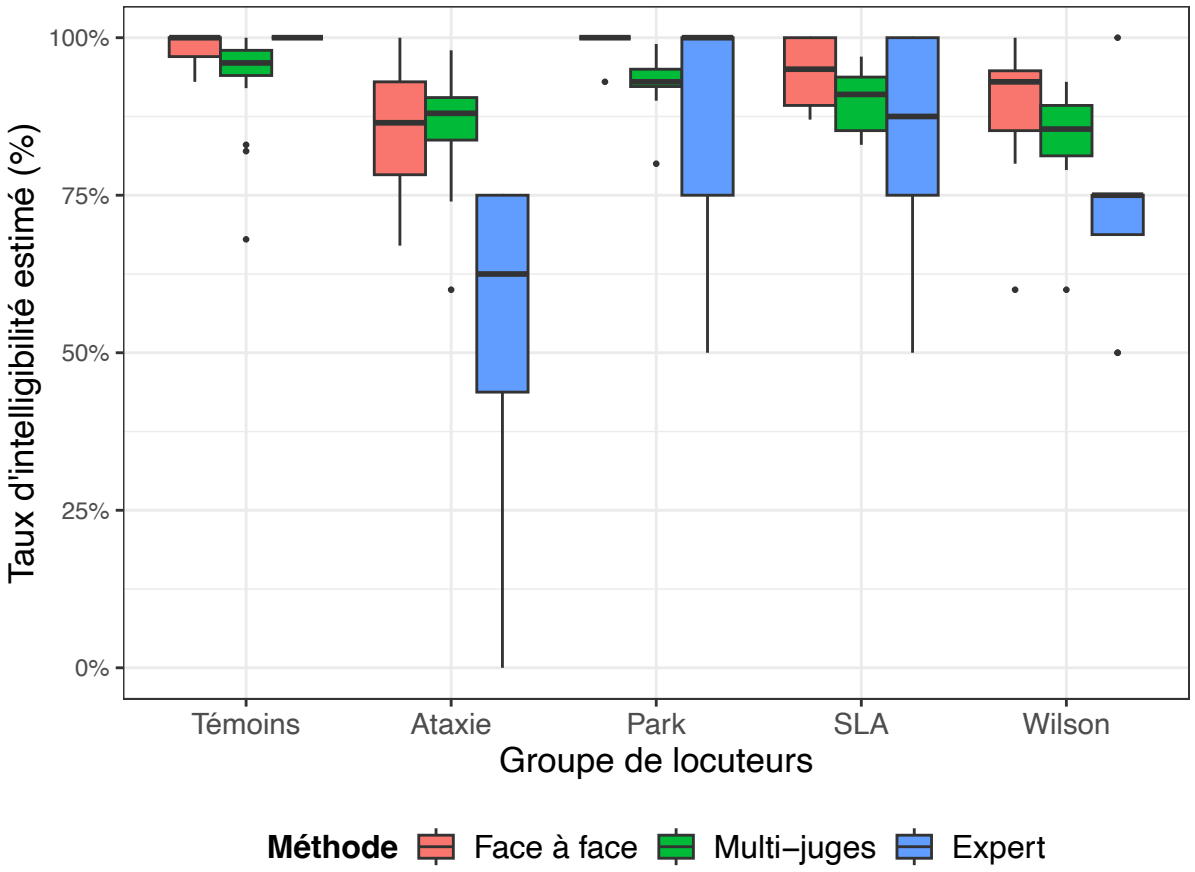


Figure 34 : Distribution des taux d'intelligibilité estimés par les trois méthodes faisant appel à l'évaluation par des juges humains, pour les 17 locuteurs témoins et pour chacun des quatre groupes de 8 locuteurs dysarthriques : Ataxie = ataxie de Friedreich ; Park = maladie de Parkinson ; SLA = sclérose latérale amyotrophique ; Wilson = maladie de Wilson. Adapté des données présentées dans Fougeron et al. (2022, [ACTI17]).

Les méthodes automatiques ont quant à elles été évaluées en les comparant aux évaluations par les juges via des corrélations de Spearman, comme illustré par la Figure 35. Dans l'ensemble, les mesures d'intelligibilité fondées sur l'extraction de caractéristiques acoustiques sont principalement corrélées avec l'évaluation « face-à-face » et avec l'évaluation experte, bien que de façon modérée ($\rho \approx .50$). Les corrélations les plus fortes avec ces évaluations par des juges ont été observées en considérant le jitter et la mesure de modulation spectrale LHMR, ce qui suggère que l'évaluation de l'intelligibilité pourrait être influencée par l'instabilité de la voix et l'altération de la dynamique temporelle, tout particulièrement lorsque l'évaluateur interagit avec le patient comme dans la condition « face-à-face », ou est amené à prendre en compte une plus grande quantité de parole continue comme dans l'évaluation experte. Pour leur part, les mesures issues de la reconnaissance automatique de la parole sont mieux corrélées avec la méthode « multi-juges » ($\rho \approx .68$), ce qui suggère que cette approche rend mieux compte des performances d'identification des mots par les auditeurs n'ayant accès qu'au signal acoustique. Enfin, on peut noter que la méthode automatique composite issue de la combinaison entre méthode « feature-based » et méthode « ASR » permet d'améliorer les performances de l'estimation automatique de l'intelligibilité au sens de corrélations plus élevées avec les évaluations par les juges ($\rho = .70$ pour les méthodes « face-à-face » et « multi-juges » à $\rho = .74$ pour la méthode experte). Si ces performances restent en-deçà de l'état de l'art en matière de systèmes d'évaluation automatique de l'intelligibilité, avec par exemple une

corrélation de r=.80 avec la référence humaine dans l'étude de Maisonneuve et al. (2024), le gain apporté par cette méthode composite suggère non seulement que le croisement de différentes approches pourrait permettre une amélioration de l'évaluation automatique de l'intelligibilité, mais aussi que l'évaluation de l'intelligibilité par l'humain est influencée par de multiples facteurs qui ne se limitent pas au décodage du message linguistique.

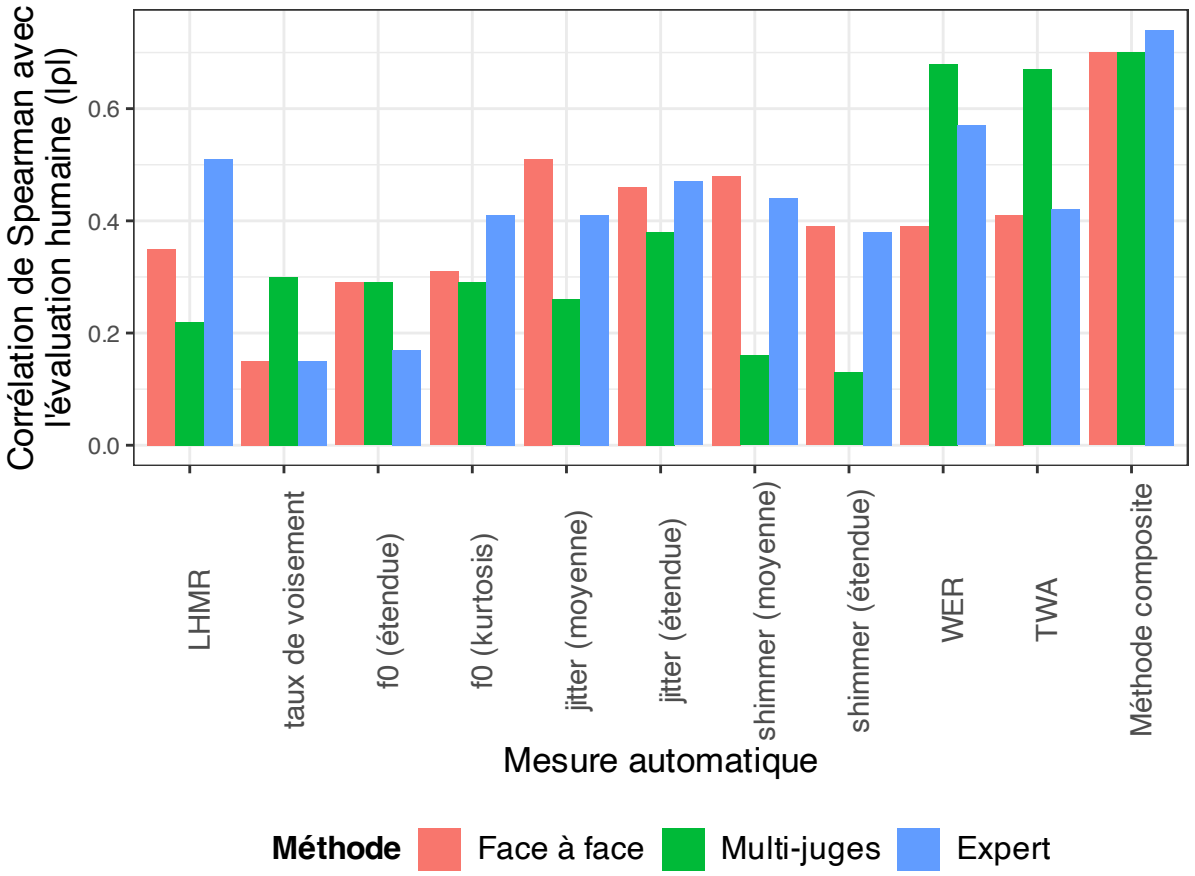


Figure 35 : Corrélations de Spearman entre estimation de l'intelligibilité à partir des différentes méthodes automatiques considérées et les trois méthodes d'évaluation par des juges humains. Les corrélations sont présentées en valeur absolue pour simplifier l'interprétation des performances relatives des différentes méthodes. Adapté des données présentées dans Fougeron et al. (2022, [ACTI17]).

5.2 Dysphonies et voix de substitution

Plusieurs de mes travaux en phonétique clinique ont porté sur les dysphonies plus ou moins sévères. La dysphonie consiste en un trouble vocal dont la définition précise est fluctuante dans la littérature, mais qui peut être considérée comme une altération du timbre de la voix qui conduit à une plainte de la part des patients (Crevier Buchman, 2001). On peut distinguer la dysphonie organique qui implique une atteinte organique des plis vocaux, par exemple suite à un cancer ou dans le cas d'une paralysie de l'un des plis vocaux, de la dysphonie fonctionnelle dans laquelle les plis vocaux ne sont pas directement atteints et qui résulte d'un geste vocal inadapté (Crevier Buchman et al., 2005). L'origine de cette inadaptation du geste vocal, souvent considérée comme liée à un déficit du contrôle musculaire au niveau du larynx, reste toutefois discutée et de nombreux cas cliniques de

dysphonies se situent à la frontière de cette définition et peuvent être considérées comme partiellement organiques (Roy, 2003).

Dans les cas de cancer touchant le larynx, une laryngectomie peut être nécessaire afin de maximiser les chances de survie du patient. Lorsque la gravité de l'atteinte n'exige pas une laryngectomie totale, une laryngectomie partielle est réalisée en préservant autant que possible non seulement le conduit respiratoire mais aussi la fonction phonatoire via le maintien d'au moins une unité crico-aryténoïdienne (Crevier-Buchman et al., 1995). Les structures anatomiques concernées varient en fonction du degré et de la nature de l'atteinte, toutefois la résection voire l'ablation d'au moins l'un des plis vocaux aboutit à l'utilisation par le patient d'une voix dite de substitution, c'est-à-dire produite sans véritable vibration des deux plis vocaux (Moerman et al., 2005).

5.2.1 Dysphonies légères chez les professeures des écoles

Publications et communications associées :

[ACL1] Pettirossi, A., **Audibert, N.**, & Crevier Buchman, L. (2024). Impact of Dysphonic Schoolteachers' Voices on Children's Reaction Times according to Phonemic Contrasts. *Folia Phoniatrica et Logopaedica*. doi:10.1159/000539562.

[ACTI4] Pettirossi, A., **Audibert, N.**, & Crevier Buchman, L. (2024). Frontières entre la perception de la voix normophonique et pathologique chez des auditeurs naïfs. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 401-411.

[ACTI11] Pettirossi, A., **Audibert, N.**, & Crevier Buchman, L. (2023). Strategies of vocal adaptation to background noise of dysphonic and control schoolteachers. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic. pp. 3932-3936.

[ACTI22] Pettirossi, A., **Audibert, N.**, & Crevier Buchman, L. (2020). Corrélats acoustiques et perceptifs de la personnalité perçue à travers la voix dans une population de dysphoniques légères. *Actes des 33èmes Journées d'Études sur la Parole*, 2020, Nancy, France, pp. 489-497.

[COM4] Pettirossi, A., **Audibert, N.**, & Crevier Buchman, L. (2019). L'âge des élèves : un facteur influençant la dysphonie chez les enseignantes. *8èmes Journées de phonétique clinique (JPC8)*, Mons, Belgique.

La thèse d'Amelia Pettirossi (2021), que j'ai codirigée avec Lise Crevier Buchman et qui a été soutenue en janvier 2021, a porté sur la dysphonie modérée dont sont fréquemment atteintes les femmes professeures des écoles en exercice, les représentations associées à ces atteintes vocales, et leur impact sur la réception du message linguistique par des enfants d'âge scolaire. Le professorat des écoles, dans lequel les femmes sont largement majoritaires, est en effet considéré comme à risque pour les troubles de la voix en raison de la surutilisation de leur voix en contexte professionnel (voir par exemple Titze (1999)) et du niveau élevé de bruit dans les écoles élémentaires et primaires (voir par exemple Shield & Dockrell (2004)) qui tend à favoriser les situations de forçage vocal.

Après une première étude menée à partir de questionnaires en ligne auprès de 709 professeures des écoles françaises en exercice qui a montré qu'une large part avaient déjà rencontré des troubles vocaux dans le cadre de leur profession avec une prévalence plus

importante de ces troubles vocaux chez celles enseignant aux élèves les plus jeunes, un panel de 61 locutrices ayant répondu à cette première enquête a été recruté. Ces locutrices étaient réparties de façon équilibrée entre les trois cycles de l'enseignement élémentaire et primaire, et majoritairement non fumeuses. Elle ont été enregistrés dans un environnement calme à l'aide d'une station d'acquisition KayPentax couramment utilisée en phoniatrie sur la production de trois /a/ tenus à hauteur et intensité confortable, deux répétitions de la lecture de paires minimales monosyllabiques avec une structure CVC insérées à la suite de la phrase porteuse « Clique sur le dessin de … », et la lecture de la fable « La bise et le soleil » dans une condition confortable et dans une condition simulée de lecture face à une classe bruyante. Une cotation selon l'échelle GRBAS (Hirano, 1981) réalisée par deux expertes sur les /a/ tenus et sur deux phrases extraites du corpus de paires minimales a permis de catégoriser les locutrices en 37 témoins et 24 dysphoniques légères en fonction du grade (G dans l'échelle GRBAS) attribué, seules deux locutrices étant évaluées au grade 2 et aucune n'étant évaluée au grade 3 qui est le plus sévère de cette échelle.

Une analyse acoustique comparative des deux conditions de lecture de la fable « La bise et le soleil » a été menée afin d'évaluer les stratégies vocales d'adaptation au bruit adoptées par les locutrices, et un éventuel impact de la dysphonie sur ces capacités d'adaptation. Pour cela nous avons extrait de chaque enregistrement les portions détectées comme voisées, et avons calculé à partir de ce matériel le spectre moyen à long terme LTAS, avec une largeur de bande de 50 Hz. Nous avons opté pour la version corrigée pour compenser les variations de fréquence fondamentale du LTAS, évaluée précédemment comme plus robuste aux variations intra-locuteur dans une étude présentée aux Journées de Phonétique Clinique (Pettirossi et al., 2017, [AFF7]), afin de mieux tenir compte des variations de durées segmentales entre conditions et locutrices et de la présence de disfluences dans certaines productions.

La Figure 36 illustre la comparaison de ces spectres moyens à long terme, moyennés pour chaque groupe de locutrices et chaque condition. Un renforcement peut être observé dans les deux groupes de locutrices entre 1 kHz et 3 kHz en condition « bruyante », consistant avec les résultats de Garnier & Heinrich (2014) qui ont mis en évidence un renforcement des fréquences autour de 3 kHz associé à l'effet Lombard caractéristique de la parole dans le bruit et qui permet de renforcer les contrastes entre la parole et le bruit, toutefois cet effet est moins fort chez les locutrices dysphoniques que chez les locutrices témoins. De plus, une énergie plus importante entre 5 kHz et 7 kHz est également observée chez les locutrices dysphoniques dans les deux conditions, en accord avec la littérature sur les caractéristiques spectrales des voix dysphoniques (Kitzing & Åkerlund, 2009), mais avec un effet plus fort en condition « bruyante ». Ainsi, dans cette condition simulée de lecture à destination d'une classe bruyante, les professeures des écoles dysphoniques montrent une adaptation au bruit similaire aux locutrices témoins, mais produiraient une voix moins efficace et plus fortement dysphonique.

Une autre question liée à la dysphonie qui touche largement les professeures des écoles est celle de ses conséquences pour les élèves auxquels elles s'adressent quotidiennement dans l'exercice de leur profession. Cet aspect a peu été traité dans des études expérimentales, toutefois l'effet délétère du bruit sur l'intelligibilité de la parole adressée aux élèves par leur enseignante qui a été mis en évidence dans plusieurs études (voir par exemple Finitzo-Hieber & Tillman (1978)) suggère que la voix dysphonique, elle aussi porteuse de bruit, pourrait impacter le décodage de la parole par les élèves. Cette hypothèse a été évaluée dans deux études récentes, concomitantes au travail de thèse d'Amelia Pettirossi. Dans la première (I. S.

Schiller et al., 2020), une tâche d'identification de non-mots produits par une locutrice normophonique dans une condition de contrôle et en imitant une voix fortement dysphonique a été soumise à des enfants d'âge scolaire et a montré des temps de réaction accru pour tous les contrastes phonémiques évalués. La seconde (Oliveira et al., 2024) a consisté en la diffusion en classe d'enregistrement de phrases produits par des locuteurs dysphoniques ou non et à l'analyse des transcriptions de ces phrases par les élèves, mettant en évidence un nombre d'erreurs plus élevé en cas de dysphonie modérée et sévère. Dans l'étude que nous avons menée, une tâche perceptive adaptée à des enfants de 6 à 10 ans a été conçue en exploitant les paires minimales insérées dans la phrase porteuse « Clique sur le dessin de … », enregistrée dans cet objectif et combinée à la présentation d'illustrations des deux éléments de la paire minimale, les enfants ayant pour consigne de répondre le plus rapidement possible via un boîtier aux dimensions adaptées à leur âge. Les résultats ont montré un effet de la dysphonie sur le décodage du contraste de voisement par les enfants, avec des temps de réaction significativement plus longs pour les locutrices dysphoniques, mais pas pour les contrastes de nasalité, d'arrondissement, d'antériorité/postériorité de la voyelle ni de lieu d'articulation consonantique également évalués.

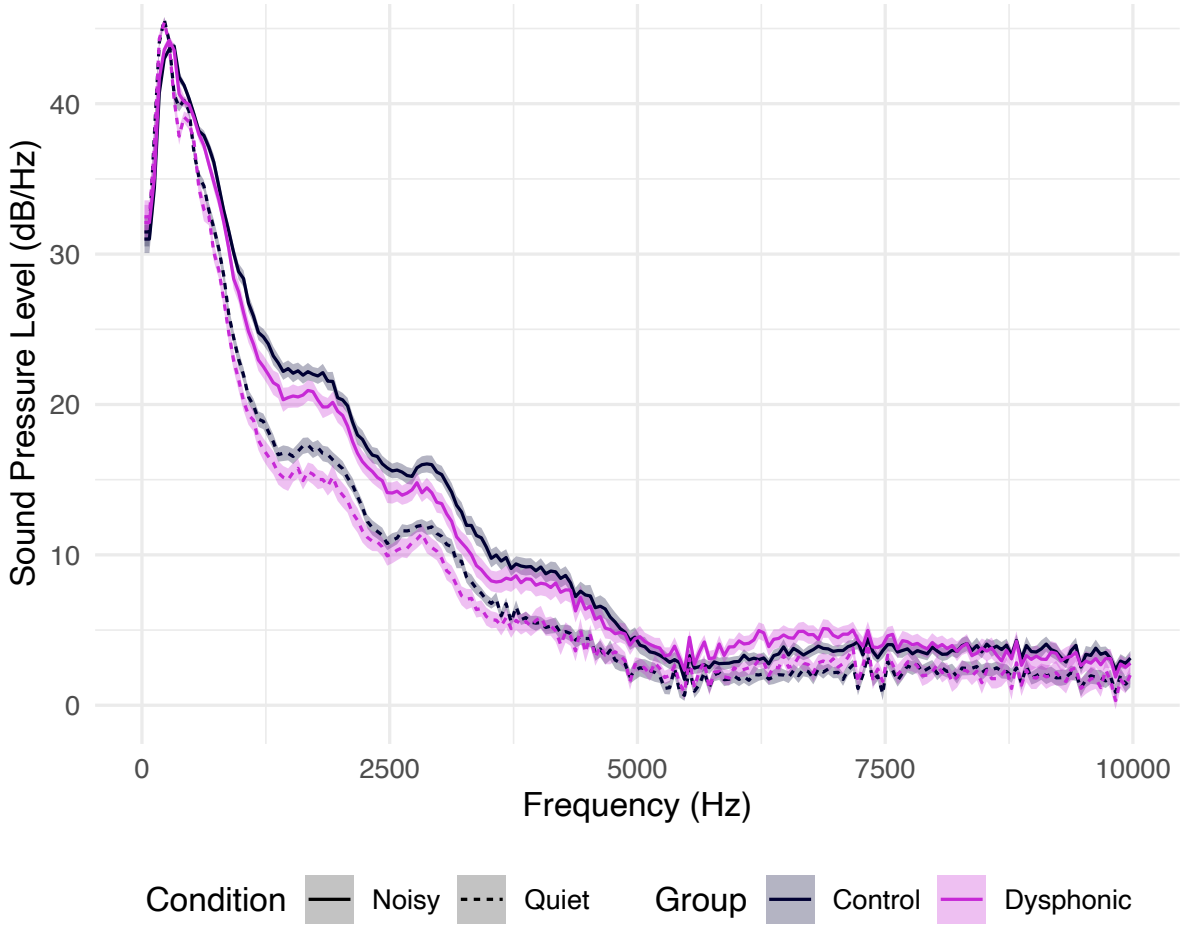


Figure 36 : Spectre moyen à long terme (LTAS) corrigé des variations de fréquence fondamentale comparé entre locutrices témoin et locutrices dysphonique lors de la lecture d'un texte, en condition calme et en simulant la lecture face à une classe bruyante. Seules les fréquences jusqu'à 10kHz sont représentées. Les courbes représentent le LTAS moyenné pour chaque groupe de locutrices et condition, et l'enveloppe colorée représente l'erreur-type. D'après Pettirossi et al. (2023, [ACTI11]).

Une évaluation perceptive a été mise en place afin d'évaluer dans quelle mesure les troubles vocaux que présentent certaines locutrices influencent la représentation de leur personnalité que des auditeurs naïfs peuvent construire à partir de leur voix. En effet, de façon consciente ou non, les auditeurs tendent à établir des liens entre la voix d'une personne et sa personnalité (voir par exemple Kreiman & Sidtis (2011) pour une revue avec une perspective historique), y compris dans le cas de voix de synthèse (Nass & Lee, 2001). Cette évaluation perceptive a été menée sur un ensemble d'échelles de Likert à cinq points, selon le principe des échelles perceptives différentielles (Osgood, 1952), les échelles sélectionnées étant issues d'un prétest de catégorisation libre suivi d'une validation pour éliminer les termes redondants ou ambigus. Les échelles différentielles suivantes relatives aux traits de personnalités attribués à partir de l'écoute des voix ont ainsi été retenues : Joyeuse/Triste, Sympathique/Désagréable, Dynamique/Molle, et Confiante/Hésitante. En complément, une échelle relative à l'évaluation par les auditeurs de la présence d'un trouble vocal (Aucun trouble vocal/Trouble vocal sévère) a été intégrée dans l'évaluation perceptive, menée sur le premier paragraphe de la fable « La bise et le soleil » en condition de lecture confortable. Les extraits présentés ont été évalués par 40 auditeurs, l'ordre et l'orientation des échelles étant randomisés de même que l'ordre de présentation des stimuli. Les scores attribués à chaque locutrice sur chacune des échelles ont ensuite été moyennés entre auditeurs pour l'analyse des résultats. Les résultats de cette évaluation ont révélé une tendance des auditeurs à attribuer des traits de personnalité plus négatifs aux locutrices également évaluées comme présentant un trouble vocal plus important, avec des corrélations comprises entre r = .59 pour l'échelle Dynamique/Molle et r = .79 pour l'échelle Sympathique/Désagréable.

La mise en relation des jugements perceptifs sur les quatre échelles correspondant aux traits de personnalité et des mesures de fréquence fondamentale moyenne, de durée syllabique moyenne retenue comme estimation du débit de parole et des mesures de périodicité du signal HNR (rapport entre énergie harmonique et énergie du bruit) et ZCR (taux de passage par zéro du signal acoustique par unité de temps) a suggéré une préférence des auditeurs pour une fréquence fondamentale élevée, un débit syllabique rapide et une périodicité moindre, avec toutefois des corrélations modérées à l'exception de l'association d'un débit rapide à une personnalité dynamique (r = .74). Si la préférence pour une voix moins périodique qui pourrait correspondre à une raucité plus importante peut sembler contre-intuitive, elle est cohérente avec les résultats de Barkat-Defradas et al. (2012) qui ont montré que les voix féminines rauques en français tendaient à être perçues comme plus attractives.

L'évaluation perceptive du trouble vocal a ensuite été complétée par une évaluation similaire par 30 auditeurs naïfs du degré d'atteinte vocal dans un groupe de dix patientes dysphoniques (dont cinq professeures des écoles) recrutées dans le cadre d'une consultation ORL suite à un trouble vocal diagnostiqué, et pour lesquelles une évaluation experte selon l'échelle GRBAS a également été effectuée. Ce groupe comprenait deux patientes évaluées au grade 1, trois au grade 2, et cinq au grade 3 le plus sévère. Pour chacun des deux groupes de locutrices, le jugement naïf du degré de trouble vocal a été mis en relation avec les différentes dimensions de l'échelle GRBAS obtenues par cotation experte. Comme illustré par la Figure 37, tandis que les jugements experts du grade de dysphonie sont liés principalement au degré de raucité suivi du degré de souffle et de serrage vocal, les jugements naïfs de degré de trouble vocal sont significativement moins liés au degré de raucité dans les deux groupes, ainsi qu'au degré de souffle dans le groupe des professeures des écoles. Ces jugements naïfs de degré de trouble vocal dépendent majoritairement du degré de serrage vocal dans le groupe des

patientes. De façon plus étonnante ce lien entre perception du trouble vocal et serrage vocal est très faible dans le groupe des professeures des écoles, mais cela peut s'expliquer par le fait que ce groupe n'inclut que quatre locutrices avec un serrage vocal léger.

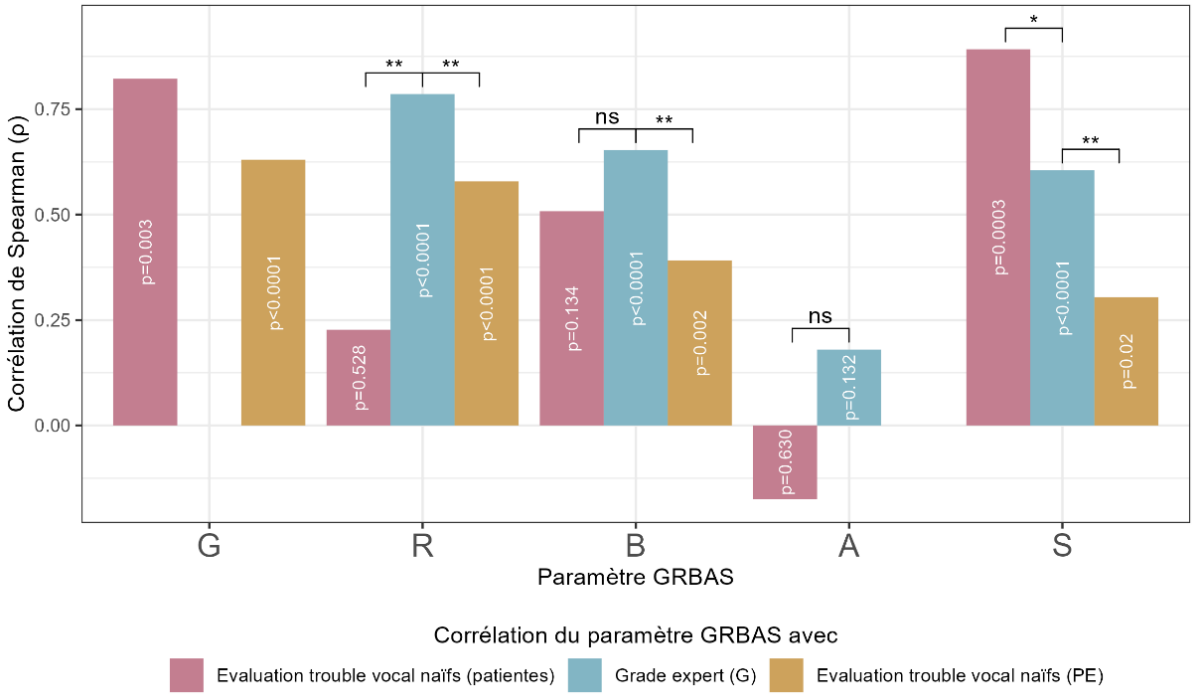


Figure 37 : Corrélation entre le grade de dysphonie attribué par l'évaluation experte et les autres dimensions de l'échelle GRBAS toutes locutrices confondues (bleu) et entre les dimensions GRBAS et l'évaluation naïve du degré de trouble vocal pour chacun des deux groupes de locutrices (rose et ocre). Les corrélations sont comparées par la méthode de Meng et al. (1992). D'après Pettirossi et al. (2024, [ACTI4]).

5.2.2 Paralysie laryngée unilatérale et expression émotionnelle

Communication associée :

[AFF1] Petrone, C., **Audibert, N.**, Haddad, R., Robert, M., Trocq, M., Mattei, A., & Lalain, M. (2024). Vocal expression of emotions in patients with unilateral vocal fold paralysis. *13th International Seminar on Speech Production (ISSP)*, Autrans, France.

Dans le cadre d'une collaboration amorcée récemment avec Caterina Petrone et Muriel Lalain du Laboratoire Parole et Langage et qui inclut également des cliniciens impliqués dans le suivi et le recrutement des patients, je me suis penché sur la question des capacités de modulation de la fréquence fondamentale par des patients atteints d'une paralysie laryngée unilatérale (PLU). La PLU consiste en une immobilité de l'un des plis vocaux, résultant d'une interruption de la commande nerveuse motrice des muscles du larynx (Alwan & Paddle, 2022). Les symptômes typiques de la PLU comprennent la dysphonie et l'instabilité de la vibration laryngée, et peuvent conduire à des ajustements compensatoires qui ont pour conséquence d'accroitre encore l'effort vocal des patients (El-Banna & Youssef, 2015). Les patients atteints de PLU ont une voix faible, soufflée et rauque, et montrent une intensité de la voix plus faible, une diplophonie et une fuite glottique (Lotto et al., 1997; Jesus et al., 2015). Sur le plan acoustique, la PLU se traduit par des valeurs plus élevées de jitter et de shimmer, des valeurs

plus faibles du rapport entre énergie harmoniques et bruit (HNR) et une plus grande variabilité de la fréquence fondamentale (f0) par rapport aux témoins sains (Jesus et al., 2015).

L'étude préliminaire que nous avons menée comme point de départ de travaux communs à venir trouve son origine dans les plaintes récurrentes formulées par les patients atteints de PLU auprès de l'équipe soignante qui les suit quant au décalage entre l'émotion que les patients ont l'intention d'exprimer et celle effectivement transmise par leur voix, le cas le plus courant étant d'être considérés à tort comme exprimant de la colère lors d'appels téléphoniques pour la prise de rendez-vous en raison de leur altération vocale. Nous avons donc cherché à comparer à partir de mesures acoustiques les capacités de modulation de la fréquence fondamentale et la qualité de voix de ces patients dans un contexte d'expressions émotionnelles simulées en les comparant aux mêmes productions par un groupe témoin.

Cinq hommes et cinq femmes francophones atteints de PLU et âgés en moyenne de 66 ans et dix locuteurs témoins appariés en âge et en sexe ont été inclus dans cette première analyse. Afin d'assurer l'homogénéité des données, tous les patients inclus présentent une PLU post-opératoire, par exemple consécutive à une thyroïdectomie ou une chirurgie cardiaque, sans dysarthrie ni troubles neurologiques ou psychiatriques. La dysphonie des patients a été évaluée sur l'échelle GRB (Hirano, 1981) et les patients ont également complété la version française du questionnaire Voice Handicap Index (Woisard et al., 2004) qui consiste en une auto-évaluation de l'incidence des troubles vocaux sur la qualité de vie. L'ensemble des locuteurs ont produit un ensemble de huit phrases courtes de cinq à neuf syllabes, de structure syntaxique comparable, dont la neutralité affective du contenu verbal a été évaluée dans une étude préalable. Chaque phrase était insérée dans trois contextes différents, évoquant trois états émotionnels différents (neutre/triste/colère), les phrases étant groupées par bloc en fonction de l'état émotionnel visé, chacun de ces blocs étant précédé d'une phase d'entrainement. L'instruction donnée aux locuteurs était de lire tous les contextes et les phrases cibles en silence, puis de produire les phrases cibles de manière la plus naturelle possible sans les lire. Si un tel protocole est susceptible d'aboutir à des productions peu écologiques et ne correspond pas à l'évolution de mon approche en matière de parole expressive que je développe en section 3, ce type d'approche s'est imposé du fait de l'impossibilité de solliciter les patients pour des sessions d'enregistrement plus longues et de la nécessité d'obtenir des données directement comparables pour cette première analyse.

Un total de 480 énoncés a ainsi été collecté. Pour cette première analyse, seule la voyelle /a/ présente dans chaque phrase a été prise en compte après une segmentation manuelle, et nous nous sommes contentés d'une analyse par groupe de locuteurs en comparant patients et témoins sans approfondir la variation individuelle en lien avec l'évaluation perceptive et les réponses au questionnaire Voice Handicap Index. La fréquence fondamentale a été extraite à l'aide de l'outil FCN-f0 (Ardaillon & Roebel, 2019), évalué comme plus fiable sur la parole pathologique que les autres algorithmes de détection de la fréquence fondamentale existants (Vaysse et al., 2022), les valeurs de fréquence fondamentale ainsi obtenues étant validées à partir de l'inspection des spectrogrammes à bande large sur les productions des patients les plus dysphoniques. Afin de caractériser d'éventuelles variations de qualité de voix, susceptible de dépendre de l'émotion exprimée (Gobl & Ní Chasaide, 2003), le rapport entre énergie harmonique et bruit (HNR) mesuré sur les fréquences supérieures à 1000 Hz (en raison de la sensibilité de cette mesure aux variations de f0, peu documentée dans la littérature mais observée par ailleurs) et la mesure cepstrale CPPS qui est considérée comme la mesure

acoustique la mieux corrélée à la dysphonie (Heman-Ackah et al., 2014) ont également été extraits au milieu de chaque /a/, de même que l'enveloppe spectrale entre 0 et 5000 Hz.

Les occurrences de /a/ analysées présentent une durée significativement supérieure chez les patients, qui pourrait découler de stratégies compensatoires en raison de l'effort supplémentaire que nécessite la production de la voix chez ces patients. Par ailleurs les valeurs de CPPS plus faibles confirment que les patients sont plus dysphoniques que les témoins, sans qu'une différence entre catégories émotionnelles ne soit observée. Les autres mesures acoustiques sont présentées dans la Figure 38, comparées entre groupes de locuteurs et catégorie émotionnelle.

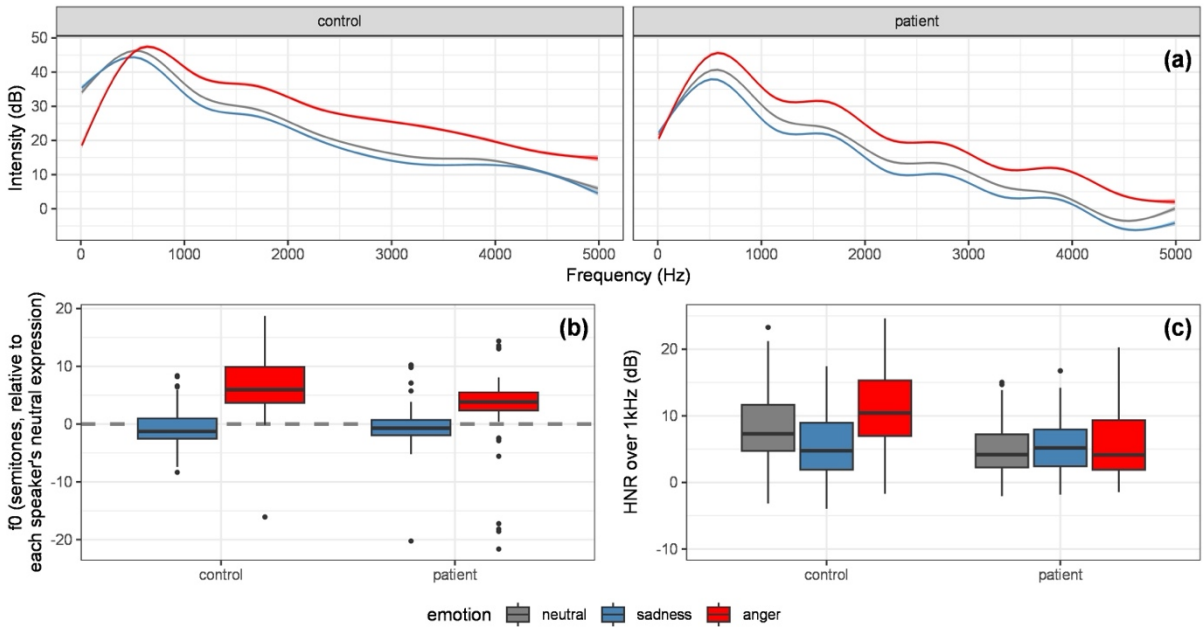


Figure 38 : Comparaison acoustique de la réalisation de /a/ en condition neutre (gris) ou d'expression de la tristesse (bleu) ou de la colère (rouge), pour les témoins et les patients atteints de paralysie laryngée unilatérale : (a) spectres moyens de 0 à 5kHz ; (b) décalage de f0 en demi-tons relativement à l'expression neutre produite par le même locuteur, la ligne horizontale pointillées matérialisant le niveau 0 qui correspond à une absence de différence par rapport à l'expression neutre ; (c) rapport entre énergie harmoniques et bruit (HNR) calculé après filtrage passe-haut pour ne retenir que les fréquences supérieures à 1kHz. D'après Petrone et al. (2024, [AFF1]).

L'inspection des spectres moyens suggère que, tandis que les expressions de colère des témoins sont associées à une forme spectrale distincte et à une intensité plus élevée que la tristesse et l'expression neutre, les expressions de colère par les patients sont caractérisées par une augmentation de l'intensité seulement. L'analyse statistique des corrélations entre spectres ont confirmé que la différence de forme spectrale entre neutre et tristesse est comparable entre les témoins et les patients, mais que la différence entre neutre et colère est plus importante pour les témoins que pour les patients. Bien qu'elle soit limitée ici à la voyelle /a/, l'analyse de la fréquence fondamentale indique un écart entre expression neutre et expression de colère plus important chez les locuteurs témoins que chez les patients, certains patients produisant même une fréquence fondamentale plus basse pour l'expression de la colère que pour l'expression neutre alors que cela n'est observé chez les témoins que pour un locuteur et une phrase. Enfin la comparaison des mesures de HNR indique une structure

harmonique moins riche chez les patients, sans différences entre émotions, ce qui suggère une moindre modulation de la qualité de voix en fonction de l'émotion exprimée chez les patients. Ainsi, ces résultats tendent à confirmer le postulat selon lequel la PLU impacte négativement la capacité des patients à transmettre leurs émotions dans la voix et la parole, et suggèrent que cela pourrait être lié à leur moindre capacité de modulation de leur fréquence fondamentale et de leur qualité de voix.

5.2.3 Voix de substitution suite à une laryngectomie partielle

Publications et communications associées :

[ACTI35] Crevier Buchman, L., Roques, E., & **Audibert, N.** (2015). Retrospective longitudinal acoustic and perceptive study of substitution voice after partial laryngectomy. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS'15)*, Glasgow, Royaume-Uni, paper 1036 (actes en ligne).

[COM5] Crevier Buchman, L., Roques, E., & **Audibert, N.** (2015). *Retrospective longitudinal acoustic and perceptive study of substitution voice after partial laryngectomy*. 3rd Congress of European ORL-HNS, Prague, République Tchèque.

[COM6] Roques, E., **Audibert, N.**, & Crevier Buchman, L. (2014). *Étude rétrospective longitudinale acoustique et perceptive des voix de substitution après laryngectomie partielle*. 70ème Congrès de la Société Française de Phoniatrie et des Pathologies de la Communication, Paris, France.

Dans le cadre du mémoire d'orthophonie d'Emeline Roques que j'ai coencadré avec Lise Crevier Buchman et des prolongements du travail réalisé dans ce mémoire, nous avons procédé à une étude longitudinale des voix de substitution utilisées par des patients ayant subi une laryngectomie partielle, afin de caractériser la façon dont ces voix de substitution se mettent en place au cours du temps. Trois grands types de laryngectomies ont été considérées. Dans la laryngectomie fronto-latérale (FL), technique chirurgicale dite verticale, un pli vocal est retiré ainsi que la commissure antérieure et la partie antérieure du pli vocal controlatéral. Dans le cas des laryngectomies partielles supracricoïdiennes qui sont des techniques horizontales, la glotte est entièrement retirée et reconstruite par suture de l'os hyoïde, soit à l'épiglotte dans le cas de la reconstruction par cricohyoïdo-épiglottopexie (CHEP), soit à la racine de la langue dans le cas de la reconstruction par cricohyoïdopexie (CHP). La voix de substitution est produite par le rapprochement entre les cartilages aryténoïdes et la structure préservée sur la face antérieure qui dépend du type de laryngectomie horizontale. Dans ces deux types de laryngectomie horizontale, l'un des aryténoïdes peut être également retiré lorsque l'état du patient l'exige, le type de laryngectomie étant alors noté CHEP1 ou CHP1 par opposition à CHEP2 ou CHP2 lorsque les deux aryténoïdes sont conservés. L'hypothèse à l'origine de ce travail est que les laryngectomies verticales, dans lesquelles le plan antéro-postérieur du vibrateur laryngé est préservé, aboutiraient à un type de voix dite « glottique » plus comparable à celle des locuteurs normophoniques et à plus forte raison des voix dysphoniques, tandis que les laryngectomies horizontales aboutiraient à une voix dite « laryngale aglottique » aux propriétés perceptives et acoustiques plus éloignées des voix normophoniques.

Les productions de 30 hommes âgés de 59 à 78 ans, répartis en trois groupes de 10 locuteurs pour chacun de ces types de laryngectomie (les types CHEP et CHP étant eux-mêmes

réparties entre chirurgies retirant ou non un aryténoïde), ont été enregistrées 3, 6 et 12 mois après avoir subi une laryngectomie partielle et ont été comparées aux productions de 15 locuteurs témoins âgés de 57 à 76 ans. Les tâches de production ont consisté en la lecture du conte « Tic-tac » et la production isolée des cinq voyelles /i, e, a, o, u/. La motivation pour l'inclusion de multiples qualités vocaliques a été la modification de la géométrie du conduit vocal qui est l'une des conséquences de ce type de chirurgie et est donc susceptible d'affecter ses fréquences de résonances, notamment dans le cas des laryngectomies horizontales qui entraînent un raccourcissement du conduit vocal (voir par exemple Schindler et al. (2005)).

L'évaluation perceptive a été effectuée sur deux échelles par cinq auditeurs à partir d'extraits de parole en condition de lecture : l'échelle GRB, simplification couramment utilisée dans les études cliniques de l'échelle GRBAS (Hirano, 1981) dédiée à l'évaluation des dysphonies, et l'échelle IINFVo (Moerman et al., 2005) spécifiquement conçue pour l'évaluation des voix de substitution. L'analyse du grade dans la cotation GRB et de l'impression générale dans la cotation IINFVo a indiqué que les voix sont significativement altérées suite aux différents types de laryngectomies, y compris dans le cas d'une laryngectomie fronto-latérale mais dans une moindre mesure pour ces laryngectomies FL, et que parmi les laryngectomies horizontales l'altération est moindre lorsque les deux aryténoïdes sont conservés. Si tous les types de laryngectomies entraînent des voix jugées plus rauques, plus soufflées, moins intelligibles, plus bruitées, moins fluentes et avec une moins bonne réalisation du voisement en comparaison au groupe témoin, les voix après laryngectomie FL sont jugées moins altérées sur l'ensemble de ces dimensions que celles après laryngectomie horizontale CHP ou CHEP. Par ailleurs pour la plupart de ces dimensions perceptives, l'altération perçue est plus importante lorsque l'un des aryténoïdes est retiré.

Au-delà de la comparaison entre groupes de locuteurs, le suivi longitudinal des patients a révélé une évolution entre temps post-opératoires sur certaines de ce dimensions perceptives, avec notamment une diminution du souffle perçu entre 3 et 6 mois post-opératoire chez les locuteurs ayant subi une laryngectomie horizontale, et pour tous les types de laryngectomies une amélioration de la fluence entre 3 et 6 mois, de l'impression d'intelligibilité entre 6 et 12 mois, et de la réalisation du voisement entre 3 et 6 mois et entre 6 et 12 mois.

En complément, cinq auditeurs francophones natifs ont passé un test d'identification en choix forcé des voyelles produites par les patients aux trois temps post-opératoires ainsi que des voyelles produites par les locuteurs témoins. Alors que les voyelles produites par les témoins sont très bien identifiées, pour tous les types de laryngectomies des confusions importantes ont été observées entre les paires /i, e/ et /u, o/, la voyelle /a/ restant bien identifiée. Tandis que dans les laryngectomies de type CHEP les confusions se réduisent entre temps opératoires, à l'inverse ces confusions s'accroissent au cours du temps pour les laryngectomies FL et dans une moindre mesure pour CHP1, l'évolution des confusions observées pour CHP2 étant dépendante de la catégorie vocalique.

Afin de mieux interpréter les confusions observées et leur évolution, les résultats de l'analyse perceptive ont été croisés avec l'analyse des fréquences des deux premiers formants des voyelles produites par les patients, via l'inspection des espaces vocaliques dans le plan F1*F2. La Figure 39 illustre les centroïdes de chacune des cinq catégories de voyelles à chacun des trois temps post-opératoires dans le cas des laryngectomies FL pour lesquelles une augmentation des confusions entre 3 et 6 mois est observée. Si les réalisations de /i/ et de /a/ restent stables, on peut observer une importante fluctuation des réalisations de /e/ et surtout

de la paire /u, o/. Plus généralement, les mesures formantiques ont corroboré les résultats du test d'identification des voyelles. Par ailleurs une augmentation des fréquences de F2 attribuable au raccourcissement du conduit vocal (Laccourreye et al., 1995) et dans une moindre mesure de F1 a été observée dans tous les types de laryngectomies, sans incidence notable sur les confusions perceptives.

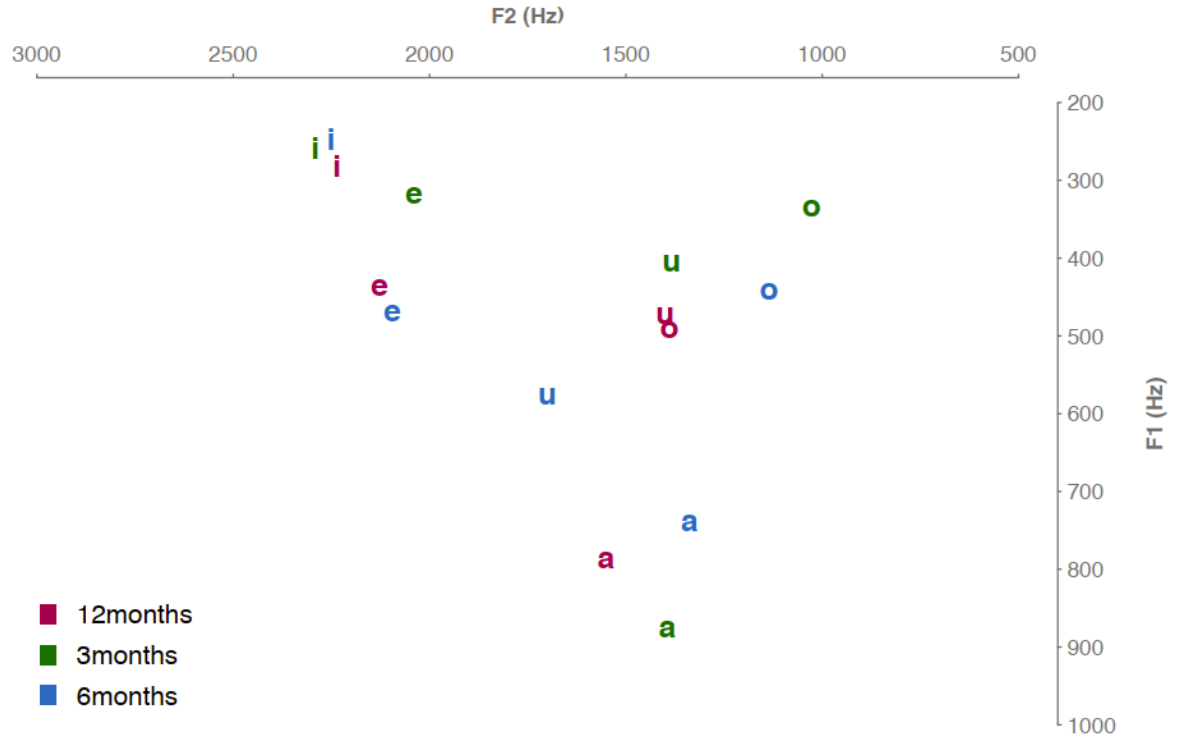


Figure 39 : Réalisation moyenne des voyelles /i, e, a, o, u/ dans le plan F1*F2 par les 10 hommes ayant subi une laryngectomie partielle fronto-latérale, enregistrés 3, 6 et 12 mois après l'opération. D'après Crevier Buchman et al. (2015, [ACTI35]).

En complément, l'analyse du spectre voisé à long de terme de l'enregistrement de la lecture de texte a montré un pic d'énergie acoustique autour de 6000 Hz pour l'ensemble des types de laryngectomies comparées aux productions des patients, avec par ailleurs une réduction de l'énergie en moyennes fréquences (et donc une augmentation de la pente spectrale) entre 3 et 6 mois pour la plupart des types de laryngectomies pris en compte.

Cette étude perceptive et acoustique des voix de substitution suite à différents types de laryngectomies a donc confirmé que bien que la raucité persiste plus que le souffle, après une période de stabilisation et d'habituation les laryngectomies fronto-latérales aboutissent à une voix similaire à celles observées dans le cas des dysphonies notamment sur le plan de l'amélioration de la réalisation du voisement. Les laryngectomies horizontales dans lesquelles les structures laryngées sont modifiées de façon plus radicale par la chirurgie restent quant à elles associées à un degré de raucité et de souffle plus important un an après l'opération en dépit d'une amélioration entre temps post-opératoires. En revanche dans les différents types de chirurgie étudiés, le timbre des voyelles est altéré et entraîne des confusions entre catégories vocaliques, sans que ces confusions puissent être directement expliquées par le raccourcissement du conduit vocal induit par la chirurgie.

5.3 Vieillissement et (co)articulation

Face au vieillissement de la population, la documentation et la meilleure compréhension des conséquences de l'âge sur la voix et la parole adulte constituent des enjeux sociétaux importants. En phonétique clinique notamment, l'étude de pathologies de la voix ou de la parole dont la prévalence est importante chez des sujets âgés nécessite de mieux délimiter la frontière entre vieillissement typique et vieillissement pathologique. D'un point de vue théorique, appréhender plus finement les causes et conséquences du vieillissement de la parole pourrait fournir des clés pour mieux comprendre le système de production de la parole en général.

Bien qu'ils ne permettent pas d'expliquer toutes les caractéristiques observées dans la parole produite par des locuteurs âgés, une partie des effets de l'âge sur la voix et la parole sont liés aux changements physiologiques inhérents au vieillissement, avec l'affaiblissement des muscles dont ceux du conduit vocal. Le vieillissement a également un effet sur la densité des tissus et provoque une érosion des articulations et la calcification des cartilages, ce qui réduit la souplesse du larynx avec des conséquences sur les caractéristiques de la voix des locuteurs (Sataloff & Kost, 2020a, 2020b). Les capacités respiratoires se réduisent également avec l'âge, avec l'affaiblissement du diaphragme et une limitation des capacités d'expansion et de contraction de la cavité thoracique, considérée comme à l'origine de la production de phrases plus courtes et de l'augmentation des pauses dans la parole (Hoit & Hixon, 1987). De plus, le vieillissement impacte également le contrôle et la coordination des mouvements, la conséquence la plus notable étant le ralentissement à la fois de l'activation et de l'exécution de ces mouvements (Seidler et al., 2002), qui dans la parole se traduit par un débit plus lent de 20 à 25% chez les personnes âgées que dans les productions de jeunes adultes (Amerman & Parnell, 1992; Fougeron et al., 2021). Toutefois ce ralentissement de la production de parole n'est pas uniforme, ainsi Hermes et al. (2018) ont mis en évidence à partir de données articulatoires une asymétrie de la vélocité des mouvements de la langue chez les personnes âgées pendant la production de consonnes, avec des phrases de décélération plus longues.

Au-delà de cet effet de l'âge sur le ralentissement du débit de parole, documenté de façon robuste dans la littérature, une baisse de la fréquence fondamentale a été observée chez les femmes âgées, la tendance inverse étant observée chez les hommes (Ramig & Ringel, 1983; Baken, 2005). Le vieillissement est également associé à des délais d'établissement du voisement (VOT) plus courts (Benjamin, 1982; Ryalls et al., 1997). En ce qui concerne la réalisation des voyelles, diverses études mentionnent une baisse des fréquences formantiques mais seule la baisse de la fréquence du premier formant semble être reproductible (Kent & Vorperian, 2018). Une explication possible des changements de fréquences formantiques avec l'âge pourrait être l'allongement du conduit vocal des locuteurs âgés (Linville & Rens, 2001), d'autres auteurs évoquant plutôt une réduction des mouvements articulatoires (Hermes et al., 2018). Par ailleurs si certains auteurs tels que Rastatter et al. (2009) font état d'un espace vocalique plus centralisé chez les locuteurs âgés, les résultats de la littérature sur les liens entre âge et espace vocalique sont peu consistants, d'autant que la majorité des études ayant considéré la réalisation des voyelles se sont concentrées sur l'analyse de formants individuels et non sur le système vocalique considéré dans son ensemble.

L'âge auquel surviennent les changements dans la voix et la parole est difficile à identifier d'après les données de la littérature, d'autant que de nombreuses études ont abordé la question des corrélats du vieillissement à travers la comparaison entre groupes d'âges, avec

une définition des groupes variable et souvent dépendante des âges des locuteurs représentés dans les données analysées. Dans une étude récente dans laquelle l'âge chronologique de 500 locuteurs âgés de 20 à 93 ans a été mis en relation avec un ensemble de mesures acoustiques, Fougeron et al. (2021) ont mis en évidence une variation entre dimensions considérées de la voix et de la parole dans les patrons d'évolution avec l'âge observés, avec des dimensions comme le débit qui évoluent de façon relativement continue tandis que d'autres et notamment celles liées à la voix subissent une inflexion à certain âge, qui peut être supérieur à 75 ans pour des dimensions comme la modulation de la fréquence fondamentale chez les femmes, mais aussi dès la quarantaine ou la cinquantaine pour l'évolution du degré de périodicité de la voix.

Peu d'études ont été consacrées aux liens entre âge et coarticulation. Du fait du ralentissement du débit observé de façon récurrente dans les productions de locuteurs, on pourrait s'attendre à observer moins de chevauchement entre gestes articulatoires correspondant aux segments produits successivement et donc moins de coarticulation chez les locuteurs âgés qui disposent de plus de temps pour atteindre les cibles articulatoires que chez les locuteurs plus jeunes dont le débit est plus rapide. Toutefois, une étude de D'Alessandro & Fougeron (2021) a montré que le lien entre débit de parole et coarticulation anticipatoire de voyelle à voyelle est moins systématique que cela, et surtout que ce lien s'estompe avec le vieillissement : en effet si ce lien est fort chez les locuteurs jeunes et d'âge moyen avec une coarticulation d'autant plus importante que le débit est rapide, le débit des locuteurs de plus de 70 ans ne permet plus d'expliquer les différences de degré de coarticulation observées, qui relèveraient plutôt d'une réorganisation de la parole.

5.3.1 Vieillissement et espace vocalique

Publication associée :

[ACTI10] Hermes, A., **Audibert, N.**, & Bourbon, A. (2023). Age-related vowel variation in French. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic. pp. 2045-2049.

Dans une étude en collaboration avec Anne Hermes pour laquelle nous nous sommes appuyés sur des données recueillies par Angelina Bourbon, nous avons cherché à caractériser l'effet du vieillissement sur l'espace vocalique et son organisation ainsi que sur les interactions entre catégories de voyelles, en prenant en compte aussi largement que possible le système vocalique du français. Pour cela, nous nous sommes appuyés sur les enregistrements de 37 locuteurs francophones natifs âgés de 23 à 90 ans (20 femmes âgées de 62,5 ans en moyenne et 17 hommes âgés de 56,7 ans en moyenne) sans trouble cognitif diagnostiqué. Ces enregistrements sont issus de l'étude de Bourbon & Hermes (2020), et consistent en la lecture d'un ensemble de phrases conçues pour faire varier la longueur et la complexité syntaxique de la phrase, inspirées de Fuchs et al. (2013) et adaptées au français. Un total de 12 phrases a été retenu, chacune répétée trois fois par chaque locuteur. Suite à une étape de segmentation automatique en phones et de vérification manuelle et à l'élimination des voyelles accentuées en position finale, les voyelles orales /i, y, e, ɛ, a, o/ ainsi que les voyelles nasales /ɔ̃, ɑ̃/ pour lesquelles le corpus comprenait un minimum de 14 exemplaires par catégorie vocalique et par locuteur ont été sélectionnées, pour un total de 15 375 voyelles incluses dans l'analyse.

En raison des difficultés que pose l'analyse formantique sur les voyelles nasales et afin de ne pas limiter l'analyse aux voyelles orale, la réalisation acoustique des voyelles a été décrite par douze coefficients cepstraux MFCC (voir section 7.2.3.3 pour des précisions sur cette approche appliquée à la description de l'espace vocalique), le coefficient 0 correspondant au niveau global d'énergie étant exclu de l'analyse, extraits sur une trame de 15 ms centrée sur le milieu de la voyelle. Les métriques DistCentroid, VDispersion et ContrastLoss (voir section 7.2.3.2) ont été calculées pour chaque locuteur dans l'espace à douze dimensions de ces coefficients MFCC afin de caractériser au niveau de chaque exemplaire de voyelle respectivement le degré de dispersion par rapport au centre de l'espace vocalique du locuteur, la variabilité au sein de chaque catégorie vocalique, et le degré de recouvrement entre catégories.

Ces métriques ainsi que les durées segmentales ont ensuite été mises en relation avec l'âge des locuteurs, afin d'évaluer dans quelle mesure les caractéristiques de l'espace vocalique et les relations entre catégories vocaliques dépendent de l'âge. En raison des résultats divergents dans la littérature concernant l'effet de l'âge sur la parole produite par les hommes et les femmes, les jeux de données correspondants ont été analysés séparément. De plus, en complément de l'analyse des données sur l'ensemble des locuteurs, nous avons analysé séparément l'évolution de ces mesures au sein du sous-groupe constitué des locuteurs âgés de 60 ans ou plus.

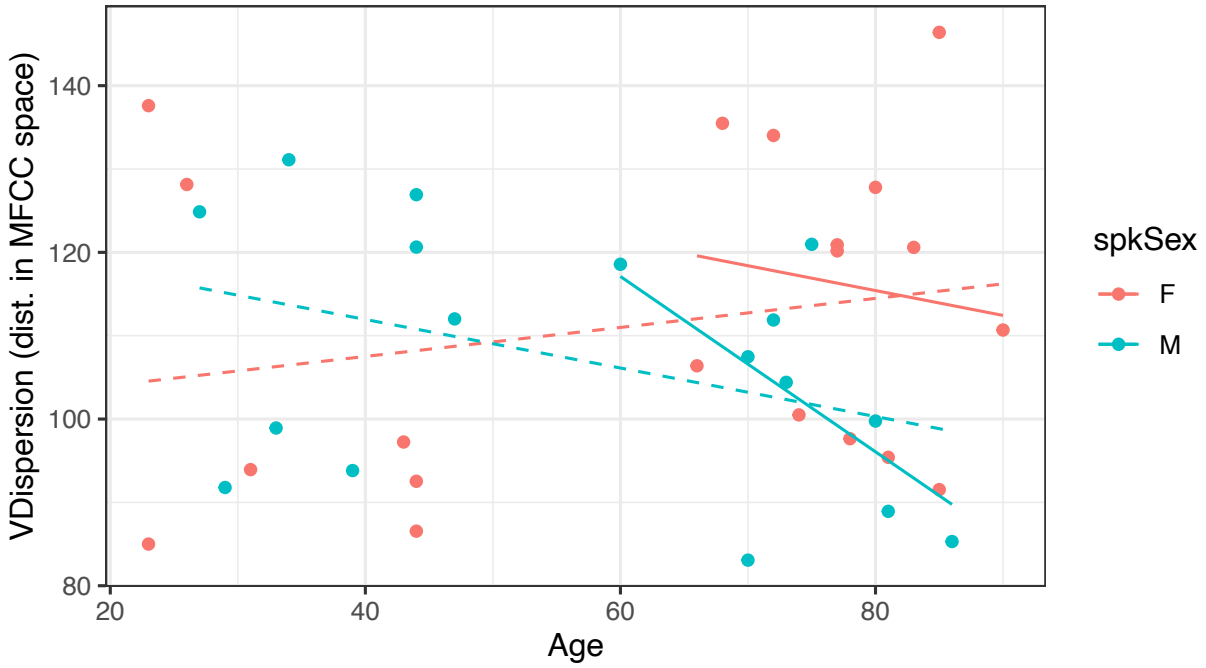


Figure 40 : Lien entre âge des locuteurs et dispersion moyenne au sein de chaque catégorie vocalique, estimée par la métrique VDispersion calculée dans l'espace à 12 dimensions des coefficients MFCC, séparément pour hommes et femmes. Les lignes pointillées correspondent à la régression linéaire sur l'ensemble des locuteurs, et celles en trait continu à la régression linéaire sur les locuteurs âgés de 60 ans ou plus. D'après Hermes et al. (2023, [ACTI10]).

Outre la confirmation du ralentissement du débit avec l'âge à travers l'augmentation des durées segmentales à la fois en considérant l'ensemble des locuteurs et dans une moindre mesure dans le groupe des locuteurs âgés, les résultats n'ont pas montré d'effet de l'âge sur la dispersion de l'espace vocalique, hormis une tendance modérée à la centralisation chez les

hommes les plus âgés (r = -.23). Comme illustré par la Figure 40, la mesure VDispersion de variabilité au sein de chaque catégorie vocalique a montré des tendances divergentes entre les femmes pour lesquelles cette variabilité tend à augmenter modérément avec l'âge (r = .21) et les hommes pour lesquels elle tend à diminuer (r = -.39). Au sein du groupe des locuteurs les plus âgés, cette variabilité intra-catégorie montre une tendance plus marquée à la diminution avec l'âge chez les hommes (r = -.57), tandis qu'elle est beaucoup plus faible chez les femmes.

Aucun effet clair de l'âge sur les confusions entre catégories estimées par la métrique ContrastLoss n'a été relevé pour les femmes, ni en considérant l'ensemble des locutrices ni en se recentrant sur celles les plus âgées. En revanche pour les hommes le recouvrement entre catégories vocaliques diminue avec l'âge, à la fois en considérant l'ensemble des locuteurs (r = -.34) et en ne considérant que les plus âgés (r = .30). En complément des valeurs de la métrique ContrastLoss qui considère de façon globale l'ensemble des recouvrements entre catégories vocaliques au sens de l'analyse linéaire discriminante (LDA), nous avons également examiné les matrices de confusions issues des catégories prédites par la LDA pour chaque locuteur et chaque catégorie de voyelle, ce qui nous a permis de remarquer un effet de l'âge sur les confusions entre voyelles orales et nasales.

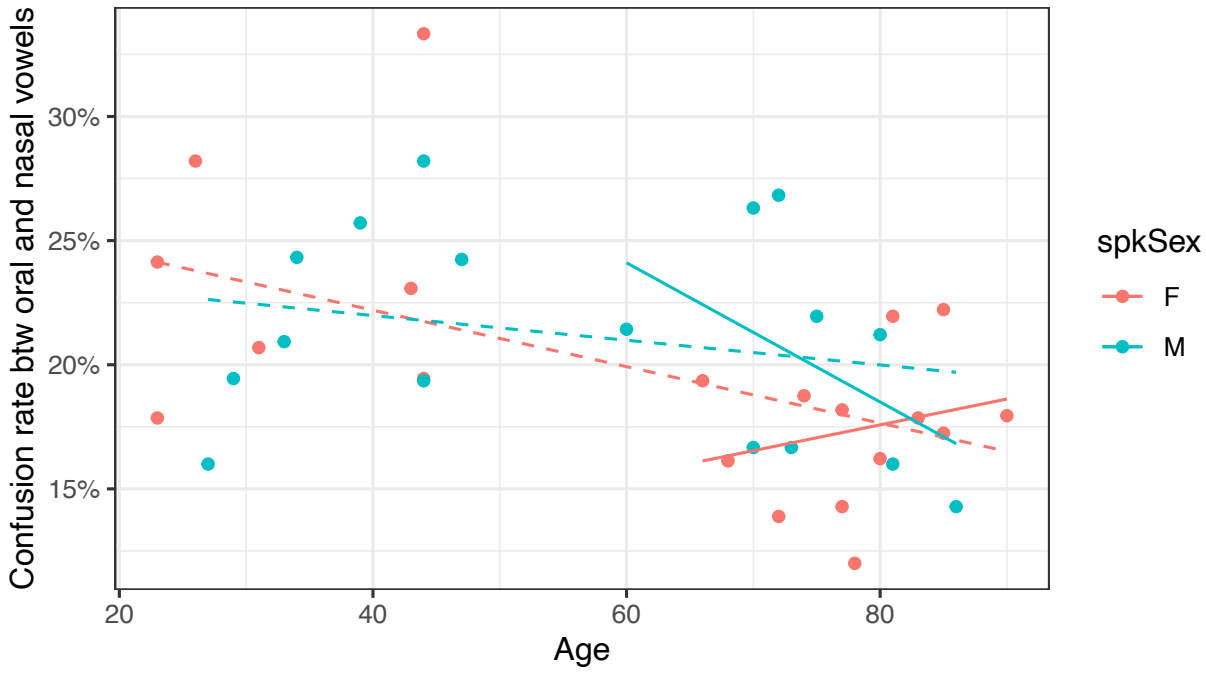


Figure 41 : Lien entre âge des locuteurs et taux de confusion entre voyelles orales et nasales d'après la classification effectuée par analyse linéaire discriminante entre catégories vocaliques sur les productions de chaque locuteur, représenté séparément pour hommes et femmes. Les lignes pointillées correspondent à la régression linéaire sur l'ensemble des locuteurs, et celles en trait continu à la régression linéaire sur les locuteurs âgés de 60 ans ou plus. D'après Hermes et al. (2023, [ACTI10]).

La Figure 41 illustre le lien entre l'âge des locuteurs et le taux cumulé de confusion entre voyelles orales et voyelles nasales. En considérant l'ensemble des locuteurs on observe une tendance à une plus grande distinction entre orales et nasales avec l'âge qui se traduit par une baisse graduelle du taux de confusion, particulièrement pour les femmes (r = -.53) et dans une moindre mesure pour les hommes (r = -.24). Cette plus grande distinctivité entre orales et

nasales est toutefois atteinte plus tôt par les femmes, ce qui peut expliquer les patrons divergents observés entre hommes et femmes dans le groupe des locuteurs les plus âgés, avec une corrélation positive modérée chez les femmes (r = .24) qui s'explique par le faible taux de confusion déjà atteint par les locutrices âgées de 65 à 80 ans, tandis que la tendance à une plus grande distinction entre orales et nasales avec l'âge est confirmée chez les hommes âgés (r = -.47).

Ainsi, les résultats que nous avons obtenus dans cette étude suggèrent que, de même que d'autres dimensions de la voix et de la parole, la production des voyelles et la structure du système vocalique qui en découle pourraient évoluer avec l'âge différemment entre hommes et femmes. Par ailleurs l'évolution des confusions observées entre voyelles orales et nasales soulève des pistes intéressantes à approfondir, éventuellement via la comparaison avec d'autres langues dans lesquelles la distinction entre voyelles orales et nasales est phonologique afin d'évaluer dans quelle mesure ce phénomène est physiologique ou générationnel.

5.3.2 Vieillissement et coarticulation

Publication et communication associées :

[ACTI6] Wohmann-Bruzzo, L., Fougeron, C., & **Audibert, N.** (2024). Effet du vieillissement sur l'anticipation d'arrondissement intra-syllabique en français. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 322-331.

[COM1] Wohmann-Bruzzo, L., **Audibert, N.**, & Fougeron C. (2024). Age effect on intra-syllabic anticipatory labialization. *13th International Seminar on Speech Production (ISSP)*, Autrans, France.

Dans une étude qui s'inscrit dans le cadre de la thèse de Louise Wohmann-Bruzzo que je codirige avec Cécile Fougeron depuis 2023, nous avons étudié l'effet de l'âge sur l'anticipation d'arrondissement dans la consonne /s/ produite avant un /y/. Nous avons pour cela adopté une méthodologique similaire à celle employée dans l'étude menée auparavant sur la variabilité interindividuelle de la coarticulation labiale et présentée en section 2.3.2, en comparant pour chaque locuteur les réalisations spectrales de /s/ en contexte droit /y/ ou /i/. Pour cette étude centrée sur l'effet du vieillissement, nous nous sommes appuyés sur les productions de 20 jeunes adultes (23 à 34 ans, 10 femmes et 10 hommes) et de 20 locuteurs âgés (72 à 86 ans, 11 femmes et 9 hommes) lisant trois fois le même texte. Le degré de coarticulation labiale et d'anticipation de la coarticulation ont été évalués via l'évolution de l'information spectrale entre le début et la fin de la réalisation du /s/, en comparant pour chaque locuteur la production de 12 occurrences de la syllabe /sy/ à 12 occurrences de la syllabe /si/ utilisée comme référence pour le locuteur considéré, soit un total de 960 occurrences de /s/ analysées. Les syllabes cibles /sy/ et /si/ ont été sélectionnées dans des mots monosyllabiques présents dans le texte lu afin de faire en sorte que le /s/ et la voyelle appartiennent toujours à la même syllabe.

Bien que l'effet d'abaissement du CoG de /s/ par l'arrondissement ait été confirmé par Koenig et al. (2013), ces derniers ont également souligné que l'abaissement du CoG peut être lié à la fois à l'allongement de la cavité antérieur et au placement plus postérieur de l'apex, ce qui les a conduits à proposer une nouvelle mesure acoustique, FreqM, définie comme la fréquence du pic d'énergie en moyennes fréquences (de 3000 Hz à 7000 Hz pour les hommes

et 8000 Hz pour les femmes) destinée à capturer plus spécifiquement la résonance de la cavité antérieure à la constriction. Leurs travaux sur les fricatives sibilantes ainsi que ceux de Shadle et al. (2023) ont confirmé que les valeurs de la mesure FreqM diminuent bien comme attendu en contexte arrondi, de façon dynamique avec une baisse de fréquence qui est maximale à la fin de la production de la fricative. Nous avons donc retenu cette mesure comme estimation du degré de coarticulation labiale anticipatoire, en considérant son évolution entre le début et la fin du /s/ afin de pouvoir caractériser le décours temporel de la coarticulation.

Après segmentation manuelle de l'ensemble des /s/ sélectionnés, douze points équidistants répartis entre le début et la fin de chaque /s/ ont été définis. Le spectre FFT a été calculé sur une trame de 25 ms découpée à l'aide d'une fenêtre de Hanning autour de chacun de ces points à l'exclusion du premier et du dernier de chaque /s/. Suite à un lissage cepstral selon la méthode utilisée par Al-Tamimi & Khattab (2015) la valeur de FreqM a été extraite de chaque trame. En complément, le débit articulatoire a été estimé pour chaque locuteur à partir de l'une des phrases lues comprenant 29 phones, en divisant la durée totale de production de cette phrase après exclusion des pauses par le nombre de phones attendu. Les productions des hommes et des femmes ont été analysées séparément en raison des différences de réalisation des consonnes fricatives et notamment du /s/ en fonction du sexe relevées dans la littérature (voir par exemple Stuart-Smith et al. (2007)).

Les résultats établis à partir de modèles de régression linéaires mixtes prenant en compte la durée du /s/ comme facteur aléatoire ont confirmé une baisse de la valeur de FreqM dans les /s/ produits en contexte arrondi aussi bien chez les femmes que les hommes, mais avec un degré de coarticulation estimé par l'écart entre /sy/ et /si/ des valeurs de FreqM moindre chez les locuteurs âgés comparativement aux locuteurs plus jeunes, tout particulièrement chez les femmes âgées pour lesquelles l'écart entre /sy/ et /si/ est non-significatif. Comme attendu, on remarque également que la durée de /s/ est plus importante chez les locuteurs âgés des deux sexes, en lien avec l'augmentation du débit avec l'âge. En revanche si les locuteurs jeunes qui produisent des /s/ plus courts coarticulent plus au sens de l'écart spectral entre leurs /sy/ et leurs /si/ ($r = .53$), ce lien est beaucoup plus faible chez les locuteurs âgés, de façon consistante avec les résultats obtenus par D'Alessandro & Fougeron (2021) sur les liens entre âge et coarticulation voyelle-à-voyelle.

L'analyse individuelle, illustrée par la Figure 42, indique que les femmes sont plus nombreuses que les hommes à peu coarticuler, particulièrement dans le groupe des locutrices âgées dans lequel on trouve 9 locutrices pour lesquelles la fréquence du pic en moyennes fréquences n'est pas significativement différente entre /sy/ et /si/, contre seulement trois locuteurs plus jeunes. Chez les hommes, seul un locuteur âgé est dans ce cas. En outre, l'observation qualitative des trajectoires individuelles indique une importante variation inter-individuelle, tout particulièrement parmi les femmes et dans une moindre mesure parmi les hommes, en termes de stratégies de coarticulation au niveau temporel avec une anticipation d'arrondissement se produisant plus ou moins tôt dans la fricative qui ne semble pas pouvoir s'expliquer par l'âge, ainsi qu'en termes d'amplitude de l'écart entre /sy/ et /si/. De plus ces différences individuelles ne s'expliquent que très partiellement par les différences de débit de parole entre locuteurs, avec notamment des locuteurs âgés dont certains présentent des valeurs proches de débit mais un degré de coarticulation anticipatoire très différent.

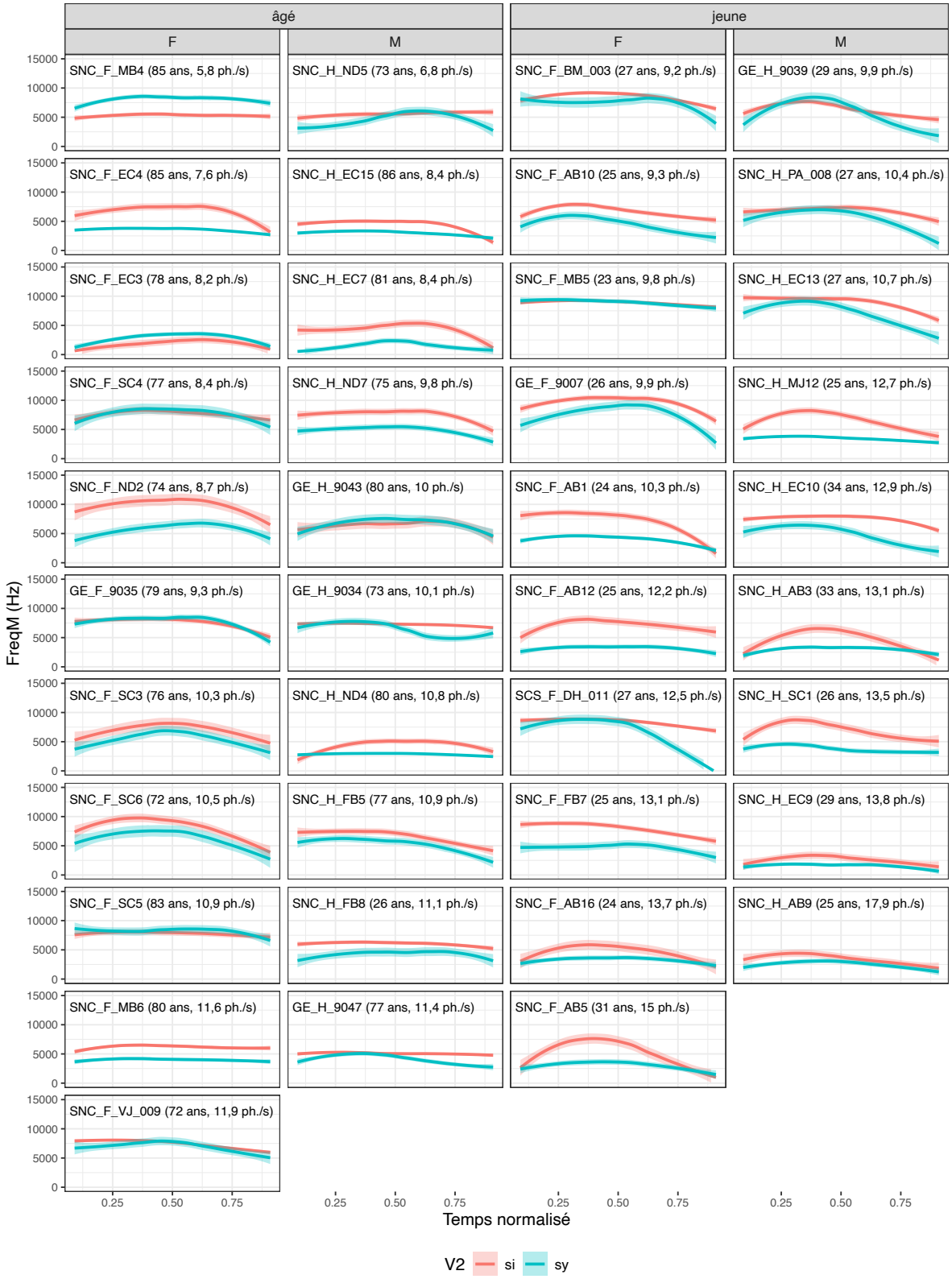


Figure 42 : Trajectoires moyennes (lissées par une régression locale *loess* après normalisation temporelle) des valeurs de la fréquence du maximum d'énergie spectrale en moyennes fréquences FreqM entre le début et la fin de la consonne /s/, comparées en fonction de la voyelle suivante pour chaque locuteur. Les locuteurs âgés sont présentés à gauche et les jeunes à droite, séparément entre hommes et femmes pour chaque groupe d'âge, par ordre croissant de débit de parole dans chaque sous-groupe. Les enveloppes colorées autour des courbes de régression représentent l'erreur-type. D'après Wohmann-Bruzzo et al. (2024, [ACTI6]).

La variabilité observée dans la coarticulation anticipatoire pourrait être liée à la variation individuelle dans la production de la consonne /s/, qui a par exemple conduit Kavanagh (2012) à proposer des mesures de discrimination entre locuteurs se fondant sur la réalisation de cette consonne. Par ailleurs, à moins de postuler une planification à une échelle encore plus réduite que celle de la syllabe chez les locuteurs âgés sains tels que ceux considérés ici, ces résultats qui suggèrent une différence entre groupes d'âge en termes de degré de coarticulation intra-syllabique sont peu compatibles avec l'une des hypothèses émises par D'Alessandro & Fougeron (2021). En effet, ces dernières ont postulé à partir de leurs résultats sur la coarticulation voyelle-à-voyelle que la planification des gestes articulatoires pourrait évoluer avec l'âge et porter chez les locuteurs âgés sur la syllabe et non plus sur des unités plus longues telles que le mot. Toutefois, cela impliquerait alors que la coarticulation intra-syllabique ne changerait pas avec l'âge au-delà des différences liées au débit de parole. Sous réserve que les résultats obtenus dans cette étude ne soit pas liés à un effet générationnel, interprétation toujours possible lorsque des locuteurs de différents âges sont comparés de façon synchrone en l'absence d'informations longitudinales sur l'évolution des mêmes locuteurs au cours du temps, l'explication des différences entre groupes d'âge pourraient être plutôt à chercher dans une évolution plus générale avec l'âge du contrôle moteur de la parole qui impliquerait un chevauchement moins important entre gestes articulatoires chez les locuteurs âgés (D'Alessandro et al., 2020).

5.4 Perception de l'âge à partir de la voix

Publication et communications associée :

[ACTI28] **Audibert, N.**, Fougeron, C., Barbier, F., Croze, L., Lavoine, C., & Rance, H. (2018). Quel est mon âge d'après ma voix ? Effets de la variété régionale et de la génération. *Actes des 32èmes Journées d'Études sur la Parole*, Aix-en-Provence, France, pp. 612-620.

[AFF5] **Audibert, N.**, & Fougeron, C. (2019). Are people as old as they sound? Acoustic, regional and generational effects. *3rd Phonetics and Phonology in Europe conference*, Lecce, Italie.

[AFF6] **Audibert, N.**, Fougeron, C., Brunot, M., Callé, S., Davallet, I., & Lechevalier, N. (2019). Liens entre âge estimé, sévérité de la dysarthrie et intelligibilité des locuteurs. *8èmes Journées de phonétique clinique (JPC8)*, Mons, Belgique.

L'âge d'un locuteur fait partie des multiples informations sociophonétiques qui peuvent être indexées dans la parole (Foulkes & Docherty, 2006; Foulkes et al., 2010; Eckert, 2017). Une question cruciale est de déterminer si l'âge doit être défini en termes chronologiques, cognitifs ou biologiques, ou encore en fonction des représentations de l'âge par les auditeurs ou interlocuteurs. Dans une série d'études communes avec Cécile Fougeron, nous avons abordé la question de la relation entre l'âge chronologique et l'âge perçu, ainsi que les différents facteurs susceptibles d'expliquer comment et pourquoi ils diffèrent, en nous concentrant principalement sur des productions de locuteurs sains de différents âges mais également en nous intéressant aux interactions entre pathologies de la parole et perception de l'âge. Ces travaux se sont appuyés sur le travail réalisé dans le cadre de stages d'orthophonie que nous avons coencadré, ainsi que sur les étapes préparatoires du mémoire d'orthophonie de Mélissa Brunot et Sandrine Callé que nous avons codirigé.

Parmi les facteurs susceptibles d'influencer la perception de l'âge, on peut mentionner les différences de dialecte et de langue : plusieurs études ont en effet mis en évidence un effet du décalage entre origine du locuteur et de l'auditeur dans l'estimation de l'âge (Foulkes et al., 2010; Nagao, 2006), tandis que Braun & Cerrato (1999) ont conclu à une absence d'effet. L'âge des auditeurs peut également influencer l'estimation de celui des locuteurs : Linville & Korabic (1986) ont ainsi montré que des auditrices âgés sont moins précises que des locutrices plus jeunes pour estimer l'âge à partir de voyelles isolées, ce résultat étant confirmé par Goy et al. (2016) dont les résultats ont également suggéré que cette différence entre groupes d'auditeurs serait spécifique à l'estimation de l'âge, les auditeurs âgés étant par ailleurs plus performants que les plus jeunes dans leur estimation du sexe des locuteurs. Sur le plan des caractéristiques acoustiques des productions, les résultats de Harnsberger et al. (2008) suggèrent que l'estimation de l'âge serait principalement influencée par le ralentissement du débit de parole, et dans une moindre mesure par le registre de fréquence fondamentale des locuteurs.

Dans une première étude nous avons comparé l'âge chronologique et l'âge perçu dans les productions de 112 locuteurs du français âgés de 50 à 89 ans extraites de la base MonPaGe_HA (Fougeron et al., 2018), répartis en quatre variétés régionales (français d'Ile-de-France, belge, suisse, québécois) et quatre décennies (50-59, 60-69, 70-79, 80-89 ans), chaque groupe étant équilibré entre hommes et femmes. Dans une tâche de perception en choix forcé, 13 auditeurs francophones jeunes (22 à 31 ans) et 13 auditeurs francophones âgés (70 à 95 ans) d'Ile-de-France ont estimé l'âge de chaque locuteur en sélectionnant l'une des quatre classes d'âge, à partir de l'écoute d'une phrase lue d'une longueur de dix syllabes.

Dans nos données, l'âge chronologique s'est avéré un prédicteur assez fiable de l'âge perçu d'un locuteur ($r = .75$). Cependant, nous avons remarqué une tendance à surestimer l'âge des locuteurs du groupe 50-59 ans, et à sous-estimer l'âge des locuteurs à partir de 70 ans, confirmant ainsi les résultats d'études précédentes, par exemple Huntley et al. (1987), voir aussi Hunter et al. (2016) pour une revue En d'autres termes, les locuteurs du groupe des 60-69 ans apparaissent comme estimés avec un âge plus proche de leur âge chronologique. Ce résultat doit néanmoins être considéré avec prudence car les réponses tendent à converger vers les catégories centrales dans ce type de tâche. Les différences entre l'âge perçu et l'âge chronologique peuvent s'expliquer par différents facteurs, liés soit au décalage régional et générationnel entre les auditeurs et les locuteurs, soit aux caractéristiques acoustiques des productions des locuteurs.

Nous avons également observé que les estimations de l'âge dépendaient de l'origine partagée ou non entre le locuteur et l'auditeur. En effet, comme illustré par la Figure 43, à âge chronologique égal les locuteurs français sont évalués par les auditeurs français d'Ile-de-France comme étant plus jeunes que les locuteurs d'autres variantes régionales du français. Les liens entre l'âge perçu et l'âge chronologique varient également en fonction des groupes régionaux : pour les locuteurs suisses, l'âge chronologique est estimé de façon plus précise ($r = .87$) que pour les autres groupes (de $r = .72$ pour les locuteurs belges à $r = .78$ pour les locuteurs québécois).

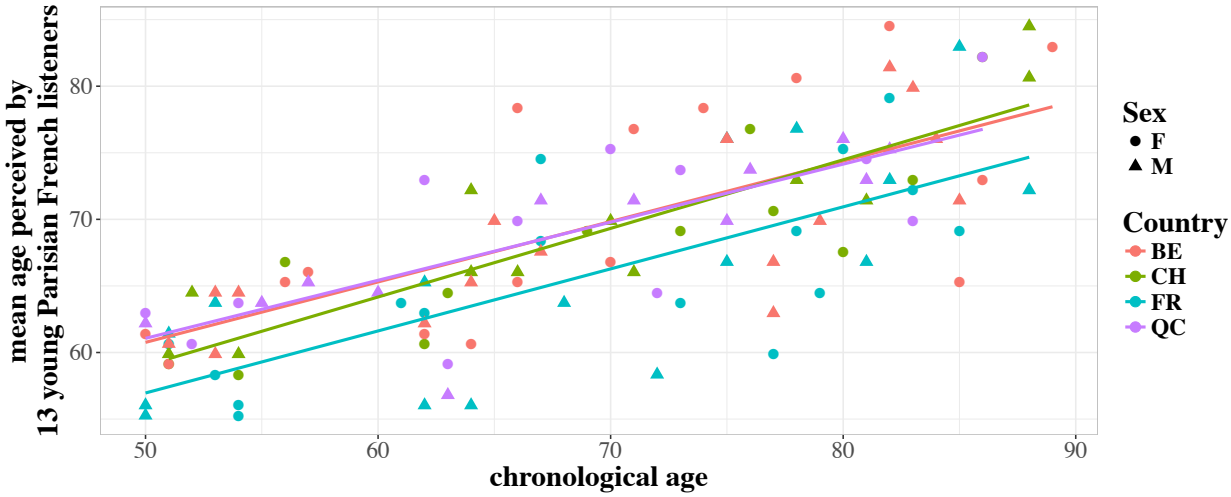


Figure 43 : Lien entre âge chronologique des locuteurs des quatre groupes régionaux (BE = belges ; CH = suisses ; FR = français ; QC = québécois) et âge perçu par les 13 jeunes auditeurs d'Ile-de-France. Les droites représentent la régression linéaire dans chaque groupe, hommes et femmes confondus. D'après Audibert & Fougeron (2019, [AFF5]).

Les estimations de l'âge chronologique dépendent également de la différence d'âge entre les auditeurs et les locuteurs. De manière inattendue, l'âge des auditeurs a plus d'effet pour l'estimation réalisée par les locuteurs les plus jeunes. En effet l'âge des locuteurs de plus de 70 ans est sous-estimé de manière équivalente par les auditeurs jeunes et plus âgés, tandis que les locuteurs de moins de 70 ans sont principalement jugés plus âgés que leur âge chronologique par les auditeurs plus âgés, comme illustré par la Figure 44. Ce résultat ne reflète pas simplement les difficultés des auditeurs plus âgés à réaliser une telle tâche, l'accord intra-juge étant équivalent à celui des jeunes auditeurs. Par ailleurs cette moindre précision des auditeurs âgés pour estimer l'âge des locuteurs de la génération précédente à la leur, ainsi que l'écart globalement plus important entre âge perçu et âge chronologique pour les jeunes auditeurs, suggèrent une difficulté à estimer l'âge de locuteurs d'une autre génération que la sienne.

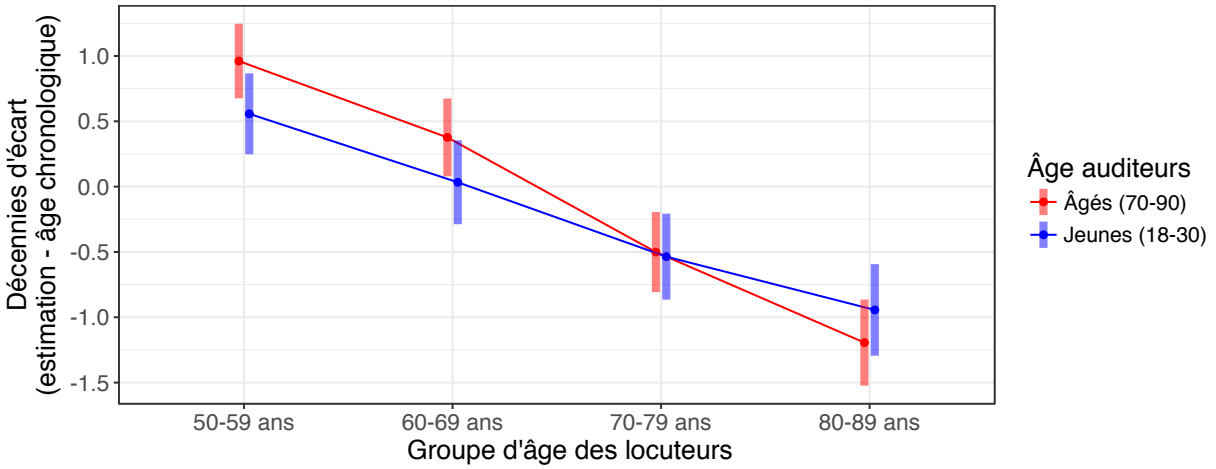


Figure 44 : Valeurs en décennies d'écart entre âge estimé et âge chronologique prédites par un modèle linéaire mixte pour chaque groupe d'âge des locuteurs et chaque groupe d'âge des auditeurs, toutes origines régionales confondues. Les barres verticales représentent les intervalles de confiance. D'après Audibert et al. (2018, [ACTI28]).

Dans nos données, l'estimation de l'âge dépend également des caractéristiques acoustiques des productions des locuteurs. Parmi les quelques indices temporels et spectraux que nous avons évalués, conformément aux résultats de Harnsberger et al. (2008) le débit de parole semble être le meilleur prédicteur de l'âge chronologique du locuteur, bien qu'il n'explique qu'une petite partie de la variation observée. On peut noter que le débit de parole prédit mieux l'âge perçu (r = -.40) que l'âge chronologique (r= -.25).

Dans une étude complémentaire, un total de 726 phrases extraites du protocole d'évaluation de l'intelligibilité de MonPaGe (Lévêque et al., 2016) et produites par 32 locuteurs dysarthriques de 25 à 70 ans et 17 locuteurs témoins âgés (77 à 88 ans) ont été évaluées par 61 auditeurs francophones natifs (33 naïfs et 28 étudiants en orthophonie), répartis en cinq groupes évaluant les productions de différents locuteurs dont quatre communs à tous les groupes. Suite à la présentation des phrases produites par chaque locuteur et donnant lieu à une évaluation de l'intelligibilité via leur transcription, l'âge à la décennie près et la sévérité de la dysarthrie du locuteur étaient évalués par l'auditeur.

Les résultats ont montré une tendance à la surestimation de l'âge des locuteurs dont la dysarthrie est considérée la plus sévère (r = .73) comme illustré par la Figure 45, et dans une moindre mesure des locuteurs les moins intelligibles (r = -0.44).

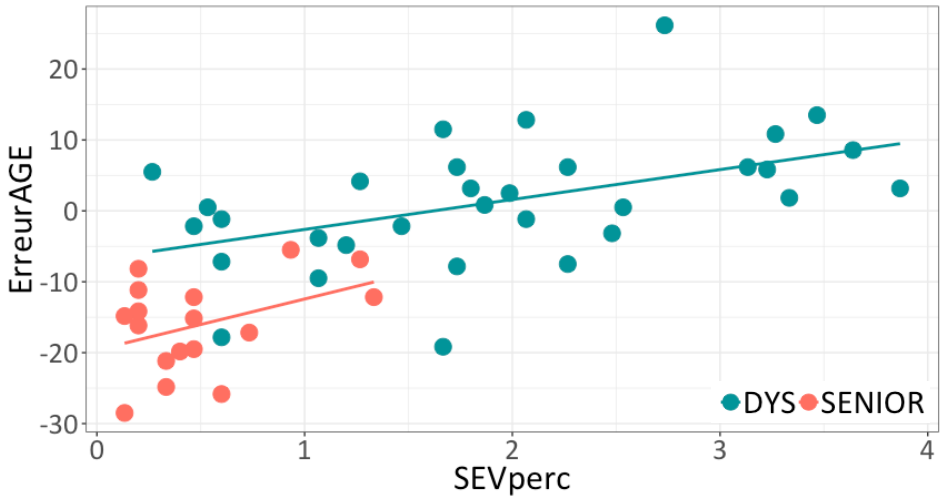


Figure 45 : Lien entre sévérité perçue de la dysarthrie (SEVperc) et décalage entre âge perçu et âge chronologique pour les 32 locuteurs dysarthriques (DYS) et les 17 locuteurs témoins âgés (SENIOR). Les droites représentent la régression linéaire dans chaque groupe de locuteurs. D'après Audibert et al. (2019, [AFF6]).

Globalement, ces travaux ont confirmé que l'estimation de l'âge d'un locuteur à partir de sa parole dépend à la fois des indices acoustiques du signal et de facteurs individuels qui peuvent être interprétés comme liés à la représentation construite par l'auditeur des caractéristiques du locuteur, et ont suggéré une meilleure aptitude à identifier l'âge de ses pairs, à la fois en termes de variante dialectale et de génération, que l'âge de locuteurs plus éloignés de soi. Les résultats soulignent également le manque potentiel d'homogénéité au sein des groupes lorsque les locuteurs sont classés ou catégorisés en fonction de l'âge chronologique plutôt que de l'âge perçu, ce qui peut avoir des implications non seulement pour la constitution de groupes appariés en fonction de l'âge pour l'étude de la parole pathologique, mais aussi pour les études sociophonétiques sur la variabilité de la parole.

6 Autres travaux sur la voix et la parole

Résumé du chapitre 6

Dans ce chapitre je récapitule les autres travaux sur divers aspects de la voix et de la parole auxquels j'ai participé et qui ne relèvent d'aucune des thématiques présentées dans les autres chapitres.

Parmi les études sur l'acquisition en L2, une étude de la qualité de voix au sens de Laver d'apprenants francophones de l'anglais comparée à des natifs de diverses régions britanniques a conclu à un arrondissement plus important et une nasalisation moindre des apprenants. Une nouvelle analyse révèle une variation individuelle importante entre apprenants mais aussi entre natifs au-delà des tendances majoritaires à l'échelle de chaque groupe. Une série d'études de l'acquisition L2 de l'intonation du français et de l'anglais à l'aide d'un outil de synthèse vocale performative a permis de valider la capacité de sujets natifs ou non à reproduire par le geste des modèles intonatifs avec une précision aussi voire plus importante que celle de l'imitation vocale, et une évolution des performances dans le cas de l'application aux modalités du français qui suggère un gain d'apprentissage lié à la synthèse performative.

Dans le cadre de la thèse de D. Yoon, une série d'études sur la variation de la voix et de la parole en fonction de la langue et du sexe des locuteurs a montré des différences de voix entre français et coréen dans l'expression du dimorphisme sexuel, ainsi qu'un lien entre taille des locuteurs et résonances.

Une étude acoustique des réalisations approximantes de /v/ en français conversationnel a montré que ces réalisations sont fréquemment présentes en positions prosodiques fortes chez certains locuteurs, suggérant une possible amorce de changement sonore au-delà d'un simple phénomène de réduction. Une étude sur grands corpus du dévoisement final en français a confirmé un effet de dévoisement avant pause, variable en fonction du lieu pour les fricatives mais sans différences systématiques entre styles de parole. Enfin, une étude sur grands corpus du phénomène de disjonctivité en français dit « h aspiré » a indiqué que la disjonction, qui concerne très majoritairement des mots en « h » et est conditionnée par le lexique mais peut aussi être déclenchée par d'autres facteurs, notamment pragmatiques.

6.1 Acquisition suprasegmentale en langue seconde

6.1.1 Qualité de voix d'apprenants francophones en anglais L2

Publication associée :

[ACTI58] Coadou-Toscano, M. & **Audibert, N.** (2009). Voice quality and English as a Foreign Language: A pilot study. *Proceedings of the 3rd International Workshop on Advanced Voice Functions Assessment (AVFA09)*, Madrid, Espagne, actes CD-ROM.

6.1.1.1 Contexte et principaux résultats

Dans le cadre d'une collaboration en 2009 avec Marion Coadou-Toscano, alors membre du Laboratoire Parole et Langage d'Aix-en-Provence, j'ai été amené me pencher sur la question de la qualité de voix d'apprenants francophones en anglais langue seconde.

Tandis que dans la majorité de mes autres travaux sur la qualité de voix c'est une définition plus étroite et centrée sur la source vocale qui a été retenue, dans ce cadre la qualité de voix est considérée comme relevant d'une définition plus large incluant à la fois configurations laryngées et articulatoires, suivant les propositions de Laver (1980) formalisées à travers le Vocal Profile Analysis scheme (Laver et al., 1981). Le Vocal Profile Analysis scheme, ci-après VPA, propose un protocole dans lequel un juge expert préalablement entraîné à cette tâche évalue séparément en leur attribuant un score de 0 à 6 (les scores de 3 à 6 étant le plus souvent associés à des configurations jugées pathologiques) un ensemble de configurations articulatoires relatives aux lèvres, à la mâchoire, à la langue, au vélopharynx, à la tension laryngée et plus largement à celle du conduit vocal, et enfin au mode de phonation employé. Le score de 0 est supposé être associé à une configuration « neutre », qui pour une application à la parole pathologique pour laquelle le VPA a été initialement développé correspondrait à une production normophonique. Dans le cadre de l'application à la caractérisation d'accents régionaux (voir par exemple Stuart-Smith et al. (1999)), ou dans notre cas à l'acquisition en langue seconde, le VPA vise également à caractériser des postures articulatoires générales plutôt que des configurations spécifiques à l'articulation de certains segments.

Les données analysées dans cette étude consistaient en des enregistrements de parole lue réalisés par la première auteure de locuteurs britanniques représentant cinq régions des îles britanniques et d'apprenants français de l'anglais avec un niveau débutant, avec dix locuteurs (cinq hommes et cinq femmes) pour chacun des six groupes. Pour les besoins de cette étude, le VPA a été recentré sur un nombre plus limité de dimensions jugées pertinentes pour la caractérisation des accents britanniques et des productions des francophones. Ainsi, les dimensions spécifiquement pathologiques telles que la fuite nasale et le tremblement vocal ont été exclues. D'autres paramètres, notamment relatifs à l'utilisation de certains modes de phonation (falsetto, *harsh voice*, voix soufflée) ont été supprimés de l'analyse en raison de l'absence totale dans les données de configurations non-neutres. Les paires de paramètres relatifs à une même dimension ont été recodées en une seule variable via le recours à des valeurs négatives. Les productions des 60 locuteurs ont donc été décrites par un ensemble de neuf variables : arrondissement des lèvres, écartement des lèvres, position de l'apex et de la lame de la langue, fermeture de la mâchoire ou protrusion, nasalité, tension du conduit vocal, tension du larynx, voix craquée et chuchotement.

Une comparaison statistique multivariée entre apprenants francophones et anglophones natifs a conclu à une différence significative entre groupes. En complément, les comparaisons univariées pour chacune des dimensions prises en compte ont mis en évidence un arrondissement labial plus prononcé et une moindre nasalisation de la part des apprenants francophones. Si ce résultat est venu confirmer ceux d'Esling & Wong (1983) sur la comparaison entre les voix de francophones et d'anglophones concernant l'arrondissement, ils vont à l'encontre de leurs résultats sur le plan de la nasalité. Cette contradiction apparente pourrait s'expliquer par la présence de voyelles nasales en français qui conduirait les francophones à moins nasaliser les voyelles non-nasales et à transférer cette caractéristique à l'anglais. Une comparaison plus directe avec l'accent de Cambridge, considéré parmi les accents britanniques représentés dans les données comme le plus proche de celui auquel les apprenants francophones sont exposés en tant que modèle, a indiqué que les francophones se distinguaient par une tension moindre du larynx et du conduit vocal et par un abaissement plus important de la mâchoire.

6.1.1.2 Effets du locuteur

Dans la version de cette étude publiée en 2009 dans les actes du Workshop Advanced Voice Functions, l'analyse était centrée sur la comparaison des moyennes entre groupes au moyen d'une analyse de variance multivariée complétée par une analyse univariée par dimension étudiée. La dimensionnalité des données qui sont constituées d'une valeur par locuteur et par dimension retenue du VFA ne permet pas une analyse statistique inférentielle prenant directement en considération l'effet du locuteur. Toutefois, une analyse descriptive plus détaillée permet d'apporter un éclairage nouveau sur la variation individuelle au sein de ce jeu de données.

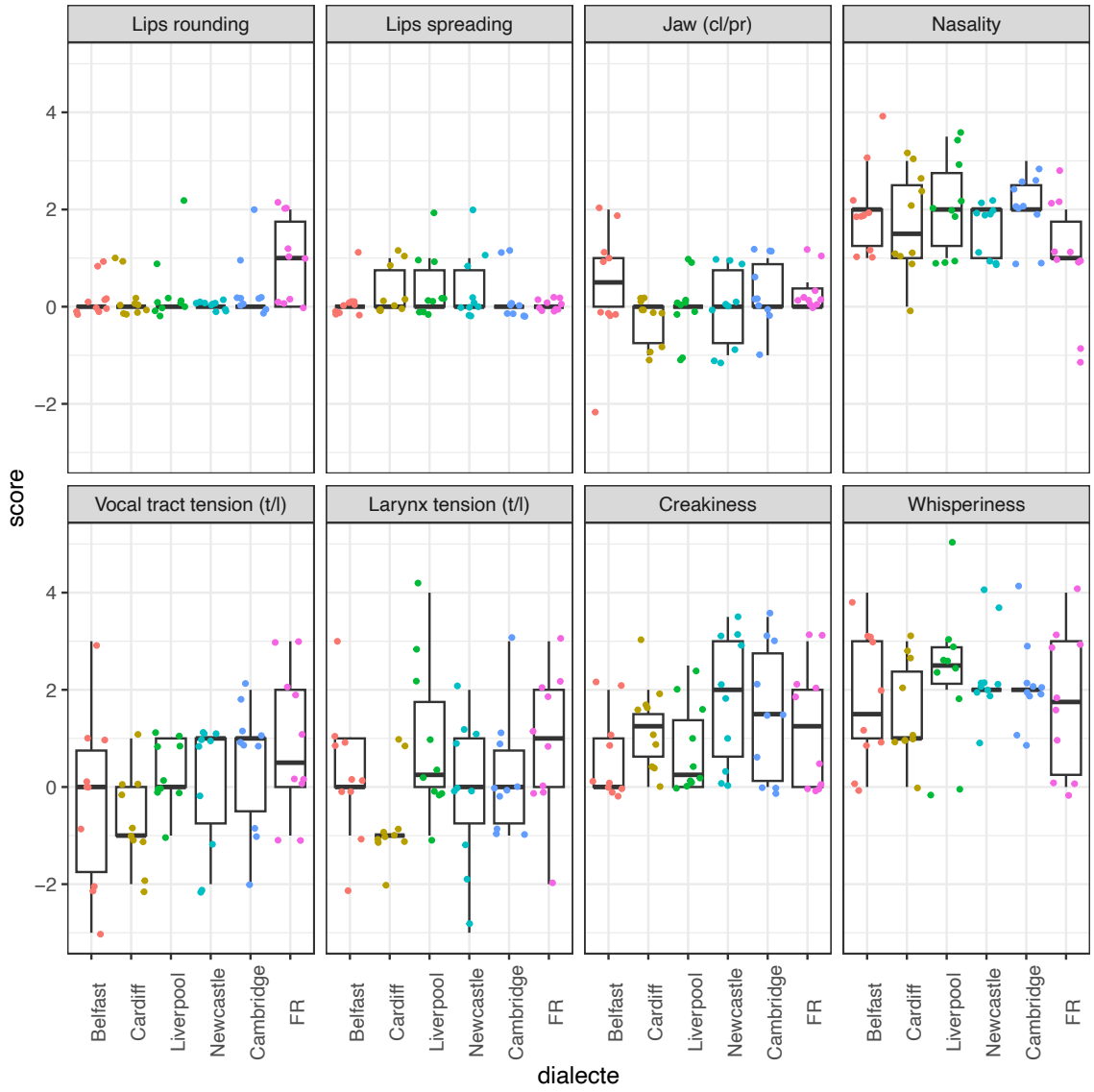


Figure 46 : Réanalyse descriptive de la distribution des scores attribués à chacun des 60 locuteurs (10 par groupe dialectal) dans l'évaluation des huit dimensions du Vocal Profile Analysis retenues dans l'étude de 2009. L'évaluation individuelle par locuteur étant représentée par les points colorés, les boîtes à moustaches qui représentent la distribution au sein de chaque groupe n'incluent pas les valeurs extrêmes (outliers) afin d'éviter les redondances. Les points sont représentés avec un léger décalage aléatoire de leurs coordonnées horizontales et verticales afin d'améliorer la lisibilité de la figure. Les apprenants francophones (FR) sont représentés en rose sur la droite de chacun des cadres correspondant à l'une des dimensions du VPA.

Comme l'illustre la Figure 46, une visualisation de ces données se concentrant plus directement sur la variabilité inter-locuteurs au sein de chaque groupe révèle une importante variabilité interindividuelle pour la grande majorité des huit dimensions du VFA prises en compte. Si pour les dimensions pour lesquelles aucun effet du groupe dialectal n'a été mis en évidence précédemment cela n'est pas surprenant en considérant l'analyse de variance comme une comparaison entre la variabilité intra-groupe et inter-groupes, une absence d'effet pouvant alors être interprétée comme le fait que la variabilité inter-groupes n'est pas supérieure à la variabilité intra-groupe, nous pouvons observer que même dans le cas des dimensions d'arrondissement et de nasalité la variation individuelle reste conséquente. En effet certains apprenants francophones sont évalués à un niveau qui correspond majoritairement à celui des anglophones natifs et inversement.

Ainsi, parmi les dix apprenants francophones inclus dans l'étude, quatre sont évalués avec un degré neutre d'arrondissement des lèvres (valeur 0) qui constitue la valeur majoritaire parmi les anglophones natifs. Notons également que seuls deux de ces dix apprenants francophones sont évalués avec un degré de nasalisation considéré comme inférieur à un degré neutre. La comparaison recentrée sur les locuteurs de Cambridge et les francophones révèle également un recouvrement entre le degré de tension générale du conduit vocal attribué à ce groupe de locuteur natifs et celui attribué aux apprenants francophones. Dans une moindre mesure, c'est également le cas pour l'évaluation du degré de tension laryngée.

6.1.2 Synthèse performative pour l'acquisition de l'intonation en L2

Publications associées :

[ACTI7] Xiao, X., Bonnet, C., Zhang, H., **Audibert, N.**, Kühnert, B., & Pillot-Loiseau, C. (2024). Enseignement de l'intonation du français par une synthèse vocale contrôlée par le geste : étude de faisabilité. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 342-350.

[ACTI13] Xiao, X., Kühnert, B., **Audibert, N.**, Locqueville, G., Pillot-Loiseau, C., Zhang, H., & d'Alessandro, C. (2023). Performative Vocal Synthesis for Foreign Language Intonation Practice. *Proceedings of CHI '23: CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany. pp.1-9.

[ACTI18] Xiao, X., **Audibert, N.**, Locqueville, G., d'Alessandro, C., Kühnert, B., Kleinberger, R., & Pillot-Loiseau, C. (2022). Évaluation de la stylisation chironomique pour l'apprentissage de l'intonation du français L2. *Actes des 34e Journées d'Études sur la Parole (JEP2022)*, Noirmoutier, France. pp. 434-442.

[ACTI20] Xiao, X., **Audibert, N.**, Locqueville, G., d'Alessandro, C., Kühnert, B., & Pillot-Loiseau, C. (2021). Prosodic Disambiguation Using Chironomic Stylization of Intonation with Native and Non-Native Speakers. *Proceedings of Interspeech 2021*, Aug 2021, Brno (virtual), Czech Republic, pp. 516-520.

6.1.2.1 La synthèse vocale performative et son application à l'intonation

Dans le cadre du projet ANR Gepeto porté par Christophe d'Alessandro (Institut d'Alembert) de 2020 à 2024, le Laboratoire de Phonétique et Phonologie a principalement été en charge de l'application à la didactique de l'intonation en langue seconde d'outils de synthèse vocale performative développés par l'équipe de l'institut d'Alembert à travers l'intégration d'un

vocodeur temps-réel à des interfaces utilisables en salle de classe, et l'évaluation de ces outils via l'élaboration de corpus et de protocoles spécifiques.

Le terme de synthèse vocale performative désigne de façon large le contrôle par le geste d'outils de synthèse vocale, et plus spécifiquement dans le cadre de ce projet le contrôle chironomique de la source vocale via les gestes manuels. Bien qu'une large part des applications préalables de la synthèse vocale performative se soient concentrée sur la voix chantée, notamment à travers l'application *Cantor Digitalis* (Feugère et al., 2017), l'efficacité de la stylisation chironomique par des locuteurs natifs de l'intonation française à l'aide d'une tablette graphique contrôlée par un stylet a déjà été démontrée (d'Alessandro et al., 2011).

Les motivations pour l'application à l'acquisition de l'intonation en langue seconde sont multiples. D'une part, le développement et l'évaluation d'applications dédiées à la didactique de l'intonation en lange seconde répond à un besoin. En effet bien que le rôle prépondérant de l'intonation L2 dans la perception d'un accent étranger et même l'intelligibilité soit bien établi (Munro & Derwing, 1995) et que les patrons intonatifs non-existants en L2 soient plus problématiques pour les apprenants (Mennen, 2015), l'enseignement de l'intonation en langue seconde reste largement négligé, à la fois en classe (Thomson & Derwing, 2015) et dans le cadre du développement d'applications informatisées (Pennington, 1999). De plus, des travaux sur l'acquisition L2 de l'intonation de l'espagnol par des apprenants sinophones ont conclu à l'efficacité de l'illustration des cibles intonatives par le geste ou des représentations visuelles (Yuan et al., 2019), ainsi qu'à l'efficacité du marquage par le geste manuel de caractéristiques suprasegmentales auprès d'apprenants catalans du français (Bailly et al., 2022).

L'entraînement à partir du geste via la synthèse performative pourrait en outre faciliter l'acquisition de contours intonatifs spécifiques à la langue seconde au-delà de l'explicitation des cibles à produire et du surcroît de motivation apportée par l'utilisation d'un outil novateur. En effet, en complément de l'illustration visuelle par des gestes manuels ou par le tracé de trajectoires intonatives, le recours à la synthèse performative permet à l'apprenant d'obtenir un retour auditif direct et immédiat. L'apprenant a ainsi la possibilité de moduler simplement à travers le tracé d'une courbe avec le doigt ou un stylet la hauteur mélodique de l'énoncé sur l'axe vertical (avec une échelle en demi-tons afin de correspondre plus directement à la perception des modulations de hauteur) ainsi que les durées sur l'axe horizontal, ce qui lui permet de prendre conscience plus efficacement des conséquences perceptives des modifications de contours prosodiques.

Dans le cadre du projet Gepeto, j'ai travaillé en étroite collaboration avec Xiao Xiao, post-doctorante spécialisée en développement d'interfaces interactives, qui a adapté les outils de synthèse performative afin de les rendre utilisables sur tablette ou smartphone dans un contexte de salle de classe et a pris en charge les premières étapes de recueil et d'évaluation des productions des apprenants. Sur le plan purement scientifique, je suis principalement intervenu en m'impliquant dans la conception des corpus et des protocoles expérimentaux en production et perception ainsi que dans l'analyse acoustique et statistique des données recueillies et l'interprétation des résultats obtenus.

En complément de cela, mon rôle dans ce projet a également été d'assurer l'interface entre les spécialistes d'acoustique et d'informatique d'une part, et les linguistes et didacticiens d'autre part, afin d'assurer la transmission des compétences techniques nécessaires pour l'utilisation de tels outils. En effet et bien qu'un effort considérable ait été accompli pour

permettre le contrôle par l'enseignant des passations à travers une interface graphique, les outils employés dans le cadre du projet reposent sur la configuration à la fois de l'ordinateur utilisé comme serveur pour transformer les gestes tracés manuellement en modulations intonatives, et du protocole réseau utilisé pour permettre la communication entre ce serveur et l'appareil mobile utilisé comme terminal. J'ai donc été amené à former les enseignants de langue impliqués dans le projet et à mettre en place un protocole simplifié afin que les apprenants puissent accéder à l'application sur leur téléphone à partir d'un simple *QR code* lors des entraînements et passations. Bien que cet aspect puise sembler relever de tâches d'ingénierie plutôt que scientifiques, il me semble utile de le mentionner dans ce document de synthèse car cela fait écho à mes activités en matière de transfert de compétences méthodologiques via le développement d'outils accessibles sur lequel je reviens dans le dernier chapitre de ce volume.

Une première étude pilote que je ne détaillerai pas ici a porté sur un ensemble de phrases françaises prosodiquement ambiguës soumises à un panel de locuteurs francophones natifs ainsi que d'apprenants non-natifs. Cette étude pilote a permis à travers la comparaison entre imitation vocale et reproduction par le geste manuel des distances mélodiques par rapport à l'énoncé de référence de valider la faisabilité de la reproduction de modèles d'énoncés par des sujets naïfs. Elle a également permis d'identifier certaines limites qui ont guidé les deux études dont les objectifs et les principaux résultats sont résumés ci-dessous. Ainsi, si la prise en main de l'interface s'est avérée aisée pour la plupart des sujets sur des énoncés courts, elle a été beaucoup plus complexe sur des énoncés de 10 syllabes ou plus. De plus, les modulations rythmiques à travers la gestion des durées syllabiques se sont révélées beaucoup plus difficiles à appréhender que celles de hauteur mélodique. Dans la suite de nos travaux nous avons donc privilégié des énoncés courts, et avons fait le choix d'afficher un guide visuel pour indiquer aux apprenants les durées syllabiques de référence afin de leur permettre de mieux se concentrer sur les modulations de hauteur mélodiques.

6.1.2.2 Acquisition par des francophones du contour fall-rise de l'anglais britannique

Cette étude s'appuie sur le cadre établi par l'école dite « britannique » d'analyse de l'intonation anglaise (Cruttenden, 1997), qui est celui le plus communément retenu pour l'enseignement de l'anglais langue seconde en France (Herment & Tortel, 2021). Une difficulté des apprenants français de l'anglais britannique observée de façon récurrente par les enseignants a été ciblée : l'acquisition du contour localisé descendant-montant (*fall-rise*), utilisé par les locuteurs natifs pour signaler une implication et ainsi mitiger le sens de l'énoncé. Dans une phrase négative, le recours au contour *fall-rise* qui limite la portée de la négation est fréquemment observé dans des énoncés ironiques mais son usage est plus large que ce celui de l'expression de l'ironie.

Le corpus recueilli pour les besoins de cette étude s'est ainsi concentré sur un ensemble de douze phrases négatives pouvant faire l'objet de deux interprétations distinctes selon qu'un contour descendant (*fall*) ou descendant-montant (*fall-rise*) était produit sur le mot-cible. Ainsi, dans la phrase "Ken didn't feed the **cat**." (*Ken n'a pas nourri le chat*), la production d'un contour descendant (*fall*) sur la voyelle /æ/ du mot « cat » signifie que Ken a oublié de nourrir le chat, tandis qu'un contour descendant-montant (*fall-rise*) indique qu'il a bien nourri un animal ou une personne mais qu'il ne s'agit pas du chat.

Les douze phrases déclinées chacune avec les deux intonations ont été produites par un locuteur natif de l'anglais britannique. Suite à une évaluation perceptive en choix forcé auprès

de locuteurs natifs de l'anglais britannique devant sélectionner l'illustration correspondant le mieux à la signification implicite de l'énoncé, quatre paires de phrases identifiées correctement par plus de 75% des auditeurs ont été sélectionnées pour l'expérience de production. Pour cette expérience de production, 12 locuteurs natifs de l'anglais britannique et 12 apprenants francophones de niveau intermédiaire ont été recrutés. Les sujets ont été soumis à trois tâches de production, guidées par les illustrations des deux significations possibles pour chaque énoncé : la lecture libre, l'imitation vocale et l'imitation gestuelle à l'aide de l'outil de synthèse performative. Dans cette dernière condition, après une phase de familiarisation les sujets devaient reproduire sur une tablette numérique les modèles présentés auditivement et visuellement des contours intonatifs produits par le locuteur natif de référence, avec la possibilité d'utiliser au choix leur doigt ou un stylet. Le système de synthèse vocale performative permettait aux sujets d'écouter en temps réel le résultat des modulations de fréquence fondamentale produites par leurs gestes manuels, à partir d'une version des productions natives modifiée par analyse-resynthèse pour correspondre à une intonation monotone (fréquence fondamentale constante) avec des durées syllabiques isochroniques. La Figure 47 illustre la présentation de l'interface pour les deux versions de la phrase "Nick doesn't listen to anyone" (interprétable comme « *Nick n'écoute personne »* avec l'intonation *fall* sur la syllabe finale, ou « *Nick n'écoute que les personnes de son choix »* avec l'intonation *fall-rise*), ici avec l'affichage du guide visuel et des durées syllabiques de référence.

Les modulations de la fréquence fondamentale normalisées en durée dans les productions des locuteurs natifs et des apprenants ont fait l'objet d'une analyse statistique comparative à l'aide de modèles GAM, en concentrant l'analyse sur la partie des énoncés contrastive entre les deux types d'intonation, située en position finale. Tandis que les locuteurs natifs ont produit des modulations significativement distinctes entre *fall* et *fall-rise* sur une partie de la portion finale de l'énoncé dans chacune des trois conditions de production, les productions des apprenants francophones n'étaient significativement distinctes entre *fall* et *fall-rise* sur une partie de la portion finale que dans les deux conditions d'imitation (vocale et gestuelle), mais pas en condition de lecture libre. Par ailleurs la comparaison entre natifs et apprenants a confirmé que l'intonation *fall* ne pose pas de problèmes particuliers aux apprenants dont les productions en lecture libre et imitation vocale n'étaient pas significativement différentes de celles de natifs, tandis qu'elles l'étaient pour l'intonation *fall-rise* avec des écarts pouvant aller jusqu'à 4 demi-tons. Enfin, bien que des contours significativement différents entre les conditions d'imitation vocale et d'imitation gestuelle aient été produits par les apprenants (contrairement aux natifs), ces différences sont peu pertinentes perceptivement avec un écart maximal inférieur à 2 demi-tons.

En complément, une évaluation perceptive par cinq auditeurs experts (enseignants de phonétique anglaise dans des universités françaises) des productions par la voix et par le geste des natifs et des apprenants a confirmé que les apprenants francophones ne maîtrisent pas l'intonation *fall-rise* et se montrent dans l'ensemble incapables de la produire en condition de lecture libre contrairement aux natifs qui la produisent naturellement de façon consistante, mais que les apprenants sont parfaitement capables de produire ce patron intonatif aussi bien en condition d'imitation vocale que d'imitation gestuelle.

Figure 47 : Illustration de l'interface utilisée pour permettre aux apprenants de reproduire par le tracé manuel sur un appareil mobile les contours intonatifs d'un modèle d'énoncé présenté auditivement et/ou visuellement. La courbe violette correspond au tracé manuel par le sujet de la trajectoire intonative, transmise en temps réel au synthétiseur. De façon optionnelle, le tracé de la trajectoire intonative de référence peut être affiché (ici en jaune) ainsi que les durées segmentales et la transcription correspondante. Les boutons sur la gauche permettent d'écouter le modèle, d'afficher ou masquer les différents guides visuels, de réécouter les modulations produites et de les conserver lorsque la production est jugée satisfaisante. Les deux exemples présentés ici correspondent aux deux versions de la phrase anglaise "Nick doesn't listen to anyone" : à gauche avec l'intonation *fall* sur la syllabe finale, à droite avec l'intonation *fall-rise*.

6.1.2.3 Acquisition par des locuteurs ukrainiens de patrons prosodiques du français

Si les résultats de l'étude de l'acquisition du contour *fall-rise* en anglais britannique se sont montrés prometteurs en confirmant la capacité d'apprenants francophones à reproduire à la fois vocalement et par le geste manuel un patron intonatif de l'anglais connu pour être particulièrement problématique pour les apprenants francophones, le protocole employé reposait sur une session unique de recueil de données qui ne permet d'obtenir un aperçu des capacités des apprenants qu'à un certain instant du processus d'apprentissage. Afin d'effectuer un pas supplémentaire vers des conditions réalistes d'enseignement de la prosodie en langue seconde, l'étape suivante a consisté en la mise en place d'une étude longitudinale dans un contexte de salle de classe, dans la limite de ce que permet un outil encore en phase de développement. Le choix s'est porté sur l'enseignement des quelques patrons prosodiques du français à destination d'apprenants ukrainiens débutants en français, ayant émigré récemment en raison de la situation politique dans leur pays sans connaissances préalables du français. Les patrons prosodiques sélectionnés ont été ceux posant des difficultés récurrentes de communication aux locuteurs ukrainiens, à savoir la distinction entre contour final déclaratif et questions polaires. Bien que les questions polaires en ukrainien soient marquées par un contour final montant, de même qu'en français, les difficultés de communication observées par les enseignants de français pourraient être liée à la présence dans certains énoncés déclaratifs ukrainiens d'un accent de hauteur lié à une focalisation large, qui peut avoir pour conséquence la réalisation d'un contour final montant (Pompino-Marschall et al., 2017). Dans le cadre de cette étude, la distinction entre déclaration et question polaire a été complétée par l'expression de l'attitude d'incrédulité, marquée en français par un contour final montant de forte amplitude (Morlec et al., 2001).

Afin de tenir compte du niveau limité en français des apprenants ukrainiens ciblés et des difficultés observées préalablement, le corpus utilisé a été composé d'énoncés courts intégralement voisés d'une à quatre syllabes, produits par un locuteur francophone natif. Chaque énoncé a été décliné sous la forme d'une déclaration, d'une question polaire et d'une expression d'incrédulité. Les énoncés ont été séparés en trois sous-ensembles, utilisés respectivement lors d'une phase initiale de pré-test en condition de lecture libre destinée à évaluer le niveau initial des apprenants, lors de quatre sessions de formation suivant un protocole comparable à la session de recueil de données réalisée dans l'étude sur l'anglais britannique, et lors du post-test en condition de lecture libre destiné à évaluer les capacités de généralisation des apprenants à l'issue de leur apprentissage. En complément, les deux dernières sessions de formation incluaient également une condition de reproduction par le geste manuel à l'aide de la synthèse performative sans affichage du guide visuel. L'objectif de cette étude était ainsi de pouvoir évaluer les progrès réalisés par les apprenants entre la session de pré-test et celle de post-test, suite à des sessions d'entraînement s'appuyant sur la synthèse performative.

La même interface que celle utilisée pour l'étude sur l'anglais britannique (Figure 47) a été employée, en utilisant un codage orthographique du découpage syllabique plutôt que des symboles API pour faciliter la prise en main de l'outil par les apprenants, qui l'utilisaient sur leur propre téléphone. En raison de la difficulté à imager de façon claire et non-ambiguë dans les trois conditions certains des énoncés retenus, la production a été contextualisée par de brèves descriptions textuelles d'un contexte fictif dans lequel chacune des trois versions de chaque énoncé pourrait être produit en complément du recours à la ponctuation.

Dix apprenants ukrainiens (sept femmes et trois hommes) ont participé à la première session d'enregistrement lors du pré-test. L'analyse des productions de ces dix apprenants a révélé des productions très peu distinctes entre modalités avec une prépondérance de contours montants quelle que soit la modalité. La seule exception relevée a été pour un énoncé déclaratif de quatre syllabes produit avec un contour plus similaire à celui du modèle natif.

En raison des mouvements sociaux survenus pendant le semestre universitaire lors duquel les sessions d'enregistrement et de formation ont eu lieu et de difficultés personnelles ayant affecté certains apprenants, seules deux femmes ont effectué l'intégralité des sessions de formations et le post-test final. L'ampleur de l'analyse des résultats de l'étude longitudinale a donc été plus limité que prévue, et s'apparente plutôt à une étude de cas ne permettant pas de généralisation. Cette analyse, effectuée également au moyen de modèles GAM, permet néanmoins de dégager des observations intéressantes quant à l'applicabilité de la synthèse performative dans un contexte de salle de classe. Ainsi des progrès notables ont été réalisés par les deux apprenantes dans la réalisation des énoncés déclaratifs, plus proches du modèle natif en condition post-test qu'ils ne l'étaient en pré-test. Ces progrès ont été plus importants pour l'une des deux apprenantes, qui était aussi plus à l'aise pour prendre en main l'interface. Les résultats concernant l'expression de l'incrédulité ont été moins probants, ce qui pourrait être lié à la variabilité plus importante des patrons intonatifs utilisés par le locuteur natif pour exprimer l'incrédulité. Par ailleurs cette hypothèse est difficilement vérifiable en l'absence d'un corpus de référence d'expressions d'incrédulité en ukrainien mais il est vraisemblable que de telles expressions soient également plus variables en ukrainien que des contours associés à la modalité, par nature plus fortement conventionnalisés.

6.1.2.4 Effets du locuteur : variation individuelle dans les données de production

Je propose ci-dessous une réanalyse descriptive des données de production obtenues dans chacune de ces deux études afin de mettre en lumière les éventuelles spécificités individuelles susceptibles d'avoir été masquées par les modèles statistiques de type GAM employés dans les études publiées.

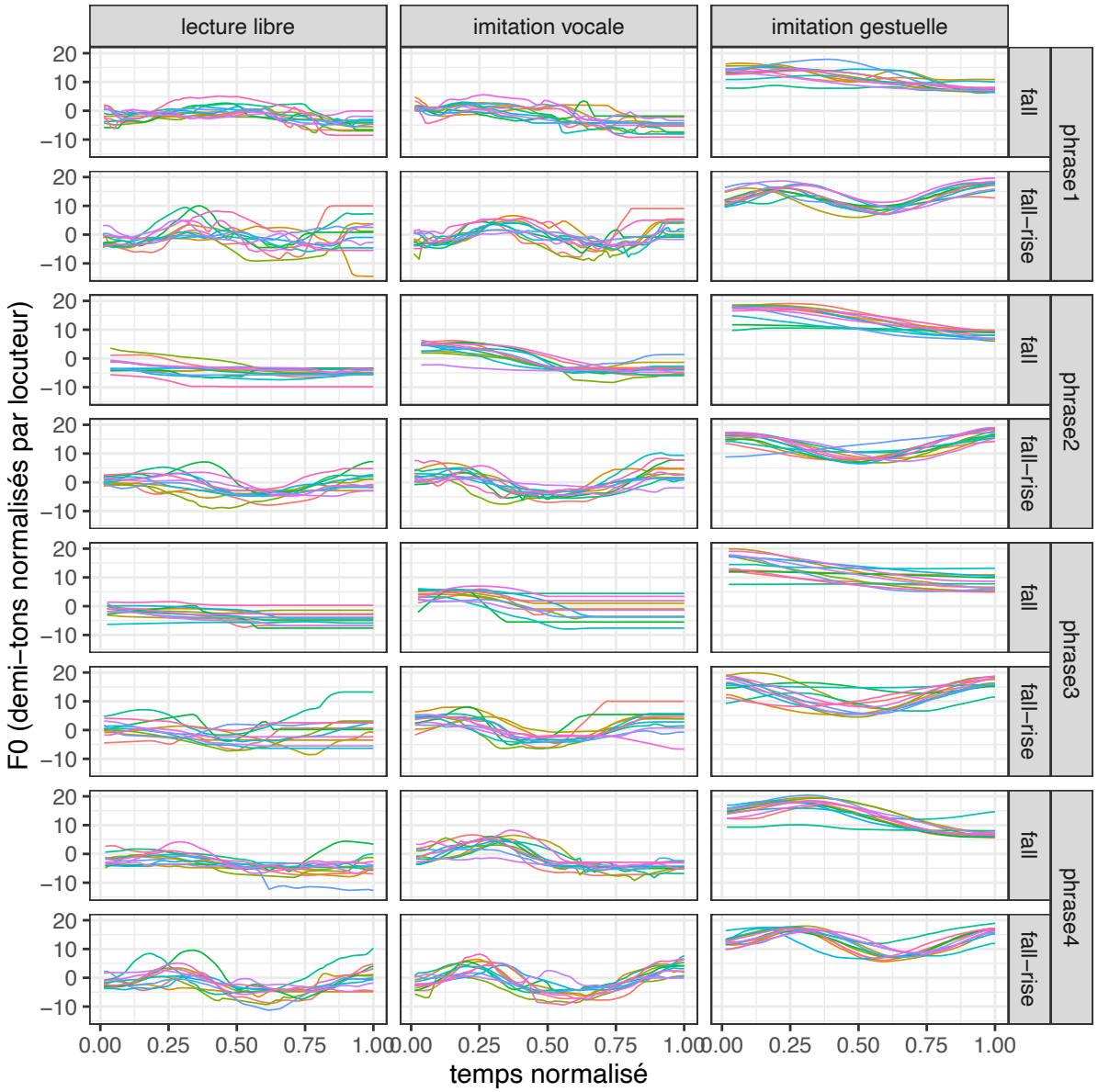


Figure 48 : Tracés individuels des contours mélodiques normalisés temporellement sur la syllabe finale correspondant aux productions des douze locuteurs natifs de l'anglais britannique avec l'intonation *fall* et *fall-rise*, pour chacune des quatre phrases sélectionnées dans chacune des trois conditions de production. Dans les conditions de lecture libre et d'imitation vocale, le niveau 0 demi-tons correspond au registre moyen de chaque locuteur.

La Figure 48 et la Figure 49 présentent le tracé des contours intonatifs normalisés temporellement produits dans l'étude sur l'acquisition de l'intonation *fall-rise* de l'anglais britannique par chaque locuteur pour chacune des phrases sélectionnées et dans chacune des conditions, respectivement pour les locuteurs anglophones natifs (Figure 48) et pour les

apprenants francophones (Figure 49). En faisant abstraction des artefacts introduits par la procédure de normalisation temporelle nécessaire pour l'analyse statistique par modèles GAM qui a pour effet de remplacer les valeurs indéfinies par une extrapolation avec des valeurs constantes à l'exception de quelques erreurs de détection (observables par exemple dans la Figure 49, pour la phrase 4 produite avec l'intonation *fall*) qui subsistent en dépit des corrections manuelles effectuées, ces productions se caractérisent par une variation inter-locuteurs remarquablement faible. C'est tout particulièrement le cas dans les deux conditions d'imitation, aussi bien vocale que gestuelle. Si la variation individuelle est plus importante en condition de lecture libre, elle reste modérée, tout particulièrement dans le cas des locuteurs anglophones natifs.

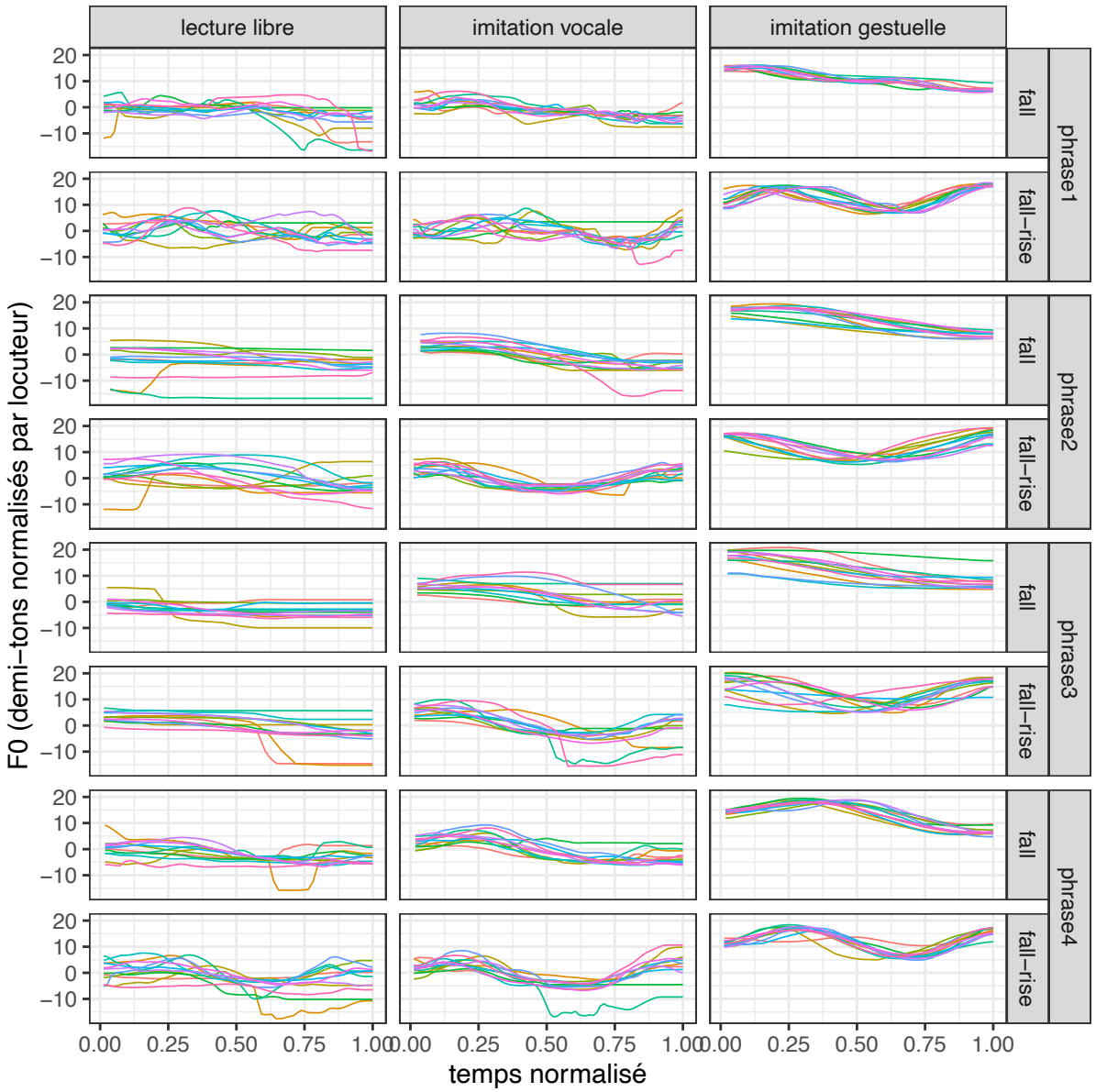


Figure 49 : Tracés individuels des contours mélodiques normalisés temporellement sur la syllabe finale correspondant aux productions des douze apprenants francophones de l'anglais britannique avec l'intonation *fall* et *fall-rise*, pour chacune des quatre phrases sélectionnées dans chacune des trois conditions de production. Dans les conditions de lecture libre et d'imitation vocale, le niveau 0 demi-tons correspond au registre moyen de chaque locuteur.

Dans le cas des locuteurs anglophones natifs en condition *fall-rise*, cette variation individuelle en lecture libre se manifeste à la fois par des différences de timing dans la réalisation des pics et vallées, et par des différences d'amplitude du contour. Ces variations dans l'amplitude des contours pourraient être liées à des différences individuelles d'expressivité, certains locuteurs tendant à se montrer plus expressifs que d'autres à la fois dans leurs productions spontanées mais aussi dans des tâches de lecture plus ou moins contrôlées. De plus il est probable que le protocole retenu pour éliciter la production d'un contour *fall* ou *fall-rise* ait favorisé une exagération des mouvements mélodiques associés à ces contours en comparaison de productions spontanées en interaction, ce qui se traduit ici par des différences individuelles d'amplitude. Dans le cas des apprenants la variabilité individuelle se manifeste majoritairement par un décalage du registre de fréquence fondamentale employé (par rapport à la valeur 0 demi-tons qui correspond au registre moyen de chaque locuteur), et dans une moindre mesure par une variation de la forme du contour qui reste toutefois presque systématiquement éloigné du modèle natif en condition *fall-rise*.

La Figure 50 présente le tracé des contours intonatifs normalisés temporellement produits en condition pré-test par les dix locuteurs ayant participé à la phase initiale dans l'étude sur l'acquisition de l'acquisition de la modalité en français par les apprenants ukrainiens, pour chacune des phrases sélectionnées et dans chacune des trois modalités. Hormis quelques excursions minoritaires qui semblent être des erreurs de détection aboutissant à des valeurs anormalement basses et qui d'après une inspection qualitative des signaux apparaissent comme liés à un dévoisement de la part de certains locuteurs, on observe pour l'ensemble des apprenants des contours montants à la fois pour les déclarations et les questions polaires, avec toutefois une pente ascendante légèrement plus marquée pour les questions polaires pour la plupart des locuteurs. On observe donc là aussi une importante consistance dans les productions des différents apprenants, alors même que les productions sont relativement peu contraintes par le protocole d'élicitation retenu. A noter toutefois qu'une locutrice n'ayant pas pris part à l'intégralité des sessions (en vert foncé dans la figure) se détache. Cette locutrice adopte en effet de façon plus consistante que les autres apprenants une stratégie pour marquer la question polaire de façon contrastive, avec une montée intonative plus importante que celle observée dans le modèle natif. Si une stratégie similaire est également observée de la part d'autres apprenants, cela ne concerne qu'une partie des énoncés et majoritairement ceux de trois et quatre syllabes.

Les productions de l'expression d'incrédulité par les apprenants sont quant à elles marquées par une grande variabilité, avec toutefois une tendance majoritaire à la production de contours montants proches de ceux de associés aux questions polaires par les apprenants. On peut noter que l'apprenante UF2 (tracé en magenta) ayant participé à l'ensemble des sessions adopte une stratégie différente d'expression de l'incrédulité lors du pré-test avec un contour final descendant, mais modifie ses productions suite aux sessions d'entraînement avec des expressions d'incrédulité en post-test (Figure 51) caractérisées par un contour final montant et proches des questions polaires produites par cette locutrice.

La Figure 51 présente les contours mélodiques individuels produits en condition post-test par les deux apprenantes ukrainiennes du français ayant participé à la l'ensemble des sessions. Les tracés présentés sur cette figure confirment l'observation faite dans la version publiée de l'étude qu'à l'issue des quatre sessions d'entraînement la locutrice UF1 (tracés en rouge) parvient mieux à produire des contours finaux déclaratifs proches du modèle francophone natif que la locutrice UF2 (en magenta) qui produit des contours déclaratifs plus fortement

descendants sur les énoncés monosyllabiques mais conserve un contour final montant sur l'énoncé de trois syllabes. La locutrice UF1 montre également une plus grande capacité de généralisation de la production des contours d'incrédulité que la locutrice UF2, avec l'utilisation des formes distinctes observées dans le modèle natif sur des énoncés de longueur variable.

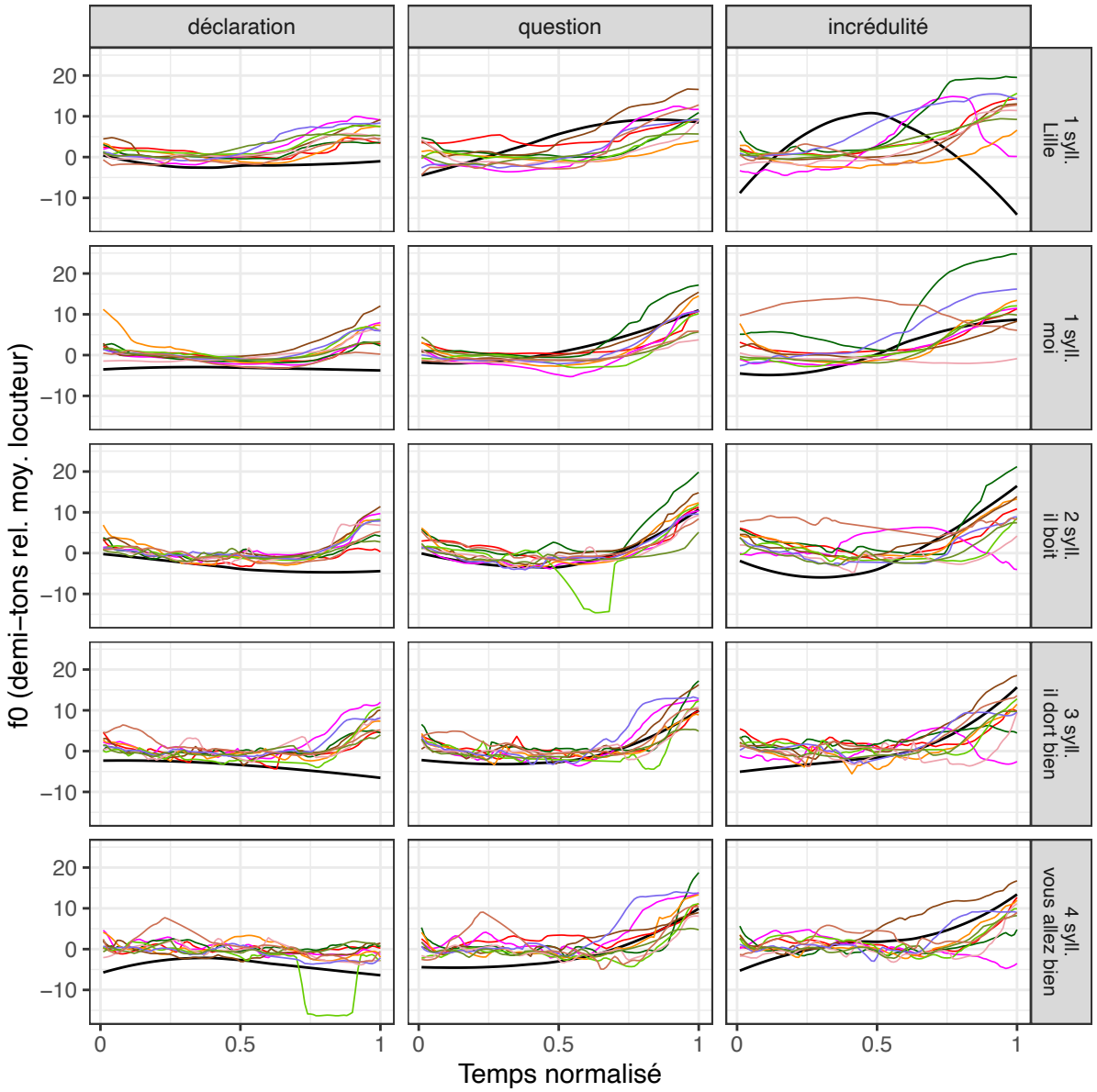


Figure 50 : Tracés individuels des contours mélodiques normalisés temporellement sur la syllabe finale, correspondant aux productions vocales en lecture libre des dix apprenants ukrainiens du français en condition pré-test, pour chacune des cinq phrases d'une à quatre syllabes incluses dans le pré-test (lignes) et chacune des trois modalités (colonnes). Le niveau 0 demi-tons correspond au registre moyen de chaque locuteur. En complément des productions des dix apprenants, la référence native est représentée en trait noir plus épais. Les deux apprenantes ayant pris part à l'intégralité des sessions sont représentées respectivement en rouge (UF1) et magenta (UF2).

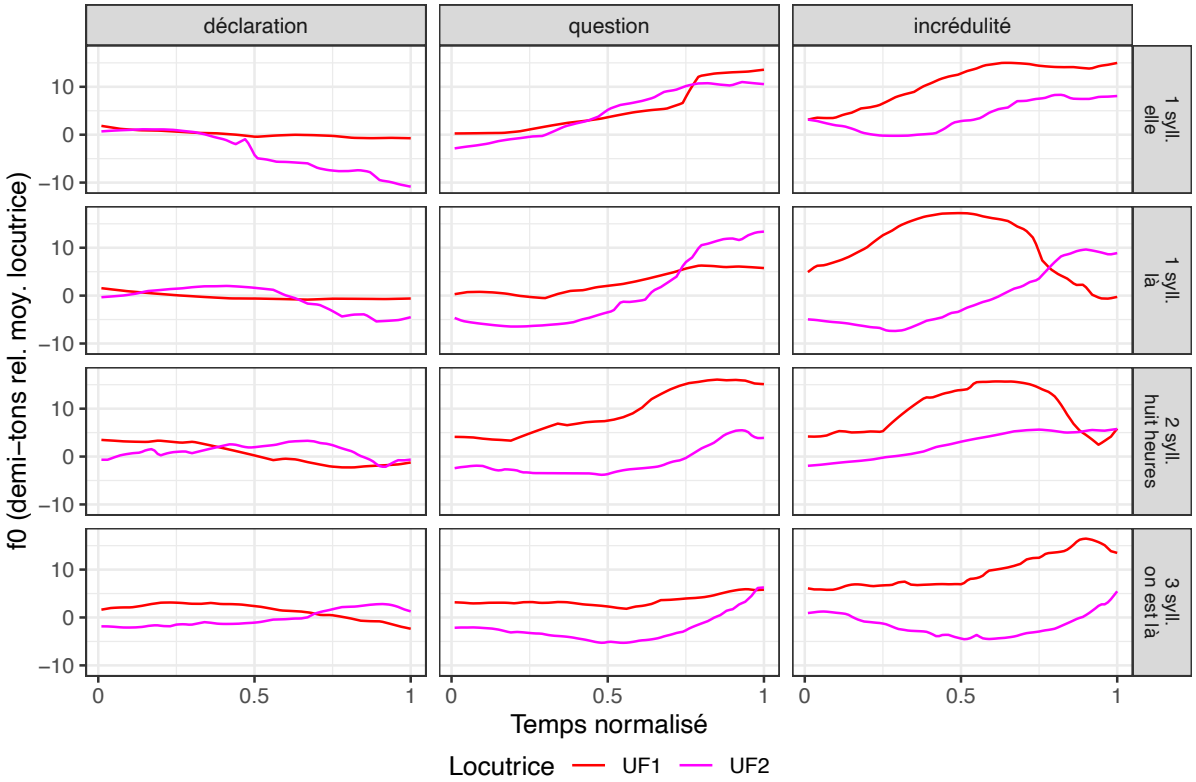


Figure 51 : Tracés individuels des contours mélodiques normalisés temporellement sur la syllabe finale, correspondant aux productions vocales en lecture libre en condition post-test des deux apprenantes ukrainiennes du français ayant participé à l'intégralité des sessions d'entraînement et d'enregistrement, pour chacune des quatre phrases d'une à trois syllabes incluses dans le post-test (lignes) et chacune des trois modalités (colonnes). Le niveau 0 demi-tons correspond au registre moyen de chaque locutrice.

En complément des modulations de fréquence fondamentale qui constituent la cible de la méthode d'apprentissage proposée, il est possible que les apprenantes adoptent des stratégies reposant sur des oppositions de durée afin de produire des contrastes entre modalités, qui seraient masquées dans l'analyse présentée ci-dessus par la normalisation temporelle effectuée sur la dernière syllabe des énoncés. La Figure 52 illustre la durée de la syllabe finale pour chacune des deux apprenantes ayant participé à l'ensemble des sessions, comparée entre pré-test et post-test pour chaque longueur d'énoncé et chaque modalité. Alors que la durée de la syllabe finale est remarquablement homogène entre modalités dans les productions du locuteur natif (non représentées ici), on remarque chez les deux apprenantes des différences de durées entre modalités, avec pour les deux apprenantes une stratégie qui évolue entre pré-test et post-test. Tandis que la locutrice UF1 qui dans la condition pré-test tendait à produire un allongement final beaucoup plus conséquent dans les énoncés déclaratifs de deux et trois syllabes réduit fortement cet allongement des énoncés déclaratifs à l'issue des sessions d'entraînement mais le conserve pour les expressions d'incrédulités, la locutrice UF2 tend vers une plus grande homogénéisation des durées entre modalités.

On peut également noter chez les deux locutrices une réduction générale de la durée de la syllabe finale, qui reflète une augmentation du débit de parole dans leurs productions en français. Cette augmentation du débit de parole, plus marquée pour la locutrice UF1 pour la

production d'énoncés déclaratifs et de questions polaires, pourrait refléter une plus grande aisance de cette locutrice dans les situations de communication en langue seconde.

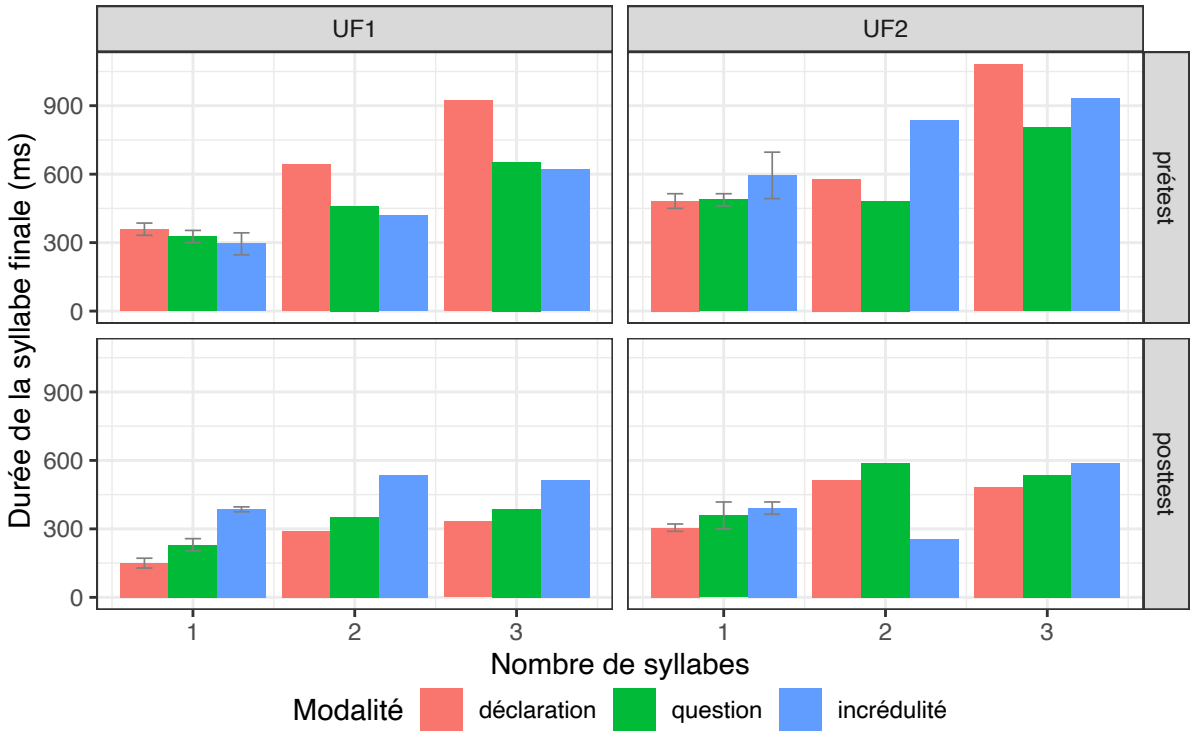


Figure 52 : Durée de la syllabe finale des énoncés produits en lecture libre en condition pré-test et post-test par les deux apprenantes ukrainiennes du français ayant participé à l'intégralité des sessions d'entraînement et d'enregistrement, pour chaque longueur d'énoncé en nombre de syllabes et chacune des trois modalités. Les barres d'erreur représentées pour les énoncés d'une syllabe représentent l'erreur-type et illustrent la variabilité entre les deux énoncés monosyllabiques inclus en pré-test et en post-test.

6.1.2.5 Quelles implications pour la didactique de la prosodie ?

La série d'études expérimentales pilotes réalisées dans le cadre du projet Gepeto a permis de valider la faisabilité technique et méthodologique de la synthèse vocale performative appliquée à la didactique de la prosodie. En effet, des sujets non-natifs sont capables aussi bien que les natifs de reproduire par le geste des contours intonatifs de la langue cible identifiés comme problématiques par les enseignants de langue. De plus dans l'étude sur l'acquisition de la modalité par les apprenants ukrainiens du français, des progrès notables ont été observés à l'issue des sessions d'entraînement. Dans le cadre de nos travaux sur la synthèse vocale performative, il serait bien entendu tentant d'en attribuer le mérite à l'utilisation de l'application proposée. En l'absence d'un groupe contrôle qui suivrait une formation fondée sur la répétition d'exemples, dont l'inclusion dans l'étude était prévue mais n'a pas été possible, voire d'un autre groupe contrôle dans lequel le recours au geste viendrait compléter les productions vocales de façon plus « classique », il convient de rester prudent quant à la portée de cette interprétation. Afin d'évaluer de façon plus probante l'impact de la synthèse vocale performative sur l'acquisition de l'intonation en langue seconde, il conviendrait non seulement de le faire auprès d'un nombre de sujets plus conséquent, mais surtout de quantifier le gain de performance attribuable à l'utilisation de cet apprentissage par le geste en comparaison de groupes contrôles d'apprenants.

Par ailleurs, les retours informels de la part à la fois des enseignants de langue et des apprenants ayant manipulé l'application font état d'un grand enthousiasme à l'égard de l'outil en dépit des quelques difficultés de prise en main observées qui s'expliquent en grande partie par le statut de prototype de l'application développée, qui peut être déroutante pour un certain nombre d'utilisateurs potentiels. Ainsi, quand bien même la synthèse vocale performative ne serait pas décisive dans les progrès réalisés, son usage pour la didactique de l'intonation en langue seconde resterait intéressant de par le surcroît de motivation que cet outil est susceptible d'apporter aux apprenants.

Outre ces difficultés de prise en main qui pourraient aisément être gommés dans une version plus aboutie de l'application, une utilisation plus large reste limitée par certains verrous technologiques, qui ne permettent pas encore d'envisager le déploiement à grande échelle d'une application dédiée à la didactique de l'intonation en langue seconde reposant sur la synthèse vocale performative. En effet, en dépit des efforts réalisés pour proposer un protocole aussi simple que possible à mettre en place au niveau technique, la nécessité d'utiliser un ordinateur comme serveur en complément d'un appareil mobile connecté au même réseau sans fil peut constituer un frein à une utilisation plus large par les enseignants de langue, tout particulièrement dans un contexte de salle de classe. Ainsi, la possibilité d'installer une telle application de synthèse vocale performative directement sur un appareil mobile, qui a été une demande récurrente des enseignants de langue avec qui nous avons été amenés à échanger dans le cadre du projet, pourrait permettre de toucher un public beaucoup plus large avec la possibilité pour les apprenants de l'utiliser en dehors du cadre strict des cours. Si la capacité de calcul des tablettes et smartphones modernes est généralement suffisante pour cela, cet objectif ne semble réaliste qu'à moyen terme car il nécessite un lourd travail de développement pour porter sur un système d'exploitation mobile le vocodeur temps-réel nécessaire au fonctionnement de la synthèse vocale performative.

Dans une moindre mesure, un autre facteur limitant en vue d'une utilisation plus large en salle de classe concerne les prétraitements nécessaires pour intégrer à l'application de nouveaux enregistrements, qui constitue l'autre demande récurrente de la part des enseignants de langue. En effet, hormis le cas particulier d'énoncés monosyllabiques, l'une des conclusions des premiers essais par des sujets naïfs a été la difficulté à appréhender la gestion du rythme, ce qui nous a conduit à proposer systématiquement un affichage des frontières syllabiques pour guider l'utilisation de la synthèse vocale performative. Si les autres prétraitements peuvent aisément être automatisés, l'intégration de nouveaux énoncés nécessitera une segmentation fine en unités plus courtes, qu'il s'agisse de frontières syllabiques, segmentales ou autre. Or en dépit des progrès importants des systèmes d'alignement forcé, ces outils automatiques ne permettent pas toujours d'obtenir une segmentation suffisamment fine des enregistrements de parole, tout particulièrement si l'utilisation de productions non-natives est envisagée pour permettre aux apprenants de produire via la synthèse vocale performative des modulations intonatives sur des enregistrements de leur propre voix.

En ce qui concerne la variation individuelle qui constitue le fil conducteur de la première partie de ce document, la comparaison des contours intonatifs individuels des apprenants indique une consistance relativement importante entre locuteurs en condition de lecture libre, aussi bien dans l'étude sur les contours *fall* et *fall-rise* de l'anglais britannique que dans celle portant sur les modalités du français. Cependant cette observation est à nuancer par le fait que les contours analysés dans ces deux études consistent en des variations locales observées

sur une seule syllabe, ce qui pourrait avoir pour conséquence de limiter la plage de variation possible. En outre et bien que la tâche de lecture libre permette une certaine marge d'interprétation et donc de variation, une telle tâche de production sur des énoncés isolés reste d'autant plus contrainte que le contexte a été fixé par le biais d'illustrations ou de descriptions textuelles.

Au-delà des résultats de ces études qui ne permettent pas d'avancer des interprétations sur ce point, on peut s'interroger sur l'ampleur de la variation individuelle des patrons prosodiques dans les productions d'apprenants non-natifs en comparaison des productions natives. Ainsi, pour une même langue maternelle et un niveau comparable atteint dans la langue seconde, il est vraisemblable que les patrons prosodiques produits en L2, qui ne semblent pas pouvoir être interprétés comme une simple reproduction de ceux de la L1 de l'apprenant, soient plus stéréotypiques et donc moins variables que ceux de locuteurs natifs.

6.2 Variation en fonction de la langue et du sexe

Publications associées :

[ACTI19] Yoon, D., **Audibert, N.**, & Fougeron, C. (2022). Différences mélodiques et spectrales entre sexes comparées chez les locuteurs coréens et français. *Actes des 34e Journées d'Études sur la Parole (JEP2022)*. Noirmoutier, France, pp. 231-240.

[ACTI21] Yoon, D., **Audibert, N.**, & Fougeron, C. (2021). Effects of Body Size on Speech Production of Male and Female Speakers of Korean. *Proceedings of the 22nd Biennial Meeting of the International Circle of Korean Linguistics (ICKL-2021)*, Taipei, Taiwan, pp. 409-412.

[ACTI23] Yoon, D., **Audibert, N.**, & Fougeron, C. (2020). Effets du sexe et de la langue parlée sur la production de la parole chez les locuteurs coréens et français. *Actes des 33èmes Journées d'Études sur la Parole*, 2020, Nancy, France, pp. 82-90.

La thèse de Dayeon Yoon, que j'ai codirigée avec Cécile Fougeron et qui a été soutenue en juin 2024 (Yoon, 2024), a porté sur les spécificités de la voix et de la parole produites par les hommes et les femmes en fonction de leur morphologie corporelle et de la langue parlée, en comparant hommes et femmes français et coréens.

Pour les besoins de cette thèse, la doctorante a recueilli une quantité conséquente de données selon un protocole que nous avons défini conjointement avec elle. Outre la production d'une voyelle /a/ tenue et de glissandi ascendants et descendants et la lecture d'une phrase porteuse à la structure comparable entre coréen et français et incluant trois logatomes constitués de la syllabe /ma/ répétée deux à quatre fois et insérée dans différentes positions prosodiques, l'une des principales originalités de ce protocole a été d'intégrer une tâche de production de transitions entre deux articulations extrêmes de vocoïdes (passage de l'articulation la plus fermée possible analogue à l'articulation d'un /i/ à celle la plus ouverte possible analogue à l'articulation d'un /a/). Ces transitions, désignées par la doctorante par le terme de « pseudo-diphtongues », étaient destinées à explorer les limites de leur système de production, de façon aussi indépendante que possible de variations dépendantes de la langue ou de l'appartenance à un groupe social, et ainsi d'obtenir une estimation indirecte d'une partie des caractéristiques du conduit vocal des locuteurs à partir des mesures formantiques en l'absence de mesures articulatoires plus directes. Le protocole initial incluait également deux autres transitions ciblant le geste de protrusion et la variation maximale sur l'axe

antérieur/postérieur, mais leur interprétation par les locuteurs s'est avérée trop variable en dépit des exemples proposés par la doctorante et elles n'ont donc pas été retenues dans le jeu de données final.

Un total de 17 femmes et 16 hommes français tous francophones natifs, et de 20 femmes et 17 hommes coréens tous coréanophones natifs ont produit ce matériel selon un protocole d'enregistrement contrôlé adapté aux restrictions sanitaires en vigueur au moment de cette campagne d'enregistrement. L'ensemble des locuteurs enregistrés se considérant comme cisgenre, les questions de genre n'ont pas été abordées au-delà des comparaisons entre hommes et femmes. En complément des informations sur l'âge des locuteurs et d'éventuelles variantes dialectales, la taille et le poids des locuteurs au moment de l'enregistrement ont été recueillis comme estimations de leurs dimensions corporelles afin d'évaluer le lien éventuel entre ces dimensions corporelles et les caractéristiques de voix et de parole des locuteurs au sujet desquels les résultats de la littérature sont contrastés, et de disposer de valeurs de référence pour l'évaluation perceptive des dimensions corporelles des locuteurs par des auditeurs naïfs.

Outre la confirmation de corrélats acoustiques du dimorphisme sexuel largement documentés dans la littérature en termes de registre de fréquence fondamentale et de fréquences de résonance du conduit vocal, les analyses réalisées sur la syllabe /ma/ ont montré un degré de souffle et de bruit dans la voix plus important chez les femmes dans les deux langues (illustré par la Figure 53 pour la mesure de différence d'amplitude entre les deux premières harmoniques H1*-H2*, corrélat acoustique de la tension laryngée dont les valeurs sont généralement considérées dans la littérature comme plus élevées dans le cas d'une voix soufflée), mais avec une différence entre hommes et femmes plus importante dans le groupe français que dans le groupe coréen. Ces différences interlangues dans le marquage du dimorphisme sexuel, comparables à celles observées entre l'allemand et le suédois (Weirich et al., 2019), pourraient s'expliquer par des différences culturelles entre la société française et la société coréenne, avec l'émergence récente chez les jeunes coréennes d'un rejet des stéréotypes traditionnels de la féminité (Zhang et al., 2022), et d'autre part une préférence des auditeurs francophones pour les voix féminines soufflées voire rauques (voir par exemple Barkat-Defradas et al. (2012)). Par ailleurs les analyses acoustiques ont également montré que dans le groupe coréen, la dénasalisation de la consonne nasale /m/ en position initiale de groupe accentuel est plus marquée chez les hommes que chez les femmes, qui quant à elles réduisent plus la durée de ces consonnes initiales, conformément à l'hypothèse selon laquelle les femmes coréennes seraient à un stade plus avancé de la dénasalisation initiale en cours en coréen (Yoo & Nolan, 2020).

La taille et le poids des locuteurs ont été mis en relation avec un ensemble de mesures acoustiques relatives à la voix, extraites des productions par les locuteurs de la voyelle tenue /a/ ainsi que des glissandi pour estimer les limites de la modulation de la fréquence fondamentale. La taille et le poids ont également été confrontés à des mesures de distance dans l'espace acoustique des deux ou trois premiers formants, extraites des vocoïdes produits avec une transition entre fermeture maximale et ouverture maximale. Pour les hommes, ces analyses ont indiqué un lien entre la taille des locuteurs et la fréquence fondamentale ainsi qu'entre la taille et l'amplitude des différences formantiques entre articulations extrêmes, ainsi qu'un lien entre le poids et degré de périodicité du signal mesuré par le rapport entre énergie harmonique et bruit HNR. Ce dernier résultat est toutefois à nuancer du fait du lien entre taille et poids dans les données. De tels liens n'ont pas été observés pour les femmes, ce qui pourrait

être en partie lié au fait que la population étudiée comprenait moins de variation de taille et de poids pour les femmes que pour les hommes.

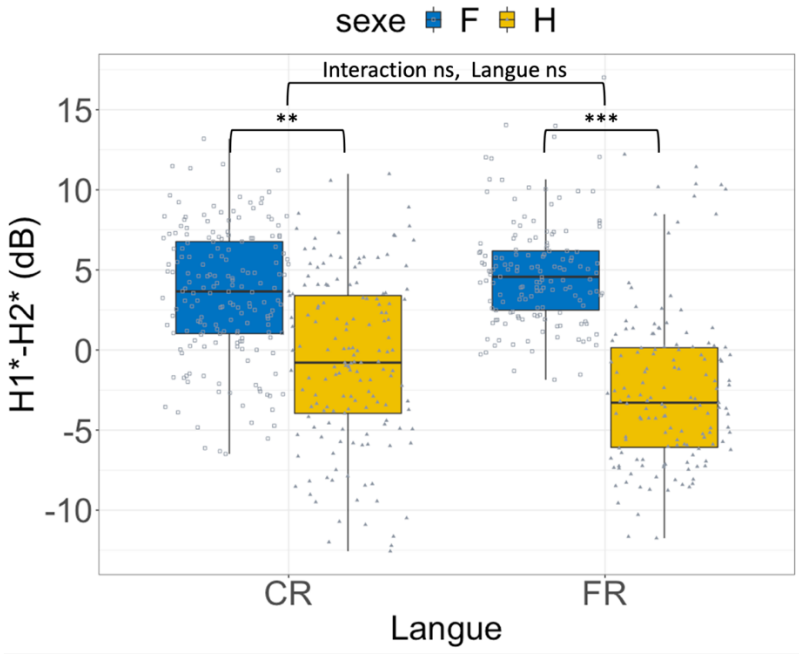


Figure 53 : Différence H1*-H2* entre l'amplitude corrigée de l'influence des formants des deux premières harmoniques, mesurée sur les occurrences de la syllabe /ma/ produites dans des logatomes insérés dans une phrase porteuse, et comparée entre hommes et femmes dans le groupe coréen (CR) et français (FR). D'après Yoon (2024).

Enfin, une étude perceptive dans laquelle 105 auditeurs francophones et coréanophones devaient estimer la corpulence et la taille des locuteurs à partir de l'écoute des logatomes a indiqué que les auditeurs étaient plus performants dans l'estimation de la taille des locuteurs dépassant une certaine taille, correspondant à la taille des hommes inclus dans l'étude. En revanche les auditeurs ne sont pas parvenus à estimer la corpulence des locuteurs. La mise en relation de la taille et de la corpulence estimées par les auditeurs et des caractéristiques acoustiques des stimuli suggère que ces estimations seraient influencées par les fréquences formantiques liées aux dimensions du conduit vocal, en lien avec les propositions d'Ohala sur le code fréquentiel (1984). En revanche la sensibilité des auditeurs à la hauteur mélodique et à la raucité de la voix était variable en fonction du sexe du locuteur.

6.3 Réalisations affaiblies du /v/ intervocalique

Publication associée :

[ACTI3] Dong, S., & **Audibert, N.** (2024). Caractérisation acoustique des réalisations approximantes du/v/ intervocalique en français spontané. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 13-22.

En collaboration avec Suyuan Dong dans le cadre de son mémoire de master que j'ai encadré et comme point de départ de son travail de thèse sur l'affaiblissement des obstruantes antérieures sonores, nous avons cherché à caractériser sur le plan acoustique les réalisations affaiblies de la consonne fricative labiodentale /v/ en français, qui en contexte intervocalique tend à être réalisée comme approximante, et à identifier certains des facteurs qui pourraient

expliquer la variation inter- et intra-individuelle de ces réalisations. Au-delà de notre contribution à la caractérisation des réalisations approximantes, l'objectif principal était d'évaluer dans quelle mesure ces réalisations affaiblies relèvent simplement d'un processus de réduction segmentale (voir par exemple Ernestus & Warner (2011)) ou pourraient être une manifestation de l'amorce d'un changement sonore en cours. Notre choix s'est porté sur une fricative voisée en raison de l'antagonisme aérodynamique entre leurs deux sources d'énergie que sont le voisement et le bruit de friction, qui les rend moins stables et favorise leur réalisation comme approximantes ou leur dévoisement lorsque l'une de ces deux sources d'énergie prend le pas sur l'autre (Johnson, 2002). Nous avons également opté pour une position intervocalique, faible en position interne de mot et donc susceptible de favoriser l'affaiblissement des réalisations du /v/ dans cette position (voir par exemple Brandão De Carvalho et al. (2008) sur les liens entre facteurs positionnels et lénition).

Bien que moins largement étudiées que les occlusives, l'acoustique des fricatives a fait l'objet d'un nombre conséquent de travaux, qui pour la plupart ont cherché à caractériser l'effet du lieu d'articulation sur leur réalisation (voir par exemple Maniwa et al. (2009)), à partir notamment des propriétés spectrales et de l'amplitude du bruit de friction, ainsi que des transitions formantiques dans les voyelles adjacentes. Quelques études se sont également penchées sur le rôle des hautes fréquences, notamment celle de Shadle et al. (2023) qui ont proposé un ensemble de métriques faisant intervenir les fréquences supérieures à 7000 Hz. L'acoustique des approximantes est moins largement documentée, néanmoins il est relativement bien établi que comparativement aux voyelles avec lesquelles elles partagent un certain nombre de propriétés, les approximantes ont une structure formantique plus faible et moins stable, avec une durée réduite (voir par exemple Reetz & Jongman (2020)).

Les analyses réalisées sur sont concentrées sur les productions de 10 femmes et 10 hommes du corpus de parole conversationnelle NCCFr (Torreira et al., 2010), ayant produit respectivement 2970 et 2534 occurrences de /v/ en position intervocalique. Après une première étape d'identification automatique de ces occurrences dans le corpus segmenté à l'aide d'un alignement forcé à partir de la transcription orthographique fine associée au corpus, l'ensemble des réalisations a fait l'objet d'une catégorisation à partir de l'écoute, l'inspection du signal acoustique et celle de l'information spectrographique, à la fois dans la bande de fréquences 0-5 kHz afin de prendre en compte la structure formantique, et dans la bande de fréquences 0-15 kHz afin de caractériser le bruit de friction. Outre la catégorisation des réalisations des /v/ intervocaliques comme fricative voisée ou approximante, les cas potentiels d'élision complète qui n'auraient pas été détectés par l'alignement forcé ont été pris en compte, de même que les cas de dévoisement dus à celui de l'une des voyelles adjacentes, et un ensemble de catégories pour lesquels les occurrences correspondantes ont été jugées inexploitables. La Figure 54 illustre la distribution observée pour les différents locuteurs, avec une importante variabilité inter-individuelle en termes de taux de réalisation comme fricative voisée ou approximante. Outre cette catégorisation, la phase d'annotation des données a inclus la vérification et éventuellement une correction manuelle des frontières segmentales, ainsi qu'un inventaire des indices acoustiques présents parmi une liste prédéfinie relative notamment à la structure et aux transitions formantiques, au bruit de friction, à la présence d'énergie en basse fréquence relative au voisement, à la périodicité et au degré de réduction de l'amplitude par rapport aux voyelles adjacentes.

En complément, la position prosodique associée à chacune des occurrences annotées a également été catégorisée parmi une liste prédéfinie. La Figure 55 illustre pour chacun des 20

locuteurs le taux de réalisation comme approximante observé parmi les occurrences de /v/ intervocalique en position finale accentuée. Comme on peut le voir, le taux de réalisations affaiblies en position forte est globalement plus important qu'attendu dans le cas d'une simple réduction segmentale qui tend à toucher principalement les positions prosodiques faibles en raison de l'effort articulatoire associé aux positions fortes (Cho, 2016), avec 37% de réalisations approximantes en position finale accentuée, et 11% en focus. On observe surtout une variation inter-individuelle très importante, avec selon les locuteurs de 2% à 71% des occurrences de /v/ en position finale accentuée réalisées approximantes. Comme l'illustre la mise en regard des deux figures, les locuteurs qui produisent le plus d'approximantes toutes positions confondues en produisent également une plus grande proportion en position forte.

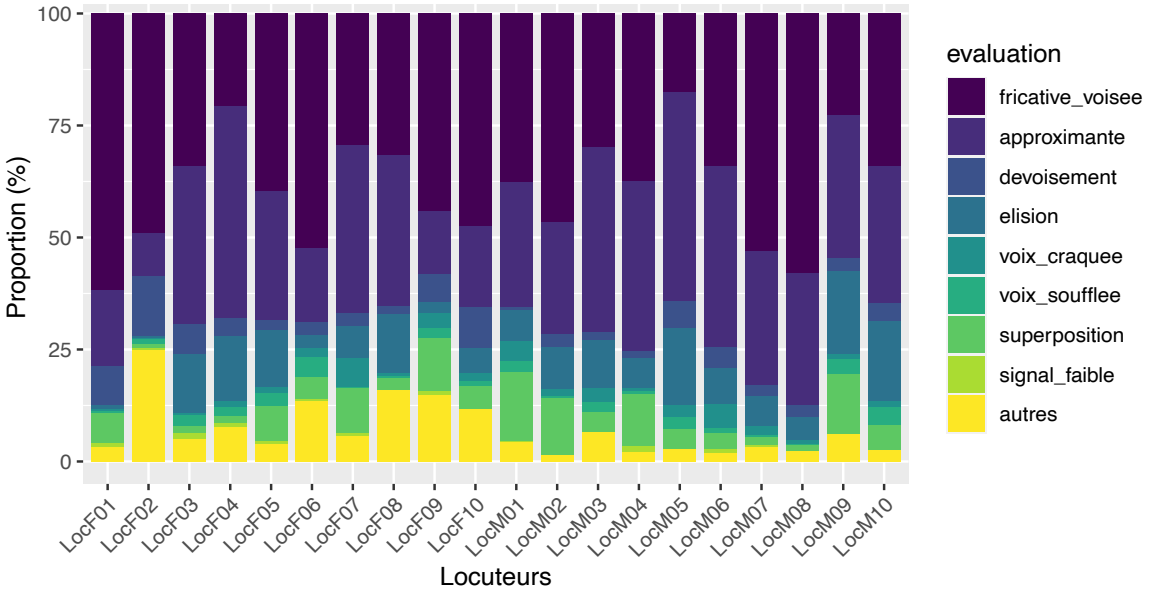


Figure 54 : Distribution par locuteur des taux de réalisation comme fricative voisée, approximante ou l'une des autres catégories considérées des 5504 occurrences de /v/ en position intervocalique produites par les 10 femmes (codes en LocF) et 10 hommes (codes en LocM) analysés. D'après Dong & Audibert (2024, [ACTI3]).

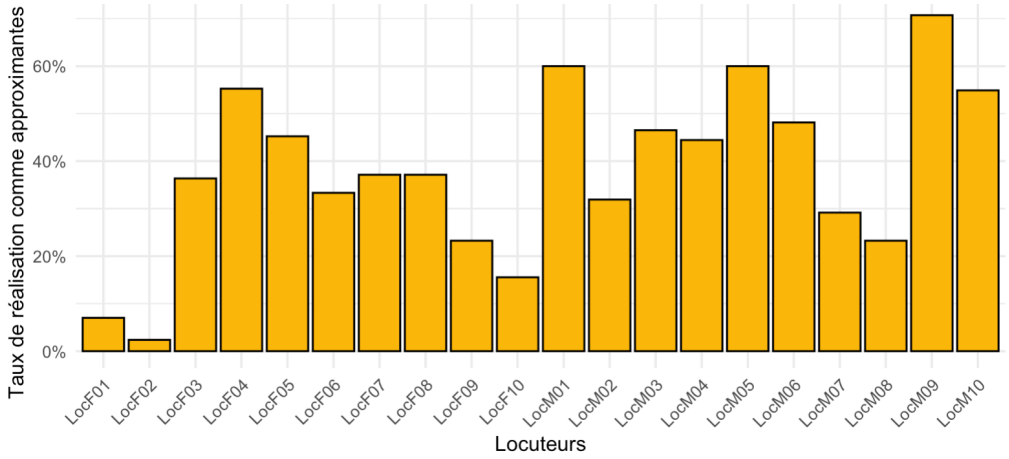


Figure 55 : Taux de réalisation comme approximante en position prosodique accentuée pour les 10 femmes (codes en LocF) et 10 hommes (codes en LocM) analysés. Adapté de Dong & Audibert (2024, [ACTI3]).

La comparaison des durées segmentales a indiqué une tendance des approximantes à être réalisées plus courtes que les fricatives voisées à la fois pour les femmes et les hommes, conformément aux données de la littérature sur les différences de durée intrinsèque entre approximantes et fricatives voisées, mais sans que les différences de durée n'expliquent les différences observées entre locuteurs ni ne distinguent fricatives voisées et approximantes au sein des réalisations d'un même locuteur.

Un ensemble de 30 mesures acoustiques ont été extraites sur 21 points équidistants entre le milieu de la voyelle précédant le /v/ et le milieu de la voyelle suivante, incluant la durée segmentale, les moments spectraux, l'énergie relative dans différentes gammes de fréquences, le rapport entre énergie harmonique et bruit HNR, les fréquences en Bark des formants F1 à F4, ainsi que les mesures en hautes fréquences spécifiques aux fricatives non-sibilantes proposées par Shadle et al. (2023). La trajectoire en fonction du temps de chacune de ces mesures, normalisées au préalable en z-scores afin de permettre leur comparaison directe, a été modélisée par un modèle mixte GAM afin de comparer entre types de réalisation (fricative ou approximante) et locuteurs ainsi qu'en fonction de la durée du /v/ et de celle des voyelles adjacentes, le contexte vocalique étant considéré comme un facteur aléatoire. Cette modélisation statistique a été effectuée séparément pour les hommes et les femmes. Par ailleurs l'enveloppe spectrale entre 0 et 14 kHz au milieu de la réalisation du /v/ a également été extraite, l'énergie acoustique étant modélisée séparément, également par un modèle GAM, en fonction de la fréquence et des mêmes facteurs.

La Figure 56 représente l'écart maximum entre fricatives voisées et approximantes tel que prédit par les modèles pour chacune des 30 mesures acoustiques normalisées. Outre la confirmation de l'intérêt de la prise en compte des hautes fréquences non seulement pour distinguer entre lieux d'articulations des fricatives (Kharlamov et al., 2023) mais également pour caractériser les différences acoustiques entre fricatives voisées et approximantes, les mesures les plus discriminantes entre les deux types de réalisation suggèrent que les approximantes seraient associées à une énergie spectrale moins diffuse et présentant moins d'énergie en hautes fréquences. Cette interprétation a été confirmée par la modélisation de l'enveloppe spectrale de chacun des deux types de réalisation, qui indique une énergie spectrale plus importante en moyennes et surtout hautes fréquences pour les fricatives voisées comparativement aux approximantes, avec un « creux » dans l'énergie spectrale autour de 6000 Hz pour les femmes et de 5000 Hz pour les hommes (voir Figure 57 pour les données des femmes). Les valeurs de centre de gravité spectral (CoG) seraient également compatibles avec une articulation moins antérieure dans le cas des approximantes, toutefois ce résultat doit être interprété avec prudence en raison des multiples facteurs, dont le voisement, susceptibles d'influencer les valeurs de CoG.

La modélisation de la dynamique des trajectoires des paramètres acoustiques entre le milieu de la voyelle précédente et le milieu de la suivante indiquent une divergence maximale entre réalisation comme fricative voisée ou comme approximante en un point temporel très proche du milieu de la réalisation du /v/, à l'exception des mesures d'intensité relative dans les bandes de fréquences les plus élevées pour lesquelles cette divergence maximale est légèrement plus tardive, avec toutefois une amplitude de la différence proche de celle relevée au milieu de la réalisation du /v/. Ce constat est donc encourageant dans l'optique de l'application de ces mesures à plus grande échelle, dans de très grands corpus pour lesquels seul un alignement forcé automatique est disponible.

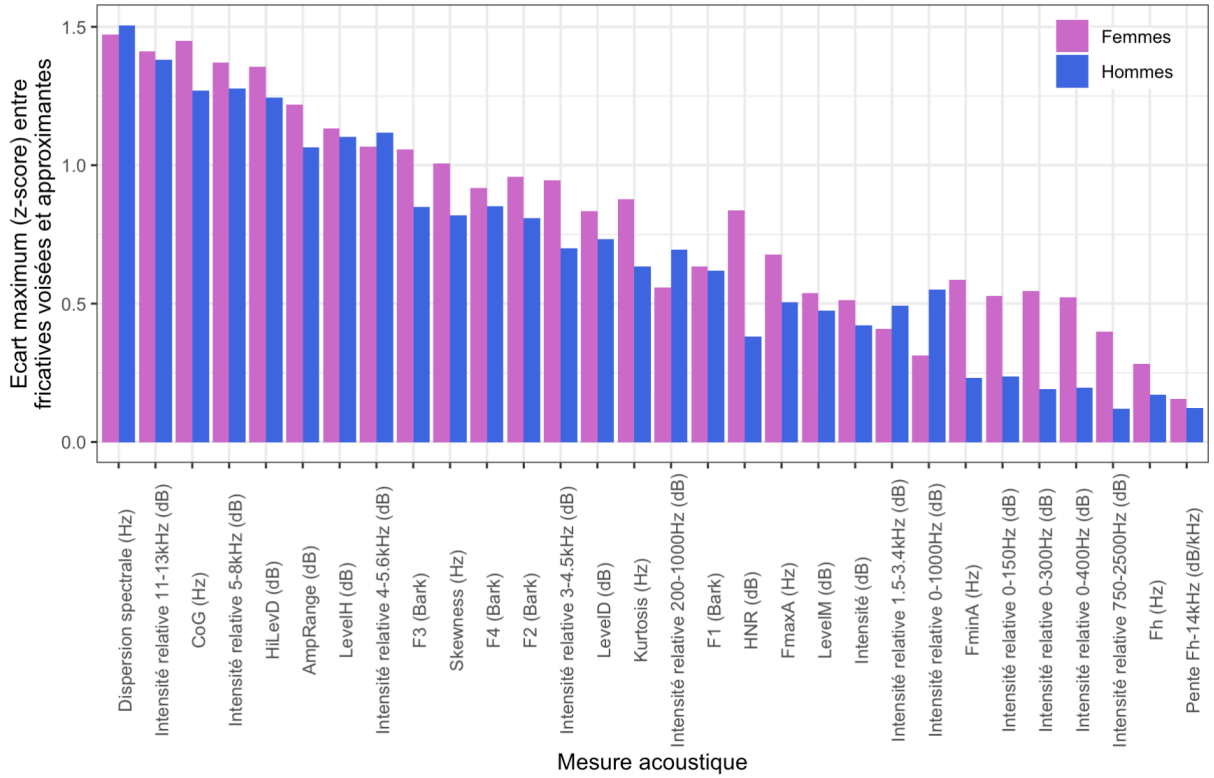


Figure 56 : Distance maximale en nombre d'écart-types observée entre les réalisations des /v/ comme fricative voisée et comme approximante à partir de la modélisation par GAM effectuée séparément pour les hommes et les femmes de la trajectoire de chacune des 30 mesures acoustiques après normalisation en z-scores. Les mesures acoustiques sont ordonnées de gauche à droite de la plus distinctive entre fricatives voisées et approximantes (en moyenne pour hommes et femmes) au sens de cette distance maximale jusqu'à la moins distinctive. D'après Dong & Audibert (2024, [ACTI3]).

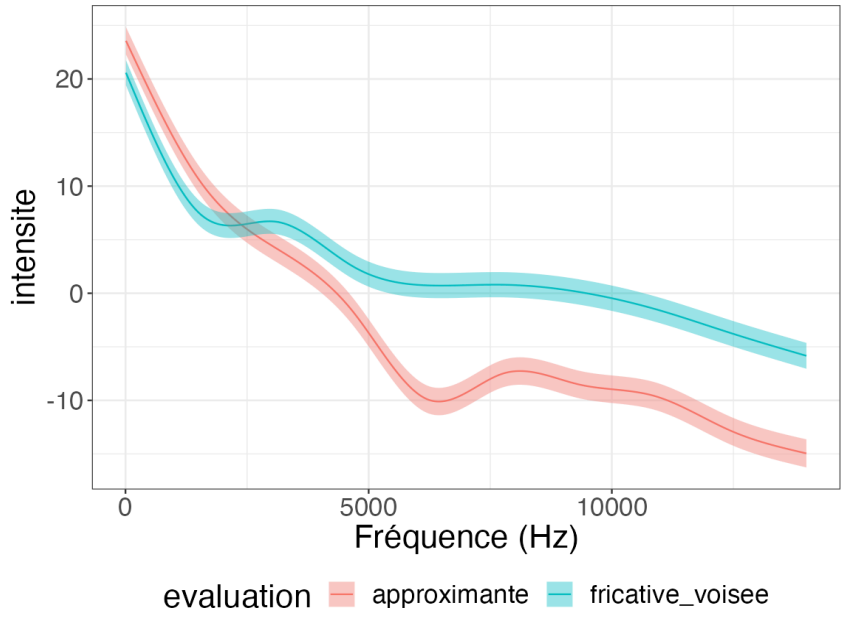


Figure 57 : Enveloppe spectrale entre 0 et 14000 Hz prédite par un modèle GAM représentant les spectres au milieu de la réalisation du /v/ pour les réalisations approximantes ou comme fricatives voisées des 10 femmes. Adapté de Dong & Audibert (2024, [ACTI3]).

6.4 Dévoisement final des obstruantes

Publication associée :

[ACTI26] 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*, pp.1726-1730.

En collaboration avec Adèle Jatteau, alors post-doctorante au laboratoire LIMSI (devenu ensuite LISN), ainsi que Ioana Vasilescu, Lorie Lamel et Martine Adda-Decker, j'ai contribué à une étude du dévoisement final des consonnes obstruantes en français standard à partir de l'analyse de grands corpus de parole continue. Le dévoisement final, attesté dans de nombreuses langues (Blevins, 2006), est souvent considéré dans la littérature comme un processus naturel pour lequel diverses explications ont été avancées comme l'anticipation de l'ouverture glottique pour permettre la respiration (Myers, 2012), la baisse de la pression sous-glottique en fin de l'énoncé qui interromprait le voisement de façon anticipée (Westbury & Keating, 1986), ou encore une conséquence de l'allongement final qui perturberait à la fois la production et la perception du voisement (Ohala, 1997; Blevins, 2006). L'objectif de cette étude a été d'évaluer à partir de l'analyse de grands corpus de parole continue en français standard l'hypothèse qui découle de ces constats selon laquelle un effet variable de dévoisement final devrait être observé dans des langues dans lesquelles la neutralisation finale n'est pas phonologisée. Tandis que le dévoisement final est déjà bien établi dans d'autres variantes régionales du français comme par exemple le français alsacien (Temple, 2000), le français standard est bien adapté pour évaluer cette hypothèse en raison du contraste de voisement en position finale de mots attesté dans de nombreuses paires minimales qui impliquent à la fois les occlusives et les fricatives et divers lieux d'articulation. Quelques études ont toutefois fait état d'une tendance minoritaire au dévoisement final également en français standard (Walter, 1977; Temple, 1999).

L'étude que nous avons menée s'est appuyée sur l'analyse du taux de voisement tel que défini par Snoeren et al. (2006), en tant que proportion de la durée de l'obstruante voisée au sens de la détection automatique de la fréquence fondamentale réalisée par l'algorithme de Boersma (1993), dans les obstruantes en position finale de mot de grands corpus de parole continue : le corpus de parole journalistique ESTER (Galliano et al., 2006), le corpus de débats ETAPE (Gravier et al., 2012) et le corpus de parole conversationnelle NCCFr (Torreira et al., 2010), segmentés automatiquement au préalable en phones à l'aide du système du LIMSI (Gauvain et al., 2002). Les obstruantes sonores en position finales de mot ont été extraites de la transcription en considérant à la fois les mots se terminant par un schwa (Cə#) ainsi que ceux dans lesquels l'obstruante se trouve en position finale absolue (C#). Après élimination des portions du corpus ESTER les plus susceptibles d'inclure des variantes non-standard du français en quantité conséquente, des mots fréquents comme « d' » composés d'une obstruante, des mots produits de façon incomplète, interjections et allomorphes ainsi que des segments particulièrement longs susceptibles de constituer des erreurs d'alignement, un total de 30 872 occurrences des consonnes /b, d, g, v, z, ʒ/ en position finale de mot ont été incluses dans l'analyse, dont 4 986 suivies d'un schwa. Ces obstruantes ont été classées en cinq catégories en fonction du segment initial présent dans le mot suivant, qui pouvait être une obstruante sourde ou sonore, une sonante, une voyelle ou une pause silencieuse. On peut noter que, comme cela est souvent le cas dans les travaux en phonétique de corpus, ces

catégories n'étaient pas réparties de façon homogène dans les données, les obstruantes avant une autre obstruante sourde ou sonore ou avant une voyelle étant plus nombreuses que celles avant sonante ou pause. Par ailleurs la proportion de schwas réalisés était plus importante avant pause.

En raison de la distribution observée des valeurs du taux de voisement v-ratio qui incluent de nombreuses obstruantes pour lesquelles ce taux est de 100% (voir Figure 58 pour le cas des obstruantes avant pause), les analyses statistiques ont été effectuées en comparant les obstruantes intégralement voisées à celles qui ne le sont que partiellement.

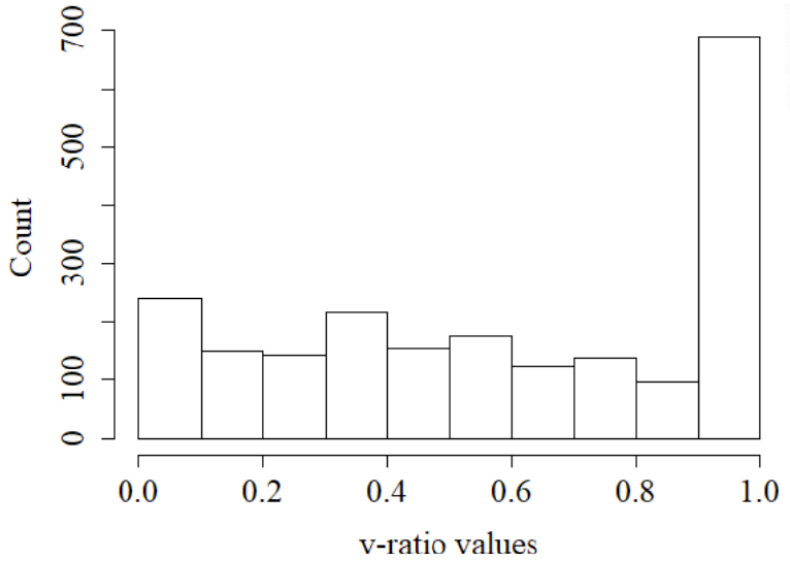


Figure 58 : Distribution des valeurs de la mesure de taux de voisement v-ratio parmi les 2129 obstruantes sonores en position finale absolue avant pause analysées. D'après Jatteau et al. (2019, [ACTI26]).

Plusieurs hypothèses ont été formulées à partir des données de la littérature. Tout d'abord, le premier postulat était que le dévoisement final serait plus fréquent avant pause que dans les autres contextes, conformément aux données de la littérature qui dans la plupart des cas concernent exclusivement le dévoisement en position finale absolue. Cette hypothèse a été validée, avec un taux d'obstruantes avant pause intégralement voisées qui n'est que de 52%, contre 74% à 80% avant un segment voisé. De plus, en n'incluant pas les cas qui comprennent un schwa final réalisé mais uniquement les obstruantes en position finale absolue, ce taux chute à 31%, soit un taux de voisement complet équivalent à celui observé avant une obstruante sourde de 24% qui s'explique par le processus d'assimilation déjà observé auparavant en français (Snoeren et al., 2006; Hallé & Adda-Decker, 2007), comme illustré par la Figure 59. Les analyses se sont concentrées sur le cas des obstruantes en position finale absolue avant pause.

La seconde hypothèse était que, les styles de parole moins formels étant plus propices à la variation, une proportion plus importante de dévoisement final avant pause serait observée dans le corpus de parole conversationnelle NCCFr, un niveau intermédiaire dans le corpus de débats ETAPE, et le niveau le plus faible dans le corpus de parole journalistique ESTER, supposé le plus formel des trois. Toutefois cette hypothèse n'a pas été vérifiée, avec une proportion d'obstruantes finales intégralement voisées plus élevée dans ETAPE (38%) que dans ESTER (28%) et NCCFr (30%). La proportion plus élevée de voisement intégral observée dans ETAPE

pourrait être liée à la nature des données, avec notamment des chevauchements fréquents de tours de parole dans les débats qui pourraient dans certains cas avoir biaisé la détection automatique de la fréquence fondamentale. Le niveau équivalent observé entre NCCFr et ESTER suggère quant à lui que la tendance au dévoisement avant pause pourrait être plus physiologique que dépendant du style de parole.

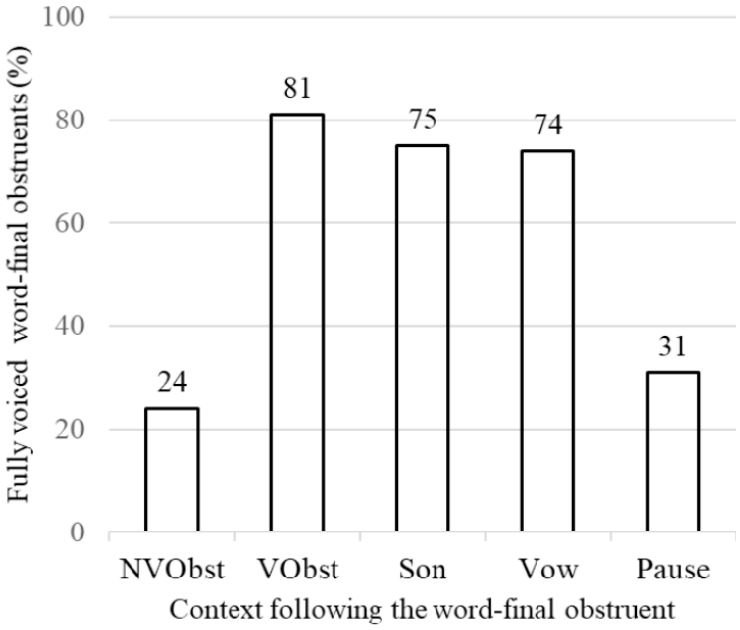


Figure 59 : Proportion d'obstruantes en position finale de mot intégralement voisées au sens de la mesure v-ratio, en fonction du contexte suivant. NVObst = obstruante sourde ; VObst = obstruante sonore ; Son = sonante ; Vow = voyelle ; Pause = pause silencieuse. D'après Jatteau et al. (2019, [ACTI26]).

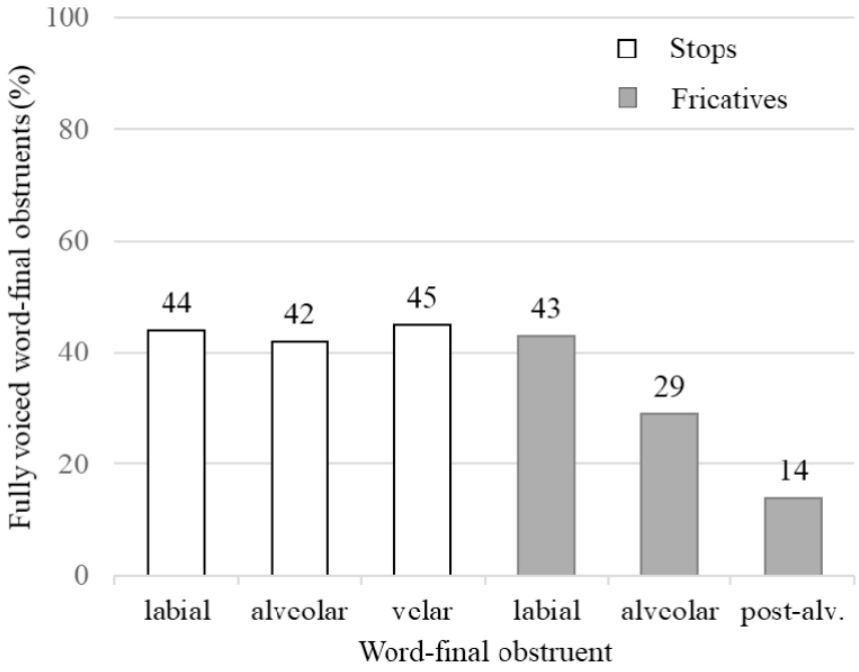


Figure 60 : Proportion d'obstruantes en position finale de mot intégralement voisées au sens de la mesure v-ratio, en fonction du mode et du lieu d'articulation. D'après Jatteau et al. (2019, [ACTI26]).

Les deux dernières hypothèses, liées à des contraintes physiologiques, concernait respectivement le mode et le lieu d'articulation : en raison de l'antagonisme entre degré élevé de pression intra-orale requis pour produire le bruit de friction et le voisement (Ohala, 1983b), les fricatives devraient être plus sujettes au dévoisement final que les occlusives. Enfin, les obstruantes postérieures devraient présenter un taux de dévoisement final supérieur à celui des antérieures, puisque la différence de pression de part et d'autre de la glotte est plus difficile à maintenir lorsque la distance entre celle-ci et la constriction est réduite (Ohala, 1997). Si comme attendu les fricatives sont bien plus largement dévoisées avant pause que les occlusives, l'effet attendu du lieu d'articulation n'est confirmé que pour les fricatives avec un dévoisement d'autant plus massif que la fricative est postérieure, tandis que toutes les occlusives montrent une proportion d'occurrences intégralement voisées équivalente quel que soit le lieu d'articulation, comme illustré par la Figure 60.

Suite à cette première analyse, en collaboration avec Adèle Jatteau nous avons étendu cette analyse en utilisant également le rapport de durée entre la consonne et la voyelle précédente comme indice du voisement, et surtout en ne considérant pas exclusivement les obstruantes sonores comme dans la plupart des travaux de la littérature mais le contraste entre sourdes et sonores à travers des mesures de taille d'effet obtenues via un modèle de régression bayésienne. Les résultats ainsi obtenus ont fait l'objet d'un article soumis à la revue *Laboratory Phonology*, qui fait l'objet de révisions mineures au moment de la finalisation de ce manuscrit.

6.5 Disjonction en parole continue

Publication associée :

[ACTI8] Jatteau, A., **Audibert, N.**, Adda-Decker, M., Lamel, L., & Bilinski, E. (2024). Le « h aspiré » à l'état sauvage : décrire la disjonctivité dans les grands corpus de parole naturelle. *Actes du 9ème Congrès mondial de linguistique française*, Lausanne, Suisse, Vol. 191, p. 09005.

En collaboration avec Adèle Jatteau du laboratoire lillois Savoirs, Textes, Langage et Martine Adda-Decker, nous avons cherché en nous appuyant sur des données alignées par des collègues du Laboratoire Interdisciplinaire des Sciences du Numérique (anciennement LIMSI) à décrire la disjonction en français telle qu'elle peut être observée dans la parole non-scriptée, au-delà des observations en laboratoire susceptibles d'être fortement influencées par la représentation de la norme.

Le phénomène de disjonction (de Cornulier, 1981), classiquement désigné en français comme « h aspiré » bien que ce terme ne corresponde pas à la réalité phonétique et phonologique, correspond aux mots à voyelle initiale qui bloquent le processus de sandhi externe en s'opposant notamment à la liaison ou à l'effacement du schwa, sans pour autant se comporter comme les mots à initiales consonantiques. Tandis que la majorité des auteurs (voir par exemple Côté (2008)) font le postulat de deux grandes catégories homogènes de mots disjonctifs ou jonctifs, la grande variabilité de réalisation observée entre éléments lexicaux mais aussi pour un même mot (Tranel & Del Gobbo, 2002) a conduit d'autres auteurs tels que Zuraw & Hayes (2017) à faire l'hypothèse d'un continuum de propension plus ou moins grande à la disjonctivité sur lequel les mots pourraient être situés à différents niveaux. Pour notre part, nous avons cherché à contribuer à ce débat à partir de l'analyse de productions en

contexte de mots susceptibles de donner lieu à une disjonction, afin de mieux identifier les facteurs qui peuvent influencer la réalisation de la disjonction.

Tandis que les études en laboratoire tendent à être biaisées par la force de la prescription qui provoque des hypercorrections de la part des locuteurs (Scheer, 2024), l'analyse dans une perspective de linguistique de corpus du phénomène de disjonction est complexifiée par sa rareté soulignée notamment par Göhring (2017). Afin de disposer d'un nombre d'exemplaires suffisant pour pouvoir donner lieu à des comparaisons, nous avons opté pour une analyse s'appuyant sur les données de plusieurs grands corpus de parole non-scriptée : le corpus de parole journalistique ESTER (Galliano et al., 2006), le corpus de débats ETAPE (Gravier et al., 2012) et le corpus de parole conversationnelle NCCFr (Torreira et al., 2010), soit un total de plus de 160 heures de parole. Ces données ont été segmentées automatiquement à l'aide du système de Gauvain et al. (2002), adapté pour considérer les réalisations jonctives ou disjonctives selon les principes de la méthode d'alignement avec variantes (Adda-Decker & Lamel, 2000).

Une première extraction a permis d'identifier dans ces corpus l'ensemble des bigrammes correspondants à un déterminant suivi d'un nom ou adjectif, ou à un pronom suivi d'un verbe, composés d'un premier mot dans lequel la liaison est fortement attendue (tel que « un », « les » ou « on »), le schwa peut ou non être prononcé (comme « le » ou « une ») ou encore qui prennent une forme différente devant une consonne ou une voyelle (comme « du », « au » ou « beau »), et d'un deuxième mot à voyelle initiale. Une vérification manuelle a été effectuée afin d'éliminer les contextes non-pertinents, aboutissant à un ensemble de 75 644 bigrammes dans lesquels l'absence de liaison entre les deux mots, la présence d'un schwa réalisé à la fin du premier mot ou une forme préconsonantique du premier mot (par exemple « du » plutôt que « de l' ») ont été codées comme correspondant à une disjonction.

Cette analyse a confirmé que la disjonction est un phénomène rare avec un total de 1 142 cas de disjonction dans les données étudiées (1,5% des bigrammes sélectionnés), qui ne permet pas une analyse de la variation entre locuteurs au-delà d'observations anecdotiques. Parmi les 3 567 lemmes différents présents en tant que deuxième mot du bigramme, 210 soit 5,7% du total présentent au moins une occurrence de disjonction, toutefois une grande partie de ces lemmes sont très peu fréquents. Huit occurrences de mots prononcés avec un [h] ou autre fricative initiale ont été éliminées suite à l'écoute des extraits, soit un total de 1 134 cas de disjonction pris en compte dans la suite de l'analyse, essentiellement qualitative en raison du faible nombre d'occurrences pour la plupart des mots.

Parmi les 35 noms communs et verbes exclusivement disjonctifs, la majorité comprennent un « h » graphique initial. On peut noter parmi les rares exceptions le terme « la une ». En revanche parmi les noms communs et verbes à « h » graphique initial majoritairement réalisés jonctifs on trouve des exceptions, notamment sur les mots « handicap » et ses dérivés ainsi que sur « hausse » et « hamburger ». Les formes exclusivement jonctives restent toutefois très largement majoritaires, y compris parmi les 167 noms communs et verbes à « h » graphique initial comme illustré par la Figure 61. Bien que parmi ces 167 lemmes, 109 soient représentés par moins de 5 occurrences, dont 57 occurrences uniques, la distribution bimodale observée est confirmée dans un sous-ensemble de données recentré sur les lemmes comptant un minimum de 20 occurrences, y compris en considérant l'ensemble des lemmes et non uniquement ceux à « h » graphique initial.

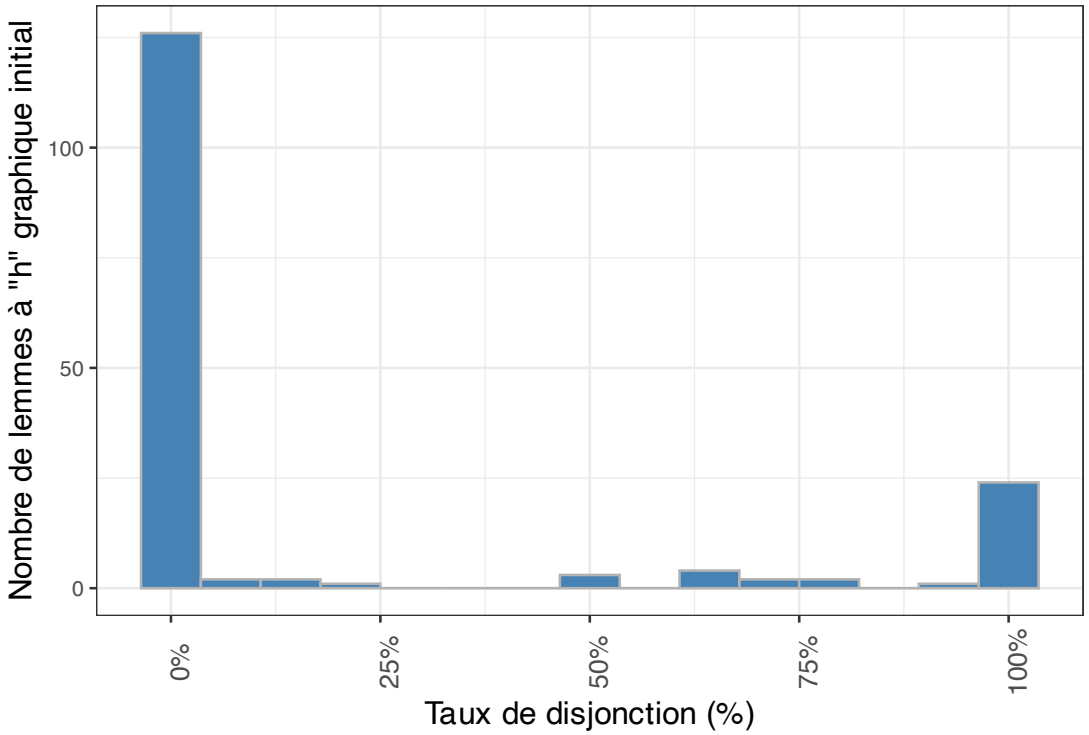


Figure 61 : Distribution du taux de réalisation disjonctive parmi les 167 noms communs et verbes à « h » graphique initial présents dans les corpus analysés. Adapté des données présentées dans Jatteau et al. (2024, [ACTI8]).

Parmi les lettres et acronymes, celles à consonne graphique initiale sont majoritairement disjonctives, avec toutefois un certain nombre d'exceptions relevées. Tandis que « onze » et « onzième » sont systématiquement disjonctifs, « un » en tant que numéral est jonctif dans 99,6% de ses nombreuses occurrences mais devient disjonctif dans certains contextes. La plupart des noms propres sont traités de façon catégorielle en étant réalisés systématiquement disjonctifs ou systématiquement jonctifs sans que des facteurs explicatifs de ce qui motive cette catégorisation puissent être identifiés, hormis une légère tendance à traiter comme disjonctifs les noms étrangers.

Enfin, quelques cas de disjonctions ont été observés sur des mots très fréquents, majoritairement jonctifs mais qui présentent également quelques cas marginaux de disjonction, qui d'après l'inspection des extraits correspondants semblent pouvoir s'expliquer par des facteurs pragmatiques, avec des cas d'emphase et d'autres qui correspondent plutôt à des cas de disjonction dus à une hésitation. Par ailleurs certains mots jonctifs peuvent devenir disjonctifs lorsqu'ils font partie d'un syntagme large.

Les données recueillies ont par ailleurs été confrontées aux résultats de Göhring (2017), qui a montré que le taux de disjonction dépend du type de sandhi. La comparaison via une régression binomiale de l'effet du type de sandhi sur le taux de disjonction a indiqué un taux de disjonction plus élevé pour les clitiques dissyllabiques (« une » ou « cette » pour lesquels la disjonction correspond à la réalisation du schwa), ce qui ne semble pas pouvoir s'expliquer par des variations dialectales, que pour la liaison et pour les clitiques monosyllabiques. Ces résultats qui vont à l'encontre de ceux de Göhring (2017) pourraient s'expliquer par des variations pragmatiques et prosodiques qu'il n'a pas été possible de contrôler dans la base de données. L'analyse de l'effet du premier mot sur le taux de disjonction a montré que le premier

mot « de » (ou « d’ » considéré conjointement) est significativement moins propice à la disjonction que « le » et les formes associées, conformément aux résultats de Zuraw & Hayes (2017). En raison des observations précédentes, ce résultat doit toutefois être considéré avec précaution.

Le résultat principal de cette étude, qui parmi les travaux auxquels j’ai contribué après ma thèse est sans doute la plus éloignée du signal de parole, est donc la tendance de la disjonctivité à être traitée de façon catégorique, chaque élément lexical étant majoritairement considéré comme exclusivement jonctif (ce qui constitue le cas de très loin le plus fréquent) ou exclusivement disjonctif.

7 Méthodes, données et outils pour la recherche

Résumé du chapitre 7

Dans ce chapitre je regroupe mes principaux apports méthodologiques en matière de développement ou d'évaluation de méthodes instrumentales et acoustiques, de recueil de corpus et de développement d'outils.

Une évaluation de systèmes d'acquisition de données fibroscopiques a conclu que seuls les systèmes avec caméra ultra-rapide offrent une finesse de synchronisation suffisante entre son et image pour l'analyse quantitative. De nouvelles mesures de nasalance obtenues à partir d'accéléromètres et de microphones ont été proposées pour pallier les limites des mesures classiques et permettre l'application aux voyelles nasales et à la parole continue. Une analyse ePGG de la production de séquences de consonnes sourdes en français a confirmé les données de la littérature sur des langues non-apparentées. Une série d'études sur l'évaluation de la capture optique de mouvements appliquée à la mesure de l'articulation labiale a conclu que les mesures de protrusion obtenues par cette méthode permettaient de rendre compte de l'opposition d'arrondissement en français, mais que les mesures d'aire aux lèvres ainsi obtenues ne reflétaient que partiellement les mouvements du contour interne. J'ai également contribué à la conception d'une plateforme multicapteurs mobile.

L'évaluation de l'alignement forcé appliqué à la parole dysarthrique a conclu à un degré d'imprécision comparable à celui observé pour des locuteurs témoins dans le cas de dysarthries légères, et à la robustesse des mesures prises au milieu des voyelles et des fricatives également dans le cas de dysarthries sévères. Une méthode fondée sur l'inspection des distributions a été proposée pour affiner la détection de la fréquence fondamentale. Des métriques destinées à capturer séparément à l'échelle de l'exemplaire la centralisation de l'espace vocalique, la variabilité intra-catégorie et le degré de recouvrement acoustique entre catégories ont été proposées. Une méthode de délexicalisation d'énoncés préservant partiellement les informations spectrales de qualité de voix en complément de la durée et de la fréquence fondamentale a été proposée et évaluée.

J'ai également contribué à la conception et au recueil d'un corpus de parole dysarthrique, de corpus multisessions destinés à l'étude de la variabilité inter- et intra-locuteur, d'un corpus d'expressions hostiles dans la parole politique, et d'un corpus de parole conversationnelle en chinois mandarin incluant une condition de contrôle via la relecture d'une partie des productions.

Je clos ce chapitre en présentant les outils en ligne que j'ai développé pour la recherche, destinés à l'exploration interactive de données quantitatives représentées sous forme d'histogrammes ou de nuages de points, et au calcul de métriques pour caractériser l'organisation de l'espace vocalique, après avoir exposé mes motivations pour le développement de ce type d'outils.

7.1 Mesures articulatoires pour la recherche en phonétique

7.1.1 Évaluation de la synchronisation de systèmes d'acquisition audio-visuelle

Publication associée :

[OS1] **Audibert, N.**, Amelot, A., Maeda, S., & Crevier Buchman, L. (2014). Évaluation de systèmes d'acquisition audio-vidéo pour la phonétique clinique. In Sock, R., Vaxelaire, B., Fauth, C. *La voix et la parole perturbées*, Travaux en Phonétique Clinique, Editions du CIPA, Collection "Recherches en PArole", pp.145-156.

En collaboration avec Angélique Amelot, Shinji Maeda et Lise Crevier Buchman, j'ai évalué la finesse de la synchronisation temporelle du flux vidéo et du flux audio de plusieurs systèmes d'acquisition audiovisuelle utilisés pour la prise de données fibroscopiques et endoscopiques en milieu hospitalier. Dans le cadre de la pratique clinique en phoniatrie, ces systèmes d'acquisition sont utilisés pour observer les structures anatomiques et les mouvements du voile du palais et du larynx. Si cette observation directe des structures internes impliquées dans la production de la parole est indispensable pour l'investigation clinique, dans un usage courant elle ne requière pas une synchronisation extrêmement fine entre le flux vidéo et le flux audio. Toutefois, pour la recherche en phonétique clinique et plus largement pour de nombreuses applications de mesures articulatoires à la recherche en phonétique, une mise en correspondance directe des informations visuelles et acoustiques est nécessaire, ce qui nécessite une finesse de synchronisation entre les deux flux de données d'autant plus importante que la résolution temporelle du flux vidéo est élevée et que la vitesse de mouvement des organes étudiés est importante. Cette question s'avère donc cruciale lorsque ces systèmes d'acquisition audiovisuels sont appliqués à l'étude du fonctionnement du larynx pendant la phonation, pour laquelle l'utilisation de caméras ultra-rapides prend toute son importance (Chevaillier et al., 2010).

Quatre dispositifs d'acquisition audiovisuelle ont été évalués, les quatre caméras étant équipées de capteur CCD. Ces dispositifs incluaient d'une part deux unités de consultation ORL couramment utilisées à des fins de diagnostic, équipées de caméras couleur haute-définition avec une fréquence d'échantillonnage vidéo de 25 trames par seconde et une fréquence d'échantillonnage audio de 22 050 Hz. En complément, deux stations d'acquisition équipées de caméras ultra-rapides dédiées à la recherche sur la parole ont également été évaluées, l'une avec caméra noir et blanc à 500 trames par seconde et une résolution de 256x256 pixels permettant l'acquisition de 4 secondes consécutives, la seconde étant équipée d'une caméra couleur pouvant aller jusqu'à 4000 trames par seconde et d'une résolution de 256x256 pixels permettant l'acquisition de 2 secondes consécutives. La fréquence d'échantillonnage audio de ces deux stations était de 44 100 Hz.

La finesse de synchronisation audiovisuelle des dispositifs d'acquisition a été évaluée au moyen d'un boîtier électronique conçu et assemblé par Shinji Maeda. Ce boîtier équipé d'un bouton-poussoir a permis de générer de façon synchrone un flash lumineux de 80 ms (diode électroluminescente) couplé à l'émission d'un son pur de fréquence 4000 Hz et de même durée. En l'absence d'un système d'acquisition audiovisuelle de référence d'une résolution très élevée, il n'a pas été possible de mesurer précisément la résolution temporelle intrinsèque du boîtier, toutefois celle-ci pouvant être estimée comme très inférieure à la milliseconde, son impact sur les mesures de décalage entre audio et vidéo a été considéré comme négligeable.

Une série de bips générés par le boîtier a été enregistrée avec chacun des systèmes évalués. Le décalage entre les trames vidéo et le signal audio a ensuite été calculé pour chacun de ces bips en mettant en correspondance les instants de début du signal audio et du signal lumineux générés par le boîtier.

Dans le cas des caméras ultra-rapides, la précision de la mesure a été limitée par le caractère progressif de l'apparition du signal lumineux. En effet bien que très rapide, l'allumage de la diode électroluminescente n'était pas instantané, en conséquence pour une fréquence d'échantillonnage vidéo de 4000 trames par seconde deux à trois trames vidéo consécutives pouvaient correspondre à l'apparition du signal (Figure 62). En considérant trois trames d'incertitude, l'incertitude temporelle estimée était donc de 0,75 millisecondes. Cette incertitude de mesure reste toutefois sensiblement inférieure à la durée d'un cycle laryngé, même dans le cas d'un registre particulièrement élevé. A titre de comparaison, la fréquence fondamentale la plus élevée produite par une chanteuse soprano dans l'étude de Sundberg (1975) était de 698 Hz, ce qui correspond à un cycle laryngé de 1,43 ms.

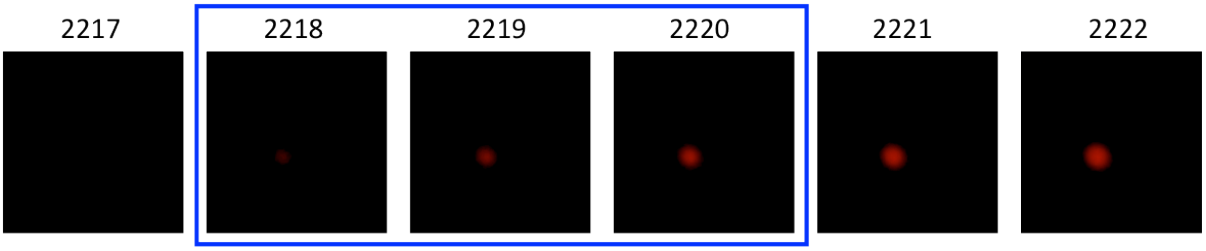


Figure 62 : Exemple de trames vidéo consécutives capturées avec une caméra ultra-rapide d'une résolution de 4000 trames/secondes. Le cadre bleu indique les trois trames pouvant être considérées comme étant celle d'apparition du signal lumineux en fonction du critère retenu. D'après Audibert et al. (2014, [OS1]).

Pour les quatre systèmes évalués, une fluctuation du décalage entre audio et vidéo a été relevée, ce qui rend impossible l'application d'une correction simple correction permettant de compenser ce décalage. L'objectif principal étant d'évaluer l'impact de l'incertitude affectant la synchronisation de chacun des systèmes évalués, le décalage maximal relevé a été retenu comme valeur de référence. Pour les deux unités de consultation ORL évaluées ce décalage maximal est conséquent et atteint même 120 ms pour l'unité KayPENTAX modèle 9295 dont l'usage est particulièrement répandu en phoniatrie. Pour l'étude du larynx et même en considérant le cas particulièrement favorable d'une voix d'homme adulte avec une fréquence fondamentale basse, cela correspond à une incertitude d'au moins 10 cycles laryngés. A l'inverse, les deux systèmes associés à une caméra ultra-rapide ont été évalués comme offrant une finesse de synchronisation bien meilleure, de 4 ms pour le système avec une résolution vidéo de 500 trames par secondes et n'excédant pas l'incertitude d'estimation d'apparition du signal lumineux de 0,75 ms pour celui avec une résolution vidéo de 4000 trames par secondes.

Ces résultats nous ont conduit à formuler des recommandations concernant l'usage des systèmes d'acquisition audio-visuelle pour la recherche en phonétique, à commencer par une mise en garde quant à l'usage d'unités de diagnostic pour le recueil de données destinées à l'analyse. De telles unités dont l'usage est beaucoup plus aisé que celui des stations d'acquisition équipées de caméras ultra-rapides demeurent extrêmement précieuses, non seulement à des fins de diagnostic mais aussi à des fins d'illustration qualitative en raison de la qualité d'image qu'elles permettent d'obtenir sans les contraintes d'éclairage et la capacité

plus limitée de stockage associées aux caméras ultra-rapides. Néanmoins, l'interprétation du lien entre image et audio peut s'avérer risquer du fait du décalage variable et potentiellement conséquent entre les deux flux. Une solution possible pour pallier ce problème, qui toutefois présente l'inconvénient de compliquer l'utilisation de ces unités, consisterait en la génération régulière de signaux brefs (généralement désignés par le terme de *triggers*) pilotés par une horloge externe et enregistrés simultanément dans le flux audio et vidéo.

7.1.2 Mesures de nasalance

Publication associée :

[ACTI45] **Audibert, N.**, & Amelot, A. (2011). Comparison of nasalence measurements from microphones and accelerometers and implications for phonetic analysis of nasality. *Proceedings of the 15th International Conference on Speech Communication and Technology (INTERSPEECH 2011)*, Florence, Italie, pp. 2825-2828.

En collaboration avec Angélique Amelot, je me suis penché sur l'évaluation comparée de mesures articulatoires non-invasives de nasalance. En effet, la complexité des corrélats acoustiques de la nasalité est bien établie en raison notamment de la présence d'anti-formants (Fujimura, 1962), ce qui conduit souvent les chercheurs en phonétiques à se concentrer sur l'analyse des segments oraux, ou à opter pour des mesures plus directes mais aussi plus invasives de l'abaissement du vélum ou du débit d'air nasal.

De nombreux instruments associés à diverses métriques ont été proposés dans la littérature afin de séparer le signal acoustique oral du signal acoustique nasal et ainsi estimer le degré de « nasalance », généralement considéré comme l'intensité relative du signal nasal par rapport au signal oral ou laryngé. Pour cette étude nous sommes concentrés sur les mesures de ces signaux à l'aide de microphones ou d'accéléromètres piézoélectriques. La plupart des métriques de la littérature ont été développées avec l'objectif de caractériser les productions dans des langues n'incluant pas de voyelles nasales (majoritairement l'anglais et le japonais) de locuteurs souffrant d'hypo ou d'hyper-nasalisation (voir par exemple Horii (1983)).

L'un des objectifs de notre étude était d'évaluer les différentes configurations instrumentales et les métriques associées sur des locuteurs sains d'une langue comme le français comportant à la fois des voyelles nasales et des consonnes nasales. Le second objectif était de déterminer la meilleure configuration expérimentale minimale pouvant être utilisée pour des études de terrain. Pour cela nous nous sommes appuyés sur des enregistrements contrôlés produits par deux hommes et deux femmes francophones natifs, avec un corpus produit deux fois par chaque locuteur, composé de voyelles tenues isolées incluant l'ensemble de l'inventaire des voyelles orales et nasales du français standard. La deuxième partie était composée de 30 logatomes de la forme CVCVCVC, avec C={t,d,n,s,z} et V={a,i,u,ɑ̃,ɛ̃,ɔ̃}. La troisième partie du corpus était composée de 30 mots français avec une séquence de deux consonnes incluant la consonne nasale /m/ ou /n/, comme par exemple le mot *apnée* (/apne/).

Comme illustré par la Figure 63, les locuteurs enregistrés ont été équipé d'un microphone oral (Vm), d'un microphone nasal inséré dans une narine (Nm), et de deux paires d'accéléromètres piézoélectriques fixés au moyen de bande adhésive double-face. L'une de ces paires d'accéléromètres piézoélectriques était fixée les arêtes latérales du nez (Na), et l'autre sur les côtés du larynx du locuteur (Va) afin de capturer les vibrations laryngées.

Afin de permettre d'obtenir des mesures comparables entre locuteurs, l'enregistrement a été précédé d'un processus minutieux de calibration des différents signaux, en prenant comme référence un /m/ tenu, supposé correspondre au maximum de nasalité produit par les locuteurs, et un /a/ tenu pour obtenir une mesure de l'amplitude orale maximale.

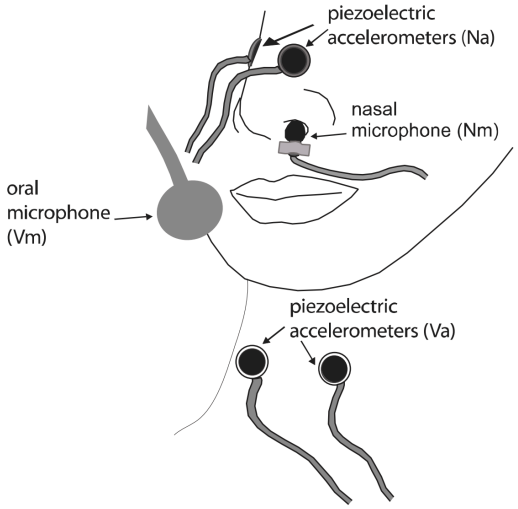


Figure 63 : Représentation schématique du dispositif expérimental retenu pour le recueil de mesures de nasalance, avec l'emplacement des deux microphones et des deux paires d'accéléromètres piézoélectriques sur le visage et le cou du locuteur (d'après Audibert & Amelot, 2011, [ACTI45], adapté de Vaissière et al. (2010)).

Quatre mesures de la littérature fondée sur des rapports entre l'énergie RMS mesurées sur certains de signaux recueillis ont été incluses dans l'évaluation. Deux de ces mesures s'appuient uniquement sur les signaux issus des microphones : la mesure TONAR (Fletcher, 1970) et la mesure de nasalance de Dalston et al. (1991). Les deux autres combinent le signal mesuré par les accéléromètres nasaux et soit celui mesuré par les accéléromètres laryngés avec la mesure HONC (Horii, 1980), soit celui du microphone oral avec la mesure N/V (Horii, 1983).

En complément nous avons proposé deux nouvelles mesures visant à quantifier le degré de nasalisation indépendamment du niveau d'énergie afin de dépasser la principale limite des autres mesures qui reposent sur un rapport entre niveaux d'énergie RMS et qui peuvent être fortement biaisées dans le cas des signaux de faible intensité. Ces deux mesures, notées respectivement LND (*Laryngeal Nasal Difference*) et OND (*Oral Nasal Difference*), consistent en la différence entre l'énergie RMS des accéléromètres nasaux normalisée pour tenir compte d'une éventuelle imprécision du processus de calibration, et respectivement l'énergie RMS des accéléromètres laryngés pour la mesure LND ou l'énergie RMS du microphone oral pour la mesure OND. Un autre avantage attendu de ces nouvelles mesures par rapport aux mesures classiques de nasalance est qu'elles permettent plus directement de définir un seuil à partir duquel les segments sont considérés comme nasals.

La Figure 64 illustre la distribution des valeurs moyennes normalisées en z-score attribuées à chaque classe phonémique pour chacune des six mesures évaluées. Si l'ensemble des mesures attribuent bien une valeur élevée aux consonnes nasales, seules les nouvelles mesures de différence entre signal nasal et signal oral permettent d'attribuer une valeur plus élevée aux voyelles nasales qu'aux segments oraux, ce qui les rend plus adaptées à l'application

à des langues incluant des voyelles nasales. On peut également noter que les mesures fondée sur des rapports tendent à attribuer une valeur excessivement élevée aux segments sourds, tout particulièrement les occlusives sourdes dans le cas de la mesure N/V, ce qui confirme les recommandations de la littérature de n'utiliser de telles mesures que sur des segments voisés (Redenbaugh & Reich, 1985). A l'inverse, cette première étude a confirmé l'applicabilité des nouvelles mesures LND et OND proposées à des segments sourds et par extension à des portions silencieuses, ouvrant la porte à l'utilisation de ces mesures sur la parole continue. De plus et bien que cela n'ait pas été évalué directement, l'inspection qualitative de l'évolution au cours du temps des valeurs prises par les six mesures sur notre corpus et leur comparaison aux données recueillies dans d'autres études sur le français suggère que les mesures LND et OND reflèteraient mieux le décours temporel de la nasalité que ne le font les mesures de nasalance exprimées sous forme de rapport.

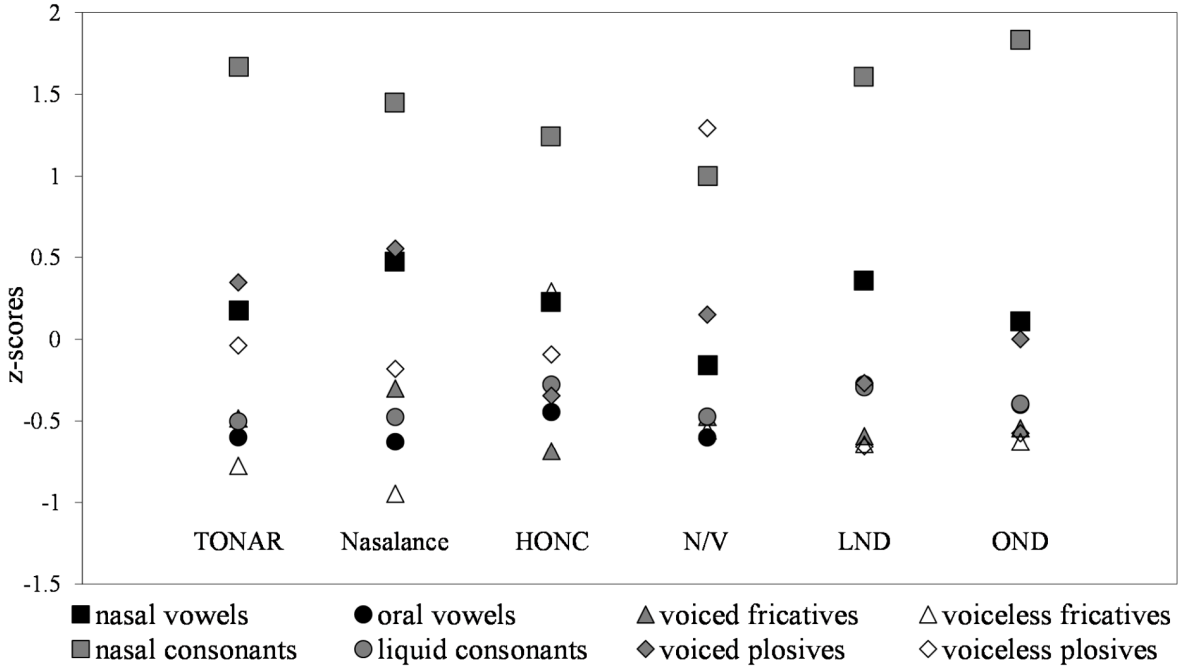


Figure 64 : Distribution des valeurs moyennes normalisées en z-scores des six mesures de nasalité, sur chacune des classes phonémiques représentées dans le corpus de logatomes, pour les quatre locuteurs regroupés (d'après Audibert et Amelot, 2011, [ACTI45]).

Néanmoins, ces nouvelles mesures se montrent particulièrement sensibles au processus de calibration et aux variations interindividuelles qui peuvent concerner à la fois la production des /m/ et /a/ tenus utilisés comme référence mais aussi probablement certaines spécificités individuelles liées à la nasalité au niveau physiologique. Dans l'analyse complémentaire de la capacité des différentes mesures à séparer entre segments nasals et oraux que nous avons effectuée séparément entre consonnes et voyelles et en tenant compte des différences individuelles, ceci s'est traduit par une incapacité des mesures LND et OND à distinguer entre voyelles nasales et orales pour l'un des deux locuteurs masculins enregistrés.

7.1.3 Mesures d'ouverture glottique par électrophotoglottographie (ePGG)

Publication associée :

[ACTI44] Ridouane, R., **Audibert, N.**, & Nguyen, V.M. (2012). Les ajustements laryngaux en français. *Actes des 19èmes Journées d'Études sur la Parole (JEP 2012)*, Grenoble, France, pp. 249-256.

En collaboration avec Rachid Ridouane et Van minh Nguyen, j'ai participé à une étude visant à comparer le degré d'ouverture glottique et son évolution dans le temps en fonction des types de segments produits en français et de leur contexte, afin de mettre ces résultats en relation avec ceux de la littérature concernant d'autres langues dans lesquelles les ajustements de l'ouverture glottique étaient mieux documentées. Les mesures d'ouverture glottique ont été réalisées à l'aide d'un photoglottographe externe (ePGG), instrument développé au Laboratoire de Phonétique et Phonologie (Honda & Maeda, 2008) et sur lequel Van minh Nguyen travaillait alors en tant qu'ingénieur d'étude contractuel spécialisé en électronique en vue d'optimiser son fonctionnement. Mes principaux rôles dans ce projet ont été de contribuer à l'élaboration du protocole expérimental, de proposer une méthode semi-automatisée via un ensemble de scripts pour le post-traitement des signaux ePGG, leur annotation via Praat et l'extraction de mesures d'ouverture glottique ainsi qu'une quantification de l'ouverture maximale et de sa position, et enfin de procéder à l'analyse statistique des résultats.

Les données recueillies ont été produites par deux locuteurs francophones natifs produisant chacun cinq à six occurrences d'un ensemble de huit mots et dix séquences de trois mots insérés dans une phrase porteuse, chaque mot ou séquence de mots incluant une obstruante sourde cible en position intervocalique ou une séquence de deux obstruantes sourde séparées par une frontière de mot. Certaines de ces séquences consistaient en deux occurrences consécutives de la même obstruante sourde (par exemple « *une nef fabuleuse* » pour la séquence cible f#f) afin d'introduire une condition de pseudo-gémination permettant une comparaison avec les données de langues avec gémination. Afin d'estimer de façon robuste l'ouverture glottique et la rendre aussi comparable que possible entre locuteurs indépendamment des différences morphologiques entre locuteurs et d'éventuelles variations de position du photoglottographe externe pendant les sessions d'enregistrement, le pic d'ouverture mesuré dans la séquence s#s (considérée comme correspondant au degré maximal d'ouverture glottique et de stabilité) a été retenu comme référence.

Les résultats obtenus ont été consistants avec ceux de la littérature sur d'autres langues non apparentées, notamment l'allemand (Hoole, 2006) et le berbère (Ridouane, 2003). En effet, ils ont montré que de façon consistante pour les deux locuteurs étudiés l'ouverture maximale glottique était plus importante pour les fricatives que pour les occlusives, et que pour les occlusives cette ouverture maximale glottique était plus importante pour les vélaires que pour les autres lieux d'articulation. L'analyse conjointe des stimuli incluant une simple obstruante intervocalique et de ceux correspondant à la condition de pseudo-gémination a indiqué une corrélation positive entre la durée de l'obstruante et le degré d'ouverture maximale glottique, légèrement plus importante pour les fricatives que pour les occlusives. Enfin, l'analyse de la localisation du pic d'ouverture glottique maximale dans les séquences occlusive-fricative et fricative-occlusive ne suivent pas les prédictions de Browman & Goldstein (1986) d'un alignement au milieu de la fricative mais varient en fonction de la position de la fricative dans la séquence.

7.1.4 Mesure de l'articulation labiale par capture de mouvement et vidéo

Publications associées :

[ACTI36] Georgeton, L., & **Audibert, N.** (2014). Mesures de protrusion par capture optique de mouvements : quelle métrique est la plus représentative de l'opposition d'arrondissement en français ? *Actes des 30èmes Journées d'Études sur la Parole*, Le Mans, France, pp. 239-247.

[ACTI41] Georgeton, L., & **Audibert, N.** (2013). Is protrusion of French rounded vowels affected by prosodic positions? *Proceedings of Interspeech 2013*, Lyon, France. pp. 3547-3551.

[ACTI43] Georgeton, L., & **Audibert, N.** (2012). Variations de la configuration labiale des voyelles /i, y, a/ : effets de la position prosodique et du locuteur. *Actes des 19èmes Journées d'Études sur la Parole (JEP 2012)*, Grenoble, France, pp. 465-472.

7.1.4.1 Contexte et système utilisé

En collaboration avec Laurianne Georgeton dans le cadre de sa thèse sur l'influence de la position prosodique initiale sur la réalisation acoustique des voyelles, nous avons cherché à extraire des mesures fiables de l'articulation labiale à l'aide d'un système de capture optique de mouvement, en l'occurrence le système Qualisys, complété par l'analyse des enregistrements vidéo de face et de profil acquis simultanément. Je détaille ici les questions métrologiques directement liées à ces mesures en complément des questionnements scientifiques quant à l'effet de la position prosodique et la variation entre locuteurs développées dans la première partie de ce document de synthèse.

Le système de capture de mouvements Qualisys permet de suivre au moyen d'un réseau de quatre émetteurs/récepteurs infrarouge disposés autour du locuteur les trajectoires d'un ensemble de marqueurs passifs réfléchissants. En captant le rayonnement infrarouge réfléchi par les marqueurs, la partie réceptrice des émetteurs/récepteurs permet de déterminer la position en trois dimensions de chacun des marqueurs avec une fréquence d'échantillonnage de 100 Hz. Des signaux brefs générés toutes les 10 ms (*triggers*) permettent d'assurer la synchronisation entre le suivi de la position des marqueurs et le flux audio acquis simultanément. Hormis une étape préalable de calibration à l'aide d'un instrument dédié au balayage du champ couvert par les émetteurs/récepteurs infrarouge tout en maintenant un écart fixe entre marqueurs, et une étape de posttraitement qui consiste en l'identification de chacun des marqueurs, l'utilisation de ce système de capture optique de mouvements permet un suivi fin (à la fois au niveau spatial et temporel) et entièrement automatisé de la position des marqueurs, tout en maintenant une importante liberté de mouvements des locuteurs pour lesquels les seules contraintes sont de rester dans le champ des émetteurs/récepteurs et de ne pas masquer les marqueurs avec par exemple des mouvements manuels.

Dans les travaux menés avec Laurianne Georgeton, ce système a été utilisé pour suivre un total de treize marqueurs disposés sur le visage des locuteurs au moyen d'une colle non-toxique, dont quatre autour du contour labial, comme illustré sur la Figure 65. Les quatre marqueurs positionnés autour du contour labial l'ont été sur les commissures des lèvres (droite et gauche), sur la lèvre supérieure au niveau de l'arc de cupidon sur le bord vermillon et à l'opposé sur la lèvre inférieure. Trois marqueurs ont été positionnés sur le menton pour permettre une approximation du mouvement mandibulaire, la position exacte pouvant être décalée par rapport à celle des marqueurs en raison du glissement de la peau pendant les

mouvements. Les autres marqueurs positionnés sur l'arête du nez et sur un casque fixé au niveau des tempes des locuteurs ont quant à eux été utilisés comme points de référence. La Figure 65 illustre également la détection par le système Qualisys de la position des treize marqueurs pour l'une des locutrices enregistrées. Sur le plan méthodologique, l'un des enjeux était de déterminer les mesures les plus adaptées pour caractériser l'articulation labiale à partir des positions des marqueurs positionnés sur le visage des locuteurs.

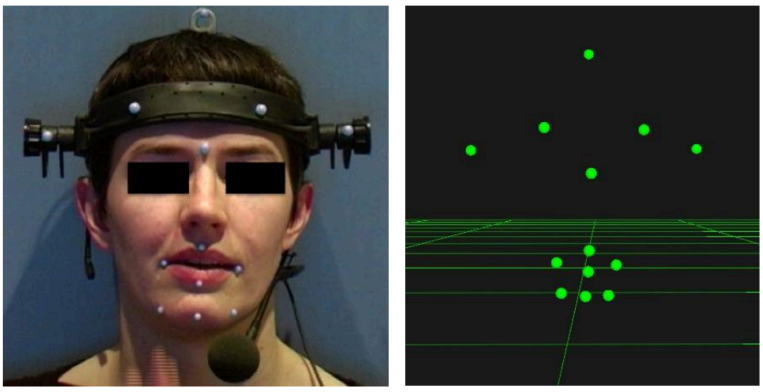


Figure 65 : Photographie de face du positionnement des treize marqueurs réfléchissants sur le visage d'une locutrice (gauche), et projection dans un plan en deux dimensions des positions en trois dimensions des marqueurs détectée par le système Qualisys (droite). D'après Georgeton (2014).

Les données analysées pour évaluer les mesures d'articulation labiale, issues des données recueillies par Laurianne Georgeton auprès de trois locutrices francophones dans le cadre de sa thèse (Georgeton, 2014), consistaient en des occurrences de voyelles en contexte $V_1C_1\#V_2C_2$ extraites de phrases construites pour placer la voyelle cible V_2 en position initiale de constituants prosodiques de différents niveaux.

7.1.4.2 Mesures de protrusion par capture optique de mouvements

Pour l'évaluation des mesures de protrusion que je commencerai par présenter ci-dessous, 491 occurrences au total des voyelles antérieures arrondies ou non /i-y/, /e-ø/ et /ɛ-œ/ ont été incluses, avec 21 à 32 occurrences par voyelle et par locutrice.

Trois mesures candidates de protrusion ont été retenues, afin de rendre compte des différentes mesures utilisées dans la littérature. En effet, si les auteurs s'accordent à considérer la protrusion comme une projection des lèvres vers l'avant, les approches retenues pour la caractériser et la quantifier divergent. Ainsi, Fromkin (1964) considère principalement l'avancement de la lèvre inférieure, tandis qu'un nombre plus important d'études se concentrent sur la lèvre supérieure (voir par exemple Tabain & Perrier (2005)). On peut également mentionner quelques études qui concluent à un poids plus important de l'avancement des commissures, soit considéré seul (Abry & Boë, 1980) soit combiné avec l'avancement des lèvres (Robert et al., 2007).

Les trois mesures de protrusion candidates ont été calculées comme l'avancement respectif de trois points par rapport à un plan perpendiculaire au plan sagittal et passant par l'arête du nez, défini à partir des six marqueurs fixes : les cinq marqueurs positionnés sur le casque et celui positionné sur l'arête du nez. Ces trois points, illustrés sur la Figure 66, sont la position du marqueur UL fixé sur la lèvre supérieure, la position du marqueur LL fixé sur la

lèvre inférieure, et le point virtuel C calculé comme la position moyenne entre les marqueurs fixés sur les commissures gauche et droite afin d'obtenir un point situé à équidistance de ces deux commissures.

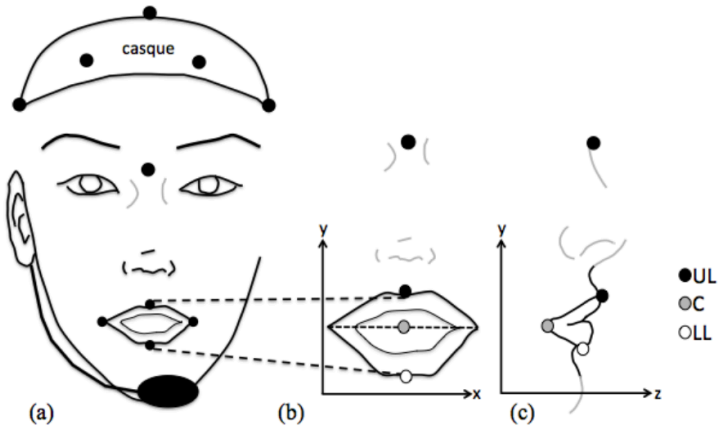


Figure 66 : Représentation schématique de la disposition des marqueurs réfléchissants sur le contour externe des lèvres et projection dans les deux plans utilisés pour l'extraction de mesures d'articulation labiale : (a) position des marqueurs, à l'exception des trois marqueurs positionnés sur le menton qui sont masqués par le microphone sur le schéma ; (b) et (c) vue de face et de profil position des marqueurs situés sur la lèvre supérieure (UL, en noir), sur la lèvre inférieure (LL, en blanc) et position équidistante entre les deux marqueurs placés sur les commissures (C, en gris) utilisée pour les mesures de protrusion. D'après Georgeton et Audibert (2014, [ACTI36]).

Afin d'évaluer la capacité de chacune de ces trois mesures de protrusion candidates de rendre compte de l'opposition d'arrondissement en français, chacune a fait l'objet d'une comparaison entre voyelle arrondie et non arrondie, séparément pour chaque paire de voyelles de même aperture et chacune des trois locutrices, une correction de type Bonferroni étant appliquée pour tenir compte des comparaisons multiples. Les résultats ont révélé que seule la mesure d'avancement du point C équidistant des deux commissures distinguait significativement la voyelle non-arrondie /i/ de l'arrondie /y/ pour l'ensemble des trois locutrices analysées. En revanche, si l'effet de l'arrondissement sur cette mesure est également significatif sur la paire mi-fermée /e-ø/ et la paire mi-ouverte /ɛ-œ/ pour l'une des trois locutrices, il ne l'est pas pour les deux autres locutrices. De plus, la mise en relation via des corrélations de chacune de ces trois mesures de protrusion avec les mesures formantiques (F2 et F3, complété par la distance F3-F2 pour /i/ et /y/ considérées comme focales) a confirmé que parmi les trois mesures de protrusion candidates, la mesure d'avancement des commissures est celle qui reflète le mieux les corrélats acoustiques de la protrusion.

7.1.4.3 Évaluation comparative : capture de mouvement vs. vidéo

Pour l'évaluation des mesures d'écartement labial et autres mesures obtenues à partir des positions des marqueurs dans le plan frontal, 306 occurrences au total des voyelles /a, i, y/ ont été analysées (104 /a/, 100 /i/ et 102 /y/). Dans la littérature, l'aire aux lèvres est considérée comme permettant une séparation parfaite entre voyelles arrondies et non-arrondies (Graillot et al., 1980). Si l'écartement horizontal est considéré également comme une mesure fiable pour distinguer voyelles arrondies et non-arrondies, le pouvoir discriminant de l'écartement vertical est variable entre locuteurs (Abry & Boë, 1980). Une autre mesure combinant écartement horizontal et vertical a également été proposée : le facteur K2 qui consiste en un rapport entre l'écartement horizontal et l'écartement vertical (Descout et al., 1980). Toutefois

à l'exception de rares études ayant considéré le contour externe des lèvres comme par exemple celle de Robert et al. (2007), la grande majorité des études articulatoires portant sur la labialité se sont concentrées sur le contour labial interne, et à notre connaissance aucune n'avait comparé directement les informations obtenues à partir du contour interne et celles obtenues à partir du contour externe. Afin d'évaluer dans quelle mesure les données obtenues à partir d'un système de capture optique de mouvements et donc relatives au contour externe du fait du positionnement des marqueurs peuvent être comparées à celles de la littérature obtenues à partir du contour interne, nous avons procédé à une comparaison directe entre les mesures issues du système de capture optique et celles issues de l'annotation des enregistrements vidéo dans le plan frontal réalisées sur les mêmes données.

Les données vidéo n'étant pas directement enregistrées par le système Qualisys, une étape de post-synchronisation à partir des enregistrements acoustiques effectués par le système Qualisys d'une part et par la caméra vidéo d'autre part a été nécessaire. Cette synchronisation a été effectuée à partir du claquement de mains réalisés par les locutrices au début de chaque enregistrement. A l'aide de scripts Matlab dédiés, la trame correspondant au milieu de chaque voyelle cible a fait l'objet d'une annotation manuelle des quatre points du contour labial interne correspondant aux commissures et aux parties internes de la lèvre inférieure et supérieure, en prenant comme référence pour ces deux derniers points l'axe reliant les marqueurs Qualisys UL et LL. A partir des quatre points ainsi délimités respectivement pour le contour externe (position des marqueurs Qualisys) et pour le contour interne (annotation des trames vidéo), les mesures de distance horizontale et verticale ont été extraites, ainsi que le facteur K2 et l'aire aux lèvres définie ici comme l'aire du polygone délimité par les quatre points.

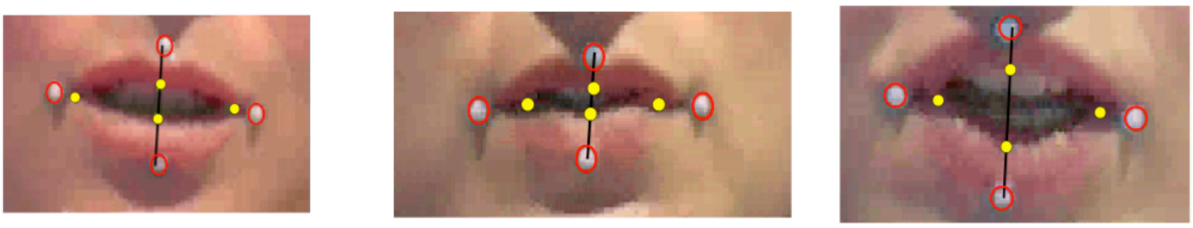


Figure 67 : Illustration de l'annotation manuelle sur les trames de l'enregistrement vidéo de face de la position des quatre points du contour labial interne (points jaunes) à partir de la position des marqueurs Qualisys (cercles rouges). Les trois images correspondent chacune à un extrait de l'une des locutrices. La position des points du contour interne utilisés pour calculer l'écartement vertical est définie à partir d'une ligne reliant les marqueurs Qualisys positionnées respectivement sous la lèvre inférieure et au-dessus de la lèvre supérieure. D'après Georgeton et Audibert (2012, [ACTI43]).

Comme l'illustre la Figure 68, la distribution des voyelles dans l'espace des distances horizontale et verticale obtenues à partir respectivement du contour interne et du contour externe est sensiblement différente entre les deux versions de ces mesures dérivées de la position des points sur le contour labial. La comparaison entre voyelles à partir de ces mesures pour les trois locutrices a de plus indiqué que si les distinctions entre les trois voyelles considérées ainsi que leur hiérarchie sont globalement préservées entre les deux versions, dans les mesures d'aire aux lèvres ou le rapport de distances, les mesures prises sur le contour externe rendent moins bien compte des différences plus subtiles entre locuteurs ou entre positions prosodiques que la référence que constitue le contour labial interne. Cette perte de

précision induite par l'utilisation du contour externe peut s'expliquer en partie par des différences individuelles anatomiques mais aussi de stratégie articulatoire, d'où un écart entre contour interne et externe qui varie non seulement entre locuteurs mais aussi entre productions d'un même locuteur. En conséquence et bien que l'utilisation de mesures obtenues à partir du contour externe reste envisageable pour d'autres études qui nécessiteraient une moindre finesse de mesure de l'ouverture labiale, le choix a été fait dans la suite de ces travaux de n'exploiter les mesures obtenues par capture optique de mouvements que pour l'estimation du degré de protrusion.

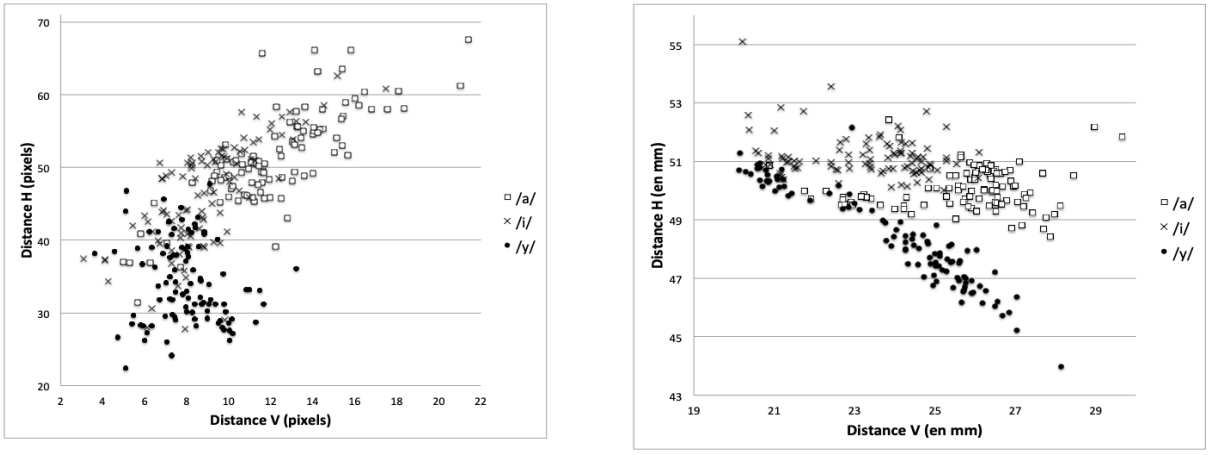


Figure 68 : Distribution dans l'espace des valeurs de distance verticale (abscisse) et de distance horizontale (ordonnée) des 306 occurrences des voyelles /a, i, y/ analysées sur les trois locutrices confondues, pour les mesures obtenues à partir du contour interne dans la partie gauche et du contour externe dans la partie droite. D'après Georgeton et Audibert (2012).

7.1.5 Utilisation simultanée de capteurs physiologiques multiples

Publications associées :

[ACL5] Crevier Buchman, L., Amelot, A., Al Kork, S. K., Adda-Decker, M., **Audibert, N.**, Chawah, P., Denby, B., Fux, T., Jaumard-Hakoun, A., Roussel, P., Stone, M., Vaissière, J., Xu, K., & Pillot-Loiseau, C. (2015). Acoustic Data Analysis from Multi-Sensor Capture in Rare Singing. *International Journal of Heritage in the Digital Era*, 4, pp. 121-132.

[ACTI37] Chawah, P., Fux, T., Adda-Decker, M., Amelot, A., **Audibert, N.**, Denby, B., Dreyfus, G., Jaumard-Hakoun, A., Pillot-Loiseau, C., Roussel, P., Stone, M., Xu, K., & Crevier Buchman, L. (2014). An educational platform to capture, visualize and analyze rare singing. *Proceedings of Interspeech 2014: Show & Tell Contribution*, Singapour, pp. 2128-2129.

Dans le cadre du projet européen iTreasures sur la préservation de l'héritage culturel intangible auquel j'ai contribué, le Laboratoire de Phonétique et Phonologie a été en charge de la collecte d'enregistrements de chants rares et de *Human Beatbox* via le recueil synchrone de données acoustiques et de mesures articulatoires multicapteurs, en collaboration principalement avec une équipe de recherche de l'université Pierre et Marie Curie. Dans ce projet qui a donné lieu au recrutement de post-doctorants spécialisés en électronique, mon rôle a principalement consisté en la participation aux choix méthodologiques pour la conception et l'évaluation des outils de recueil et d'analyse, notamment concernant la question cruciale de la synchronisation de signaux hétérogènes.

Afin de permettre un recueil contrôlé de données relatives à l'articulation labiale et linguale ainsi qu'à la nasalité et au contrôle laryngé synchronisées avec l'acoustique tout en laissant aux chanteurs une aussi grande liberté de mouvements que possible, le choix s'est porté sur une structure légère permettant de déplacer les capteurs avec les mouvements de tête du locuteur sans modification de distance susceptible d'impacter le recueil de données, sous la forme d'un casque baptisé *HyperHelmet*. La Figure 69 illustre ce dispositif qui combine microphone, sonde échographique et caméra vidéo, complétée par un électroglottographe et une paire d'accéléromètres piézoélectriques.

Outre le développement de l'outil de recueil proprement dit, des outils logiciels ont été développés grâce à l'expertise des post-doctorants Patrick Chawah et Thibaut Fux afin de piloter les enregistrements en disposant d'une visualisation simultanée des signaux issus des différents capteurs, permettant ainsi de contrôler en temps réel ces enregistrements. La Figure 70 illustre le suivi du recueil multicapteurs pendant une session d'enregistrement de chant corse (*Cantu in Paghella*) à l'aide du logiciel dédié développé à cet effet. En outre, des outils d'analyse ont également été développés afin de simplifier l'annotation des contours linguaux obtenus à l'aide de la sonde électroglottographique en l'automatisant partiellement, mais aussi de proposer un affichage simultané des signaux vidéo, de spectrogrammes et du tracé d'analyses acoustiques courantes (fréquence fondamentale, formants) et de mesures dérivées du signal électroglottographique pour permettre une interprétation plus directe des variations observées de la part des utilisateurs phonéticiens.

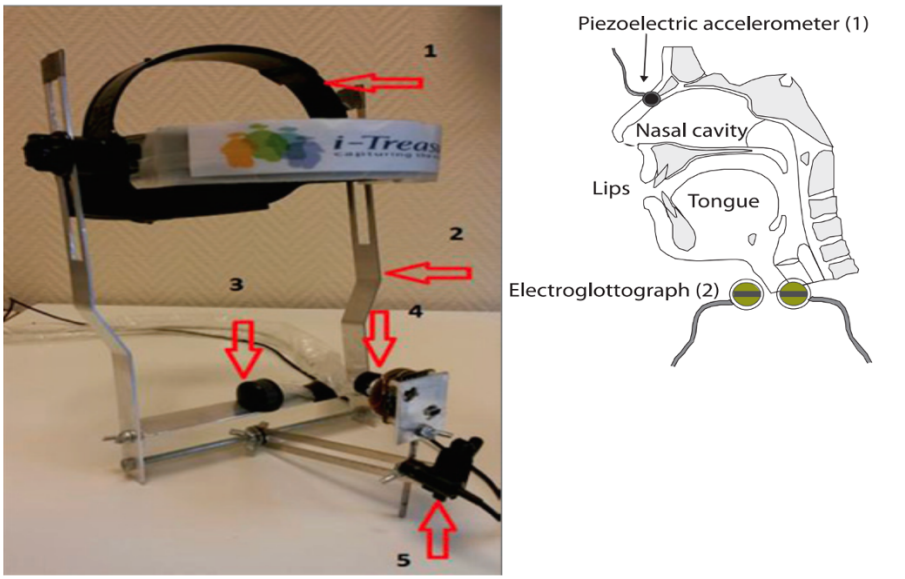


Figure 69 : Photographie du système multi-capteurs mobile réglable *HyperHelmet* (gauche), complété par une représentation schématique (droite) du positionnement des deux capteurs additionnels non directement intégrés au casque : accéléromètre piézoélectrique sur l'arête nasale du locuteur et électroglottographe autour du cou au niveau du larynx. Les flèches rouges numérotées sur la partie gauche de la figure correspondent respectivement à : (1) la bande ajustable permettant de fixer le casque autour du crâne du locuteur ; (2) la barre verticale de réglage de la hauteur ; (3) le support ajustable de fixation de la sonde échographique ; (4) la caméra labiale avec réglage du focus et de l'orientation ; (5) le microphone.

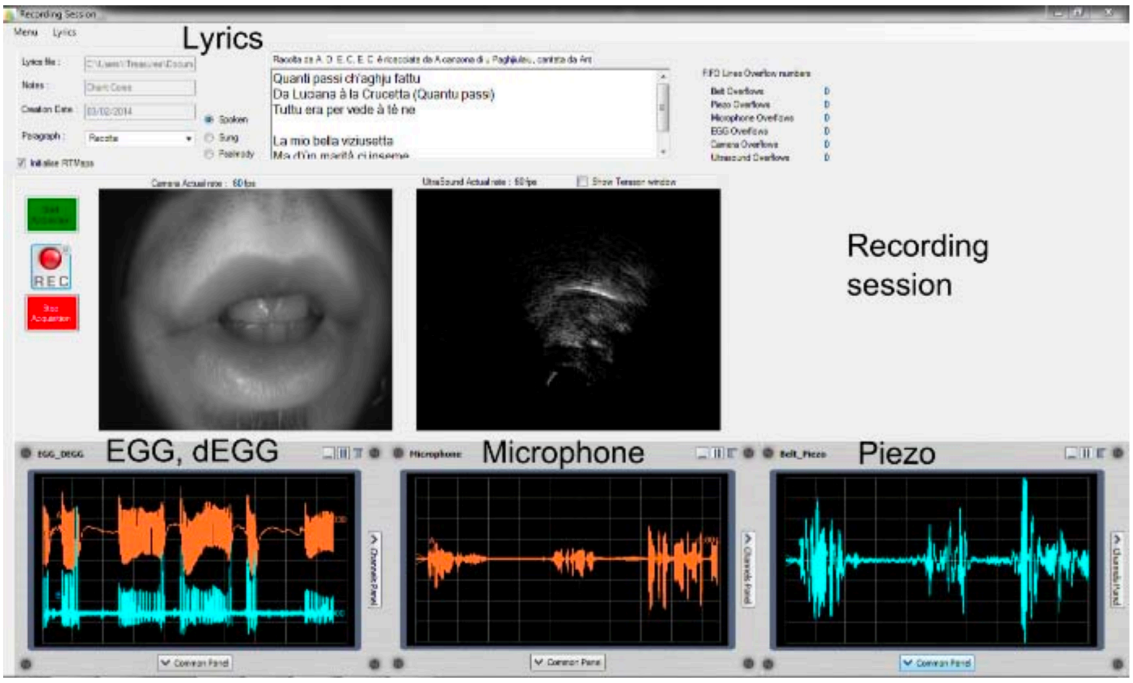


Figure 70 : Illustration du monitoring simultané des différents canaux d'enregistrement et de la transcription du texte lors d'une session d'enregistrement de chant corse « Cantu in Paghjella » à l'aide du système multi-capteurs *HyperHelmet*

7.2 Mesures et méthodes acoustiques pour la recherche en phonétique

7.2.1 Alignement forcé appliqué à la parole dysarthrique

Publications associées :

[ACTI54] **Audibert, N.**, Fougeron, C., Fredouille, C., Meunier, C., & Panseri, O. (2010). Évaluation d'un alignement automatique sur la parole dysarthrique. *Actes des 18èmes Journées d'Études sur la Parole (JEP 2010)*, Mons, Belgique, pp. 353-356.

[ACTI56] Fougeron, C., **Audibert, N.**, Fredouille, C., Meunier, C., Gendrot, C., & Panseri, O. (2010). Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique. *Actes des 18èmes Journées d'Études sur la Parole (JEP 2010)*, Mons, Belgique, pp. 365-368.

7.2.1.1 Principe général de l'alignement et application à la dysarthrie

Si les progrès techniques des dernières décennies ont permis le développement d'analyses à grande échelle en phonétique de corpus (M. Y. Liberman, 2019), que j'aborde dans la première partie de ce document de synthèse pour introduire mes propres contributions en la matière, de telles études nécessitent dans la grande majorité des cas le recours à des outils d'alignement en phones pour segmenter le signal acoustique et identifier la position des segments qui font l'objet d'analyses comparatives. En pratique, afin que ce processus de segmentation puisse être considéré comme suffisamment fiable pour exploiter et interpréter les mesures résultantes, la méthode utilisée est généralement celle de l'alignement forcé. A partir de la spécification d'une transcription supposée correspondre au signal, le plus souvent fournie sous forme orthographique, cette méthode permet de contraindre les possibilités

d'appariement entre signal acoustique et séquence de phones et ainsi de limiter les erreurs de segmentation.

Bien que les outils d'alignement forcé soient principalement développés par les unités de recherche en traitement automatique de la parole qui utilisent à ces fins les modèles acoustiques entraînés pour les systèmes de reconnaissance automatique de la parole, un certain nombre d'outils performants et accessibles aux phonéticiens ont été développés et diffusés. Parmi ceux qui intègrent des modèles pour le français et sont largement utilisés, on peut ainsi citer WebMAUS (Kisler et al., 2017), SPPASS (Bigi, 2015) ou encore Montreal Forced Aligner (McAuliffe et al., 2017). L'usage de tels outils s'est très largement démocratisé dans la communauté de recherche en phonétique, et ils sont désormais couramment employés également sur des données issues de corpus de parole de laboratoire afin d'obtenir une première version automatique de la segmentation en phones (et éventuellement en mots) qui est ensuite contrôlée et corrigée manuellement.

Lorsque les conditions favorables au fonctionnement de ce type d'outil sont réunies, ils permettent d'obtenir aisément une segmentation fiable, et ainsi de traiter des volumes de données trop importants pour pouvoir faire l'objet d'une segmentation manuelle. Des conditions idéales impliquent que la qualité du signal soit suffisante et exempte de bruits de fond ou de chevauchements de tours de parole, que la transcription corresponde fidèlement aux productions des locuteurs et que la nature des données de parole à aligner ne soit pas trop éloignée de celles utilisées pour entraîner les modèles acoustiques. Les modèles sont généralement entraînés sur les variantes de la langue cible les mieux fournies en ressources numériques, typiquement le français standard produit par de jeunes adultes dans le cas du français. Ce dernier point peut s'avérer plus problématique pour l'application à des données atypiques, comme par exemple des enregistrements de parole non-native ou à plus forte raison des enregistrements de parole pathologique, notamment dans le cas d'une pathologie telle que la dysarthrie qui dans le cas d'atteintes sévères peut induire des distorsions très importantes de l'articulation des segments de parole.

Dans le cadre du projet ANR DesPho-APaDy qui visait à documenter la parole dysarthrique via la collecte d'enregistrements d'un nombre conséquent de patients et à en fournir une description phonétique, la question de l'applicabilité de méthodes de traitement automatique et notamment de l'alignement forcé s'est avérée cruciale. J'ai donc été amené à mettre en œuvre une évaluation de la fiabilité de l'alignement forcé obtenu à l'aide du système développé au Laboratoire d'Informatique d'Avignon et appliqué à des enregistrements de parole dysarthrique. Le système d'alignement utilisé repose sur le principe des modèles de Markov cachés (HMM) complété par le décodage par un algorithme de type Viterbi, classique en matière d'alignement forcé (Brugnara et al., 1993). Les modèles HMM utilisés ont été entraînés sur le corpus de parole radiophonique ESTER (Galliano et al., 2006). Afin de limiter les sources potentielles d'erreur, pour l'application à la parole dysarthrique le dictionnaire phonétisé a été adapté pour le recentrer sur les formes présentes dans les enregistrements à aligner, parfois éloignées des formes canoniques déjà intégrées dans le dictionnaire phonétisé utilisé par défaut par le système. Je reviens en section 7.3.1 sur les conventions de transcriptions adoptées et leur incidence sur l'alignement forcé.

7.2.1.2 Evaluation de l'alignement forcé

Le corpus utilisé pour cette évaluation consistait en la lecture du texte « Tic-tac » (parfois désigné sous le nom « Le petit cordonnier »), issu de la base de test développée par Claude

Chevrie-Muller et sélectionné pour sa densité en séquences articulatoirement complexes susceptibles de mettre en lumière les difficultés articulatoires des patients dysarthriques. Le processus de segmentation phonémique fine étant particulièrement chronophage, cette évaluation s'est concentrée sur un sous-ensemble restreint. Quatre patients locuteurs du français standard atteints de dysarthrie modérée (degré 1 dans la Batterie d'Evaluation Clinique de la Dysarthrie proposée par Auzou & Rolland-Monnoury (2006)) ou sévère (degré 2) enregistrés en milieu hospitalier dans un environnement calme ont été inclus, en équilibrant entre hommes et femmes et entre degrés de sévérité. En complément, un enregistrement dans les mêmes conditions de la lecture de ce texte par un homme et une femme non-dysarthriques (degré 0) d'âge proche de celui des patients a été ajouté en tant que condition de contrôle. Les productions de ces six locuteurs ont fait l'objet d'une transcription orthographique selon la convention définie dans le cadre du projet DesPho-APaDy afin de tenir compte des disfluences mais aussi d'adaptations spécifiques aux locuteurs dysarthriques observées dans les productions.

Pour les besoins de cette évaluation, deux experts phonéticiens ont corrigé indépendamment l'un de l'autre les alignements automatiques générés par le système d'alignement automatique. Ces corrections consistaient non seulement en une modification des frontières temporelles des phones, mais aussi dans certains cas en des substitutions, ajouts ou délétions d'unités appariées à tort avec le signal acoustique par le système d'alignement. Bien que les deux experts sollicités aient été formés à la même école et se soient appuyé sur les mêmes critères acoustiques pour déterminer les frontières segmentales, par exemple l'apparition du second formant pour segmenter le début et la fin des voyelles, les choix de segmentation peuvent toutefois varier entre annotateurs comme documenté par exemple par Pitt et al. (2005) sur la parole conversationnelle en anglais américain. Cette variabilité inter-annotateurs est particulièrement susceptible d'être observée sur des séquences entre voyelles et consonnes présentant une structure formantique (notamment dans le cas des approximantes précédées ou suivies de voyelles, pour lesquelles le choix de placement de la frontière est en partie arbitraire). Dans le cas de la parole dysarthrique qui peut dans les cas sévères présenter des continuums de vocoïdes difficiles à identifier précisément, une telle fluctuation dans les choix de segmentation est d'autant plus probable. Bien que les cas les plus extrêmes de production de tels continuums jugés inintelligibles et impossibles à segmenter aient été exclus de l'analyse, de nombreux cas intermédiaires subsistent. Afin d'évaluer l'impact de la variabilité inter-annotateurs sur les choix de segmentation, les productions de deux des patients ont ainsi été corrigées par les deux annotateurs.

Hormis le cas des continuums jugés insegmentables, la grande majorité des divergences observées entre la segmentation issue du système d'alignement automatique et les corrections des annotateurs portait sur le décalage des frontières temporelles de début et de fin de segment. L'analyse de la fiabilité du système s'est donc focalisée sur ces décalages temporels, après élimination des cas d'insertion ou de délétion entre versions de l'alignement comparées et des segments voisins. Les comparaisons entre alignement automatique et correction manuelle ont porté respectivement sur 1319 (77% du total) et 1917 (86% du total) segments pour chacun des deux annotateurs, en analysant non seulement le décalage des frontières initiales et finales, mais aussi celui du point central des segments, fréquemment retenu comme point de mesure dans les analyses phonétiques acoustiques, notamment pour l'analyse acoustique des voyelles. Les décalages moyens mesurés entre alignement

automatique et correction manuelle pour chacun des locuteurs et chaque annotateur sont illustrés par la Figure 71.

En complément de l'évaluation de la significativité statistique des comparaisons, peu informative en raison du nombre de segments comparés, nous avons également analysé la proportion de segments pour lesquels le décalage du point central était supérieur à 20 ms. Bien que nous ne l'ayons pas détaillé dans l'étude publiée pour des raisons de format, ce seuil de 20 ms était inspiré des critères retenus précédemment par (Nefti, 2004) pour l'évaluation de l'alignement automatique de la parole. Il était également motivé par les observations de Pitt et al. (2005) sur l'anglais américain, qui ont relevé une variabilité inter-annotateur moyenne de 16 ms, arrondie dans notre cas à 20 ms pour tenir compte de la résolution temporelle de 10 ms du système utilisé. Une autre motivation qui m'a conduit à faire ce choix d'analyse des décalages est la présence, typique du fonctionnement des systèmes d'alignement forcé dans les cas d'échec de l'appariement entre modèles acoustiques et trames du signal à aligner, de segments subissant un décalage très important (jusqu'à 300 ms dans ces données) et donc susceptibles de biaiser les mesures de décalage moyen.

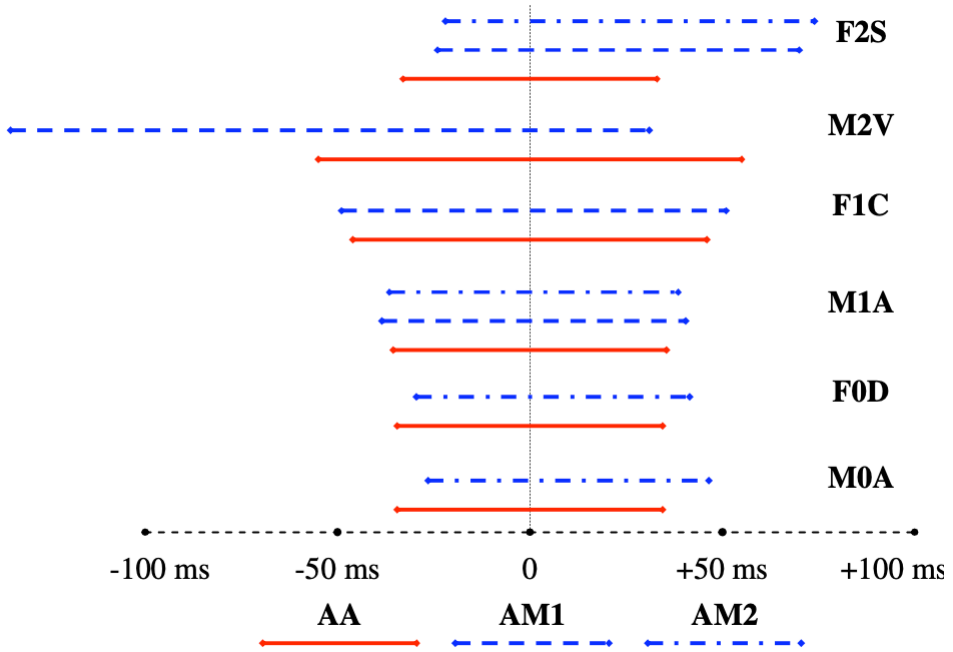


Figure 71 : Durées segmentales et décalages moyens relevés entre l'alignement automatique (AA, centré sur la valeur 0) et les segmentations manuelles de références corrigées par les deux experts phonéticiens, AM1 et AM2, pour chacun des six patients dysarthriques évalués. Les codes en F correspondent aux femmes, ceux en M aux hommes, le chiffre indiquant le degré de sévérité de la dysarthrie (0 = non-dysarthrique ; 1 = dysarthrie légère ; 2 = dysarthrie sévère). Les alignements des patients F2S et M1A ont été corrigés par les deux experts pour évaluer la consistance inter-juges. D'après Audibert et al. (2010), [ACTI54].

Conformément à ce que suggèrent les décalages moyens représentés sur la Figure 71, les résultats ont indiqué un effet de la sévérité de la dysarthrie sur le décalage entre alignement automatique et référence manuelle, avec respectivement 53% et 66% de segments dont le point central était décalé de plus de 20 ms pour les deux patients atteints de dysarthrie sévère, tandis que cette proportion était comprise entre 14% et 26% pour les locuteurs contrôles et les patients atteints de dysarthrie modérée. Par ailleurs la comparaison entre annotateurs sur

les données des deux patients ayant fait l'objet d'une double correction indiquait une proportion de segments dont le point central est décalé de plus de 20 ms de 3% pour le patient atteint de dysarthrie modérée, et de 7% pour la patiente atteinte de dysarthrie sévère. L'analyse par classe phonémique de ces décalages a de plus révélé un décalage particulièrement important de la barre d'explosion dans le cas des occlusives sourdes, segmentées en tenue et relâchement dans nos données, en dépit d'un décalage modéré des frontières de début et de fin des occlusives considérées dans leur ensemble, qui exclut la possibilité d'extraire des mesures telles que le VOT à partir de la segmentation automatique. En revanche les durées des consonnes liquides /l/ et /ʁ/ ont été mieux préservées par l'alignement automatique.

Par la suite, l'utilisation combinée de deux ensembles de modèles acoustique pour affiner les performances du système d'alignement forcé a permis d'améliorer les performances en faisant baisser la proportion de segments avec le point central décalé de plus de 20 ms à 15% pour les locuteurs contrôles non-dysarthriques, 23% pour les locuteurs modérément dysarthriques et 44% pour les locuteurs sévèrement dysarthriques. En dépit de ces améliorations notables qui ont permis d'envisager l'exploitation de l'alignement forcé sur les productions de locuteurs dysarthriques, les décalages observées pour les patients sévèrement dysarthrique restaient conséquents. Ces résultats nous ont permis de cadrer l'utilisation possible de l'alignement forcé appliqué à la parole dysarthrique, en limitant la généralisation de son usage aux cas de dysarthries modérées. En effet, une application sans vérification et correction manuelle à des productions de patients sévèrement dysarthriques nécessiterait une adaptation des modèles acoustiques à ces productions très différentes de celles sur lesquelles les modèles sont entraînés, pour laquelle une quantité conséquente de parole fortement dysarthrique segmentée manuellement serait nécessaire, ce qui semble peu réaliste au vu de la difficulté de la tâche sur des productions aussi dégradées. Par ailleurs et bien que les performances obtenues sur ces données puissent être encore affinées, le taux élevé de décalage observé peut amener à s'interroger sur l'intérêt dans le cas des dysarthries sévères d'une première passe obtenue à partir d'un alignement automatique avant correction manuelle, en comparaison d'une approche entièrement manuelle sur ces données.

A la lumière des résultats issus de cette évaluation, nous avons également pu formuler des recommandations quant à l'utilisation de segmentations obtenues automatiquement sur la parole dysarthrique, et plus généralement sur la parole atypique, la principale étant de filtrer l'alignement obtenu en tant compte des scores de confiance attribués par le système lors du processus d'appariement des modèles acoustiques et du signal à aligner.

7.2.1.3 Impact des erreurs d'alignement sur les mesures phonétiques

En complément de l'évaluation des décalages temporels entre l'alignement forcé automatique et la correction manuelle de référence, nous avons également évalué sur les données des quatre locuteurs dysarthriques l'incidence de ces décalages sur les mesures phonétiques envisagées pour caractériser la parole dysarthrique. Outre les mesures temporelles de durées segmentales ainsi que de durée et de distribution des pauses qui découlent directement de la segmentation issue de l'alignement forcé ou de sa correction manuelle, un ensemble de mesures acoustiques ont été extraites et comparées. Ces mesures acoustiques se sont concentrées sur la fréquence du centre de gravité spectral (CoG) mesuré dans le bruit des fricatives ainsi que sur les positions des deux premiers formants des voyelles, toute ces mesures étant prises au point central défini par chacun des deux alignements.

La comparaison entre alignement automatique et correction manuelle a montré que si les pauses annotées par les correcteurs étaient bien détectées comme telles par l'alignement forcé, l'alignement automatique insère également une proportion conséquente de pauses inexistantes en raison du codage dans le lexique utilisé par l'aligneur de la possibilité d'insérer une pause après chaque mot. Ces insertions de pauses, qui par ailleurs sont à l'origine d'une partie des décalages mesurés entre alignement automatique et alignement manuel, peuvent donc introduire un biais important si elles ne sont pas corrigées.

Afin de disposer d'un nombre d'exemplaires suffisant dans chaque catégorie, les mesures de centre de gravité spectral ont été comparées entre catégories larges pour chaque locuteur. Les fricatives alvéolaires /s/ et /z/ ont été considérées séparément comme dentales, tandis que les labiodentales /f/ et /v/ et les post-alvéolaires /ʃ/ et /ʒ/ ont été regroupées en tant que non-dentales, afin d'évaluer si la différence de CoG entre dentales et non-dentales était bien préservée par l'alignement automatique. Cette comparaison est illustrée par la Figure 72 : bien que des fluctuations des valeurs moyennes entre les deux versions de l'alignement aient été observées pour les deux locuteurs sévèrement dysarthriques, dans l'ensemble des cas le CoG des fricatives dentales restait supérieur à celui des non-dentales quelle que soit la version de l'alignement utilisée.

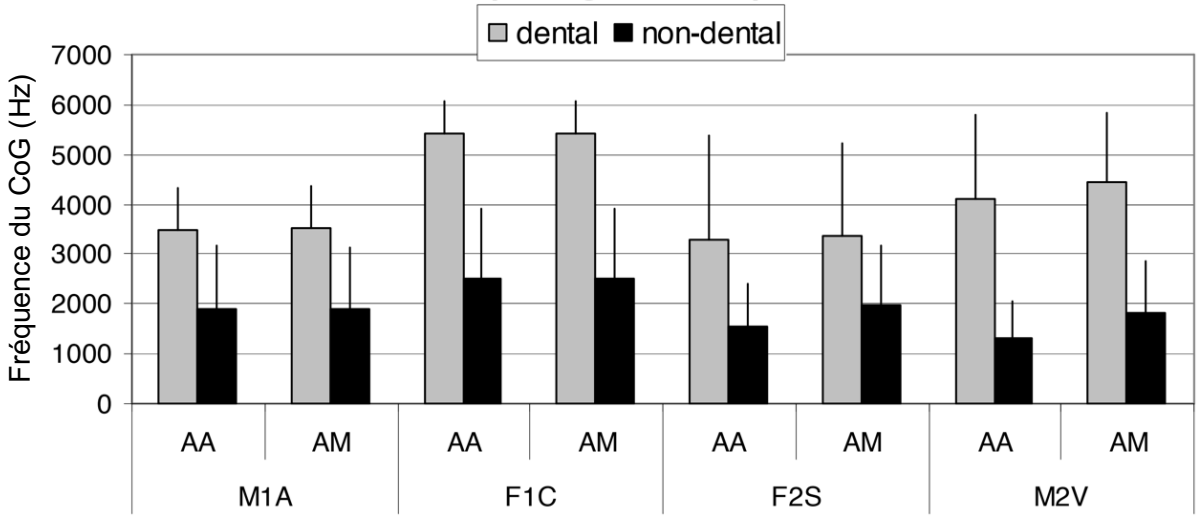


Figure 72 : Mesures de centre de gravité spectral (CoG) moyen du bruit des fricatives comparées entre alignement automatique (AA) et alignement manuel (AM) pour les quatre locuteurs dysarthriques analysés, comparées entre lieux d'articulations dentaux et non-dentaux. Les barres d'erreur représentent l'écart-type. Les codes en F correspondent aux femmes, ceux en M aux hommes, le chiffre indiquant le degré de sévérité de la dysarthrie (0 = non-dysarthrique ; 1 = dysarthrie légère ; 2 = dysarthrie sévère). D'après Fougeron et al. (2010, [ACTI56]).

Enfin, quatre catégories de voyelles orales suffisamment représentées dans le texte lu par les locuteurs ont été prises en compte pour la comparaison des valeurs des deux premiers formants, les voyelles moyennes /e, ɛ/ étant fusionnées en une même catégorie de même que /o, ɔ/ : /a/, /e, ɛ/, /i/ et /o, ɔ/. Les mesures formantiques ont été évaluées comme globalement robustes aux erreurs d'alignement, le changement d'alignement n'ayant pas d'effet sur les valeurs de F1 ni de F2. De plus, la corrélation entre les valeurs formantiques mesurées à partir de l'alignement automatique et celles mesurées à partir de l'alignement manuelle était élevée (r = .9 à la fois pour les mesures de F1 et les mesures de F2).

Si l'importance de certains décalages relevés entre l'alignement automatique et celui de référence est incompatible avec une analyse fine et que l'insertion de pauses inexistantes peut s'avérer problématique, la comparaison de ces mesures acoustiques a confirmé que l'utilisation d'un alignement forcé automatique sur la parole dysarthrique était envisageable à condition de s'accompagner de précautions.

7.2.2 Affinage de la détection de la fréquence fondamentale

Publications associées :

[ACTI27] **Audibert, N.**, & Falk, S. (2018). Vowel space and f0 characteristics of infant-directed singing and speech. *Proceedings of the 9th International Conference on Speech Prosody*, Poznan, Pologne, pp. 153-157.

[ACTI2] **Audibert, N.** (2024). iHist et iScatter, outils en ligne d'exploration interactive de données : application aux valeurs aberrantes de f0 et de formants. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 598-607.

Les outils logiciels dédiés (exclusivement ou en complément d'autres méthodes) à l'extraction de la fréquence fondamentale sont nombreux. Au cours des dix dernières années, plusieurs études ont proposé une évaluation comparative des performances de différents algorithmes de détection. Ainsi Jouvet & Laprie (2017) ont confronté 17 algorithmes aux valeurs de référence obtenues à partir de signaux électroglottographiques et ont évalué l'influence de l'ajout de bruit. Dans leur évaluation des outils gratuits applicables à l'analyse de productions de patients parkinsoniens à partir de signaux bruités, Illner et al. (2020) en ont retenu dix après avoir éliminé certaines méthodes candidates jugés inadaptées à la tâche évaluée ou trop proches d'autres méthodes incluses par ailleurs, en s'appuyant sur des corrections manuelles pour obtenir des valeurs de référence. Enfin, Vaysse et al. (2022) ont évalué les performances de douze algorithmes sur des productions de patients parkinsoniens ou atteints de cancer, également à partir de valeurs de référence corrigées manuellement.

En combinant les choix effectués par ces trois études, un total de 25 algorithmes ont été évalués, pour des résultats divergents d'une évaluation à l'autre mais globalement en faveur d'outils en ligne de commande tels que SWIPE (Camacho & Harris, 2008) ou FCN-f0 (Ardaillon & Roebel, 2019), la plupart suggérant de combiner les mesures issues de différentes méthodes pour obtenir une détection plus fiable. Ces résultats s'expliquent en partie par le choix d'opter pour un paramétrage par défaut, tandis que les algorithmes de détection de la fréquence fondamentale intégrés dans Praat (Boersma, 2001) et notamment l'algorithme fondé sur l'autocorrélation du signal acoustique utilisé dans la plupart des cas sur des signaux de parole (Boersma, 1993) requièrent un affinage des paramètres pour obtenir des résultats optimaux. Cette nécessité d'optimisation des paramètres concerne notamment les seuils haut et bas de valeurs de f0 considérées comme acceptables pour un locuteur dans une condition donnée plutôt que l'utilisation d'une plage large de valeurs possibles telle que celle proposée par défaut, qui a tendance à déboucher sur des sauts d'octave et autres erreurs grossières de détection dès lors que les signaux analysés sont bruités ou peu périodiques (dévoisement, voix craquée, etc.).

Ce constat a été à l'origine de la méthode proposée par De Looze (2010) dans le cadre de sa thèse, à travers une détection en deux étapes visant à automatiser ce processus de détection via des critères statistiques en l'absence de connaissances explicite des valeurs

minimale et maximale effectivement présentes dans les données analysées. Le principe de cette méthode est de réaliser une première passe de détection avec une plage large de valeurs possibles (de 60 Hz à 600 Hz) spécifiées à l'aide de l'algorithme de détection de Praat (Boersma, 1993), puis de s'appuyer sur la distribution des valeurs ainsi obtenues pour déterminer une plage de valeurs plus réaliste utilisée pour contraindre la détection lors d'une seconde passe. A partir d'une annotation manuelle des valeurs minimales et maximales de fréquence fondamentale dans un corpus de données de parole en anglais et en français, De Looze (2010) a proposé de fixer la valeur minimale $0.83*Q_{15\%}$, et la valeur maximale à $1.92*Q_{65\%}$, $Q_{15\%}$ et $Q_{65\%}$ désignant ici respectivement le quantile à 15% et le quantile à 65% issus des valeurs exprimées en Hertz obtenues à l'issue de la première passe de détection. Ces seuils empiriques, qui permettent une amélioration notable de la précision des estimations comparativement aux mesures issues du paramétrage par défaut, ont par la suite été repris dans de nombreuses études pour lesquelles la fréquence fondamentale a été détectée à l'aide de Praat. Toutefois l'efficacité de l'utilisation des mêmes seuils sur d'autres jeux de données repose sur le postulat implicite que la forme générale de la distribution des valeurs obtenues suite à la première passe serait transposable à d'autres locuteurs et condition, et que les proportions respectives de valeurs détectées comme anormalement basses et anormalement hautes suite à cette première passe le seraient également.

Dans le cadre de ma collaboration avec Simone Falk sur l'analyse de données de parole et de chant dirigés vers l'enfant produites par des locutrices allemandes que je présente plus en détail dans la section 4.3, j'ai pu constater que sur des conditions de productions aussi éloignées de celles représentées dans les données utilisées par De Looze (2010) pour déterminer les valeurs limites proposées dans sa méthode, le décalage pouvait dans certains cas être conséquent. En effet, la parole et le chant dirigé vers l'enfant sont produits avec des valeurs de fréquence fondamentale à la fois plus élevées que celles habituellement relevées dans la parole, mais aussi plus variables. La méthode employée n'étant décrite que brièvement dans les publications issues de ces travaux, je reviens ici sur les choix effectués et leur implication.

Suite à ces premières observations, j'ai opté pour une méthode de détection itérative inspirée de celle de De Looze (2010), mais dans laquelle les valeurs minimales et maximales de fréquence fondamentale utilisées pour la seconde passe de détection étaient déterminées visuellement à partir de l'inspection de la distribution des valeurs issues de la première passe de détection pour chaque locutrice et chaque style de parole ou de chant, complétée par la vérification des signaux correspondant à la plage de valeur proche des seuils déterminés par cette méthode. Dans ce cas particulier, la valeur maximale de 600 Hz pour la première passe a été conservée, mais la valeur minimale a été relevée à 110 Hz pour tenir compte du registre des locutrices. Notons que ce choix aurait pu impacter la valeur des seuils inférieurs et supérieurs calculés avec la méthode de De Looze (2010), mais son incidence est ici négligeable. La Figure 73 illustre la distribution des valeurs de fréquence fondamentale obtenues lors de la première passe de détection pour l'une de ces locutrices germanophones en condition de chant et de parole, ainsi que les valeurs limites de fréquence fondamentale pour la seconde passe de détection avec chacune des deux méthodes.

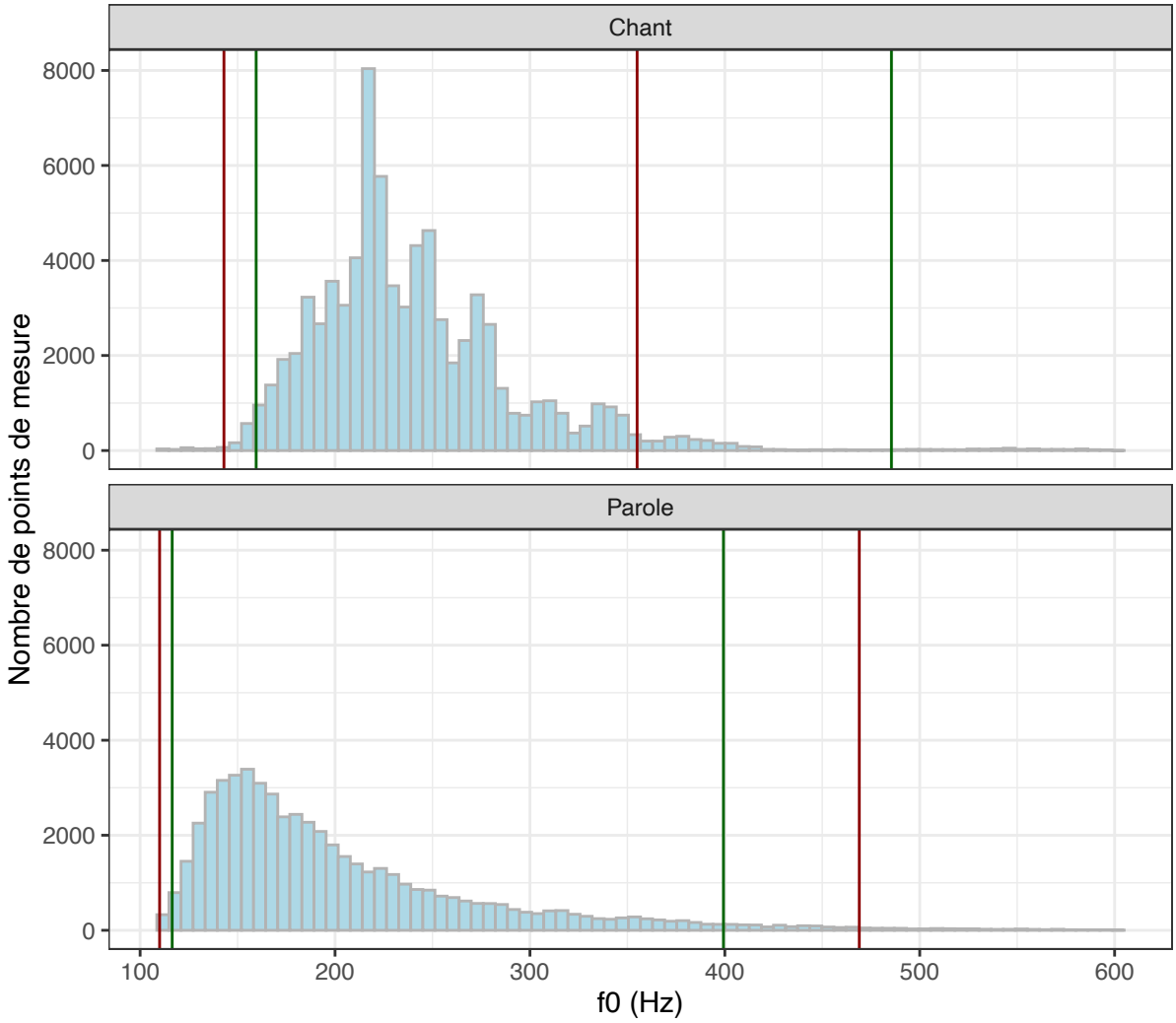


Figure 73 : Distribution des valeurs de fréquence fondamentale en Hertz extraites par Praat lors de la première passe de détection avec une valeur minimale de 110 Hz et une valeur maximale de 600 Hz, pour les productions chantées et parlées de l'une des locutrices des études de Audibert & Falk (2018, [ACTI27]) et de Falk & Audibert (2021, [ACL3]). Les lignes verticales vertes représentent les seuils bas et haut de fréquence fondamentale définis par la méthode de De Looze (2010), les lignes verticales rouges représentent ceux définis à partir de l'inspection des distributions complétée par la vérification des signaux correspondant aux plages de valeurs proches de ces limites.

Cette méthode fondée sur l'inspection visuelle des distributions, qui est à la base de l'exemple que je présente pour illustrer un usage possible de l'application iHist dans mon article sur lequel je reviens dans la partie suivante de ce document de synthèse, est en partie à rapprocher de celle proposée par Frid & Ambrazaitis (2010). Ces derniers ont proposé une méthode entièrement automatique fondée sur l'appariement entre la forme de la distribution des valeurs transformées en logarithme obtenues suite à la première passe de détection et une distribution-type et dont les premiers résultats ont suggéré un avantage pour cette méthode sur des données de parole spontanée comparativement à l'utilisation directe des seuils proposés par De Looze (2010). Si cette méthode n'a été appliquée jusqu'alors qu'à la détection de fréquence fondamentale effectuée avec l'algorithme de Praat (Boersma, 1993), elle pourrait également s'avérer utile pour affiner la détection réalisée avec d'autres outils

précédemment évalués comme plus performants pour la détection automatique de f0 et qui permettent également de spécifier une plage de valeurs pour contraindre la détection, comme c'est le cas des cinq algorithmes de détection intégrés dans l'outil en ligne de commande SPTK (Yoshimura et al., 2023). Dans le cas d'une méthode fondée sur l'utilisation d'un réseau de neurones convolutif comme FNC-f0 (Ardaillon & Roebel, 2019), évalué comme plus précis que les autres méthodes testées sur la parole pathologique par (Vaysse et al., 2022) mais qui ne permet pas un tel paramétrage, l'inspection de la distribution des valeurs obtenues pourrait être utilisée pour filtrer les valeurs, le cas échéant en combinant cette inspection avec celle des valeurs obtenues par le biais d'autres méthodes.

7.2.3 Mesures de l'espace vocalique

Publications associées :

[ACTI34] **Audibert, N.**, Fougeron, C., Gendrot, C., & Adda-Decker, M. (2015). Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS'15)*, Glasgow, Royaume-Uni, paper 0753 (actes en ligne).

[ACTI42] **Audibert, N.**, & Fougeron, C. (2012). Distorsions de l'espace vocalique : quelles mesures ? Application à la dysarthrie. *Actes des 19èmes Journées d'Études sur la Parole (JEP 2012)*, Grenoble, France, pp. 217-224.

[ACTI46] Fougeron, C., & **Audibert, N.** (2011). Testing various metrics for the description of vowel distortion in dysarthria. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong-Kong, Chine, pp. 687-690.

[ACTI10] Hermes, A., **Audibert, N.**, & Bourbon, A. (2023). Age-related vowel variation in French. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic. pp. 2045-2049.

Comme évoqué dans la première partie de ce document de synthèse, par souci de consistance thématique je fais le choix de présenter dans cette partie les aspects qui relèvent de la métrologie dans les travaux que j'ai mené avec Cécile Fougeron depuis 2011 sur la caractérisation de l'espace vocalique et son application à différents facteurs de variation, en complément de la présentation en section 4 des questions scientifiques et des moyens mis en œuvre pour y répondre. Par ailleurs, le développement d'un outil en ligne destiné à rendre aisément accessible le calcul de certaines de ces métriques est présenté en section 7.4.3.

Les métriques proposées dans la littérature et celles que nous avons introduit visent à capturer les variations entre locuteurs ou entre conditions de production de la structuration du système vocalique à partir de mesures formantiques, le plus souvent limitées aux deux premiers formants. Ces variations peuvent se subdiviser en trois principales dimensions. La principale de ces dimensions, qui est aussi la plus largement ciblée par les métriques existantes consiste en une mesure du degré d'expansion ou au contraire de centralisation du système vocalique considéré dans son ensemble. Ce degré de dispersion ou de centralisation peut être estimé sur l'ensemble des formants mesurés, ou spécifiquement sur l'une des dimensions acoustiques dans le but d'estimer les variations sur une dimension articulatoire particulière. En complément, il est possible d'estimer la dispersion des réalisations vocaliques au sein de

chaque catégorie de voyelle, et enfin le degré de chevauchement entre les réalisations de voyelles appartenant à des catégories différentes.

Bien que la centralisation de l'espace vocalique s'accompagne généralement d'un degré de recouvrement plus important entre catégories, qui en est la conséquence évidente à degré constant de dispersion égal au sein de chaque catégorie, la covariation n'est pas systématique entre ces trois dimensions qui ne sont que partiellement interdépendantes l'une de l'autre

7.2.3.1 Projection vers une valeur unique pour chaque dimension considérée

Dans nos premiers travaux communs (Fougeron & Audibert, 2011, [ACTI46] ; Audibert & Fougeron, 2012, [ACTI42]), nous nous sommes concentrés sur des métriques utilisées dans la littérature, en les adaptant si nécessaire au système vocalique du français, et en accordant une attention toutes particulière à celles dédiées à la caractérisation des distorsions vocaliques liées à la dysarthrie. Le point commun entre ces métriques est de proposer d'estimer par une unique valeur une dimension supposée représentative de l'organisation du système vocalique d'un locuteur, éventuellement recentrée sur une condition de production particulière.

La mesure la plus classiquement utilisée a été l'aire de l'espace vocalique, proposée initialement par (Fant, 1973), qui désigne l'aire de la région délimitée par les lignes reliant les coordonnées des voyelles dans l'espace bidimensionnel défini par les fréquences du premier et du second formant. Cette mesure est souvent désignée dans la littérature par l'acronyme VSA (de l'anglais Vocalic Space Area). Une telle définition laisse toutefois une certaine marge d'interprétation concernant les catégories de voyelles prises en considération dans le calcul de l'aire de l'espace vocalique, et la prise en compte de la variabilité des réalisations au sein de chacune de ces catégories. L'approche la plus courante, qui présente les avantages de cibler les voyelles les répandues dans les systèmes vocaliques des langues du monde (Schwartz et al., 1997) et d'être quantifiable par le simple calcul de l'aire d'un triangle, s'appuie sur les positions des deux premiers formants des voyelles périphériques /a, i, u/. Dans la majorité des études, les positions formantiques de référence qui sont utilisées dans le calcul de l'aire du triangle sont obtenues en considérant les centroïdes de chaque catégorie vocalique, c'est-à-dire la valeur moyenne ou médiane de F1 et de F2 par catégorie. Une variante fréquente pour les applications à l'anglais est de considérer à la fois les voyelles /æ/ et /ɑ/ pour représenter les valeurs de F1 les plus élevées, l'aire de l'espace vocalique étant alors estimée par l'aire du quadrilatère (voir par exemple Flipsen & Lee (2012)).

Pour nos applications à la parole dysarthrique, en complément de l'aire du triangle nous avons considéré le pentagone formé par les voyelles /a, E, i, u, O/, les notations /E/ et /O/ correspondant ici respectivement aux archiphonèmes /e, ɛ/ et /o, ɔ/. L'aire du pentagone vocalique, notée pVSA pour la distinguer de l'aire du triangle /a, i, u/ notée tVSA, peut alors être calculée directement par la formule ci-dessous :

$$\frac{1}{2}\sum_{i=V}(F1_i \cdot F2_{i+1} - F1_{i+1} \cdot F2_i)$$

Toutefois, le calcul direct de l'aire du pentagone, voire d'un polygone constitué d'un nombre plus élevé de points, nécessite pour obtenir un résultat fiable que l'ordre des points soit bien conforme à celui attendu. Or cette condition est d'autant moins garantie que la nombre de catégories vocalique prises en compte est élevé et que les exemplaires de voyelles analysées s'éloignent de productions canoniques hyperarticulées par des locuteurs sains. De plus, si elle peut permettre l'identification d'erreurs de détection des valeurs formantiques, la

vérification systématique de l'ordre des points préalablement au calcul de l'aire du polygone peut s'avérer fastidieuse et sujette à erreurs pour l'analyse de jeux de données comptant un nombre important de locuteurs. Ce constat a conduit certains auteurs à opter pour une méthode de calcul de l'aire vocalique plus centrée sur les données et automatisable, s'appuyant sur le calcul de l'enveloppe convexe à partir des centroïdes de catégories vocaliques (voir par exemple Sandoval et al., 2013).

Dans le contexte de l'étude de la parole dysarthrique, la mesure FCR (Formant Centralization Ratio) a également été proposée par Sapir et al. (2010) pour quantifier le degré de centralisation en ciblant plus spécifiquement certaines voyelles en fonction de l'impact attendu de la centralisation de l'espace vocalique sur leurs valeurs formantiques. Cette métrique repose sur le calcul du rapport entre les valeurs formantiques supposées augmenter avec la centralisation (F1 et F2 de /u/, F1 de /i/) et celles supposées diminuer avec la centralisation (F1 de /a/, F2 de /i/). Pour nos travaux sur le français dans lesquels cinq catégories de voyelles orales ont été considérées, nous avons proposé une version modifiée de cette métrique, baptisée cFCR (custom Formant Centralization Ratio) dans lequel la valeur de F2 de /o/ supposée augmenter avec la centralisation est également prise en compte dans le numérateur, de même que la valeur de F2 de /e/ supposée diminuer avec la centralisation dans le dénominateur.

Bien que nous ne l'ayons pas directement incluse dans nos travaux, on peut également mentionner la métrique VAI (Vowel Articulation Index) proposée par Sapir et al. (Sapir et al., 2011) comme l'inverse du FCR et évaluée ensuite par Caverlé & Vogel (2020) comme plus stable que les mesures d'aire de l'espace vocalique et de FCR, avec toutefois une différence marginale comparativement à FCR.

Enfin, nous avons également inclus dans nos analyses d'autres mesures de centralisation destinées à capturer des modifications spécifiques à l'un des deux premiers formants proposées dans la littérature : la métrique spécifique au premier formant F1RR (F1 Reduction Ratio) qui se concentre sur l'écart de valeurs de F1 entre /a/ d'une part et /i/ et /u/ d'autre part, et la métrique spécifique au deuxième formant F2RR (Sapir et al., 2010) qui évalue l'écart de valeurs de F2 entre /i/ et /u/. La Figure 74 illustre les dimensions capturées par ces deux métriques ainsi que par l'aire du pentagone pVSA.

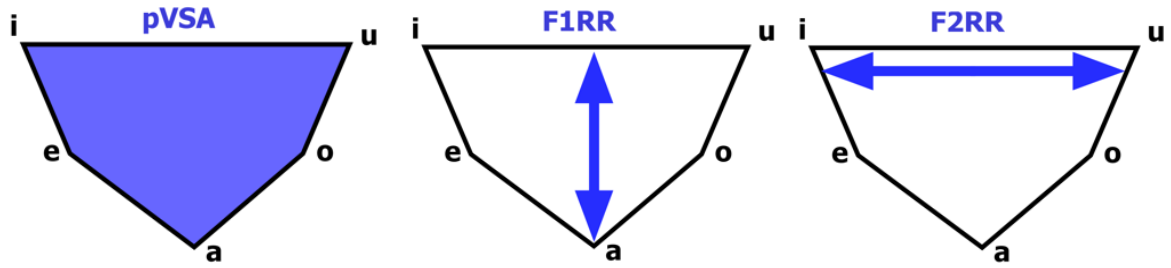


Figure 74 : Représentation schématique des métriques de centralisation pVSA, F1RR et F2RR. Adapté de Audibert et al. (2015, [ACTI34]).

Les métriques destinées à caractériser la dispersion des voyelles au sein d'une catégorie vocalique sont moins nombreuses dans la littérature. On peut principalement mentionner l'indice ϕ proposé par Huet & Harmegnies (2000), dont l'objectif est de caractériser de façon globale le degré d'organisation du système vocalique. Cet indice est calculé comme le rapport entre deux mesures intermédiaires CM_{inter} et CM_{intra}, conçues pour quantifier respectivement

le degré général de dispersion du système via la mesure des distances cumulées entre le centre de l'espace vocalique et les centroïdes de chaque catégorie vocalique, et la variabilité intra-catégorie à travers la distance cumulée entre le centroïde de catégorie et les exemplaires de voyelles appartenant à la même catégorie, pondérées par le nombre d'exemplaires et de catégories. Dans nos travaux, afin de permettre de caractériser séparément le degré de dispersion/centralisation et la variabilité au sein de chaque catégorie nous avons pris en compte ces deux mesures intermédiaires en complément de l'indice ϕ lui-même.

Enfin, afin de caractériser le degré de recouvrement entre catégories vocaliques nous avons introduit une métrique spécifique baptisée tOverlap, calculée comme la valeur cumulée de l'aire de recouvrement entre ellipses de dispersion estimée par échantillonnage pour chaque paire de voyelles du système. En complément, nous avons également pris en compte dans certaines de nos analyses l'aire de recouvrement entre paires de voyelles spécifiques.

On peut relever que dans la littérature, ces métriques sont majoritairement calculées à partir de valeurs formantiques exprimées en Hertz, ce qui pour les métriques calculées à partir des deux premiers formants a pour effet d'accorder un poids plus important aux valeurs de F2 qu'à celles de F1 en raison de leur plus grande variabilité, et dans une moindre mesure de déboucher sur des valeurs plus élevées de ces métriques pour les femmes pour lesquelles les fréquences de résonance sont plus élevées que pour les hommes. Toutefois certaines études ont calculé les métriques relatives à l'espace vocalique à partir de mesures formantiques converties en Bark ou ERB afin que les écarts calculés soient plus conformes à la perception des différences de fréquence, comme par exemple l'étude de Weirich & Simpson (2013) qui a utilisé l'échelle Bark pour le calcul de l'aire de l'espace vocalique. Dans nos travaux, nous avons également fait le choix de calculer les métriques à partir de valeurs formantiques converties en Bark en utilisant la formule de Traunmüller (1990).

7.2.3.2 Métriques calculées à l'échelle de chaque exemplaire

Si elles présentent l'avantage de permettre une interprétation plus directe des valeurs en lien avec le style de parole ou plus généralement la condition dans laquelle les voyelles analysées ont été produites, ou encore les caractéristiques du locuteur, les métriques présentées ci-dessus qui associent à un système vocalique une valeur unique restent des projections obtenues à partir d'un ensemble d'exemplaires. A ce titre elles ne permettent pas d'interpréter si un exemplaire donné doit être considéré comme plus ou moins périphérique, plus ou moins typique de la catégorie vocalique à laquelle il appartient, ou encore si cet exemplaire est plus ou moins susceptible d'être confondu avec un exemplaire d'une autre catégorie. Par ailleurs dans le cadre d'études menées à relativement faible échelle en termes de nombre de locuteurs en raison des critères d'inclusion (typiquement pour les études en phonétique clinique) ou des contraintes expérimentales et quel que soit le nombre de voyelles par locuteur, la représentation de tout un système vocalique par une valeur numérique unique est parfois incompatible avec l'analyse statistique des résultats.

Nous avons donc proposé l'utilisation de métriques destinées à refléter chacune des trois grandes dimensions de la variation vocalique (dispersion/centralisation, variation intra-catégorie et recouvrement entre catégories) en associant une valeur à chaque exemplaire de voyelle présent dans les données analysées. Ces métriques, dont le calcul est intégré dans l'application en ligne présentée en section 7.4.3 et qui pour deux d'entre-elles sont illustrées par la Figure 75, sont définies comme suit :

- La métrique DistCentroid, destinée à rendre compte du degré de centralisation du système vocalique, est définie comme la distance euclidienne de chaque exemplaire au centroïde de l'espace vocalique considéré dans son ensemble. Le choix effectué pour définir la position de ce centroïde global de l'espace vocalique, supposé correspondre à une articulation neutre, a été de considérer la moyenne entre les centroïdes des différentes catégories vocaliques, plutôt que des valeurs formantiques théoriques supposées correspondre à l'articulation d'un schwa qui auraient nécessité de disposer d'informations sur le conduit vocal des locuteurs. Il aurait par ailleurs été envisageable de moyenner les valeurs formantiques de l'ensemble des voyelles du système indépendamment de leur catégorie mais cela aurait impliqué de décaler le centroïde global vers les catégories vocaliques les plus représentées, typiquement /a/ et /e, ɛ/ pour le français dans un corpus non-contrôlé.
- La métrique V-Dispersion, destinée à rendre compte de la variabilité intra-catégorie, est définie comme la distance euclidienne de chaque exemplaire au centroïde de sa catégorie vocalique, c'est-à-dire à la position moyenne de l'ensemble des voyelles du locuteur appartenant à la même catégorie vocalique.
- La métrique ContrastLoss, destinée à rendre compte du recouvrement entre catégories, est calculée à l'aide d'une analyse linéaire discriminante (LDA) selon une méthode inspirée d'Harmegnies & Poch-Olivé (1992). Notons toutefois une différence dans l'exploitation des résultats de la LDA puisque les valeurs attribuées à la métrique ContrastLoss au niveau de chaque exemplaire ne correspondent pas à un taux de confusion, calculable uniquement à l'échelle d'un ensemble d'exemplaires, mais à 1 – la probabilité a posteriori d'appartenance à la catégorie qui correspond à la voyelle (c'est-à-dire à la probabilité de non-appartenance à la catégorie de référence). Ainsi, plus la valeur de ContrastLoss est élevée, plus la probabilité que l'exemplaire soit confondu avec une autre catégorie de voyelle est élevée. On peut donc s'attendre à ce qu'au niveau de la catégorie vocalique ou du locuteur, la valeur moyenne de la métrique ContrastLoss soit fortement liée au taux de mauvaise classification.

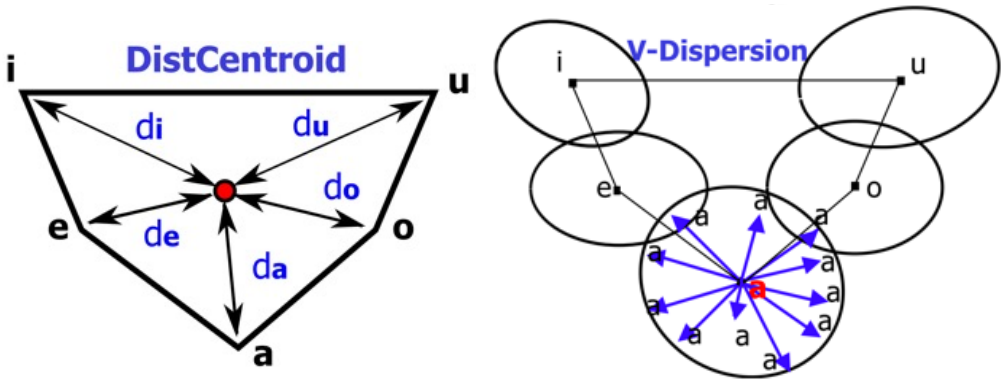


Figure 75 : Représentation schématique des métriques calculables à l'échelle de l'exemplaire DistCentroid qui mesure le degré global de centralisation/dispersion du système vocalique, et V-Dispersion qui mesure la dispersion intra-catégorie. Adapté de Audibert et al. (2015, [ACTI34]).

7.2.3.3 Prise en compte des nasales : les coefficients MFCC comme alternative aux formants

Les métriques DistCentroid, V-Dispersion et ContrastLoss, qui reposent sur des calculs de distance euclidienne et sur l'analyse linéaire discriminante à partir de coordonnées dans un espace multidimensionnel, sont illustrées sur la Figure 75 dans le cas de l'application à l'espace

bidimensionnel des fréquences des deux premiers formants mais peuvent être calculées de la même façon dans l'espace des trois premiers formants pour rendre compte par exemple de l'opposition d'arrondissement des voyelle antérieures du français. Elles peuvent aussi être appliquées à d'autres représentations acoustiques dans un espace multidimensionnel d'un ensemble de voyelles voire de consonnes à la seule condition que la réalisation de l'ensemble de sons pris en considération puisse être décrit par un même jeu de paramètres acoustiques.

En raison de la présence d'anti-formants (Fujimura, 1962), les voyelles nasales rendent l'analyse formantique particulièrement complexe (voir par exemple de Wet et al. (2004), ou Kent & Vorperian (2018) pour une revue des principales sources d'erreur dans la détection des formants). Dans les langues comme le français qui incluent des voyelles nasales, les analyses de l'espace vocalique se concentrent donc généralement sur les voyelles orales. Si la limitation aux voyelles orales n'empêche pas a priori l'estimation de la centralisation, elle peut ne fournir qu'un aperçu partiel de la variation intra-catégorie et surtout ne pas rendre compte fidèlement du recouvrement entre catégories en ne considérant qu'un sous-ensemble du système vocalique.

Afin de dépasser cette limitation, les caractéristiques spectrales des voyelles peuvent être représentées par un ensemble de coefficients MFCC mesurables automatiquement, selon une approche similaire à celle adoptée par Ferragne & Pellegrino (2010).

7.2.4 Délexicalisation pour l'étude perceptive d'énoncés non-contrôlés

Publications associées :

[ACL2] Petrone, C., Carbone, F, **Audibert, N.**, & Champagne-Lavau, M. (2024). Facial cues to anger affect meaning interpretation of subsequent spoken prosody. *Language & Cognition.* Published online 2024:1-24. doi:10.1017/langcog.2024.3

[ACTI9] **Audibert, N.**, Carbone, F., Champagne-Lavau, M., Said Housseini, A., & Petrone, C. (2023). Evaluation of delexicalization methods for research on emotional speech. *Proceedings of Interspeech 2023*, Dublin, Ireland. pp. 2618-2622

7.2.4.1 Contexte et motivations pour le développement d'une méthode de délexicalisation

En 2020, j'ai été sollicité par ma collègue Caterina Petrone du Laboratoire Parole et Langage d'Aix-en-Provence pour apporter ma contribution au projet COMEDIA dont elle était porteuse avec Maud Champagne-Lavau et dans lequel une post-doctorante venait d'être recrutée. Dans le cadre de ce projet qui portait sur l'expression émotionnelle produite par des acteurs expérimentés et inexpérimentés, elles avaient supervisé le travail d'étudiantes en cinéma chargées de sélectionner des expressions de colère chaude produites dans des films par des acteurs francophones, complétées par des expressions émotionnellement neutres produites par les mêmes acteurs. Toutefois, de même que dans le cas d'expressions émotionnelles spontanées, l'analyse acoustique et perceptive de tels énoncés se heurte au manque de comparabilité directe entre expressions neutres et expressions émotionnelles (ou dans un cadre plus large, entre catégories émotionnelles). En effet, outre les problèmes posés par l'interprétation des mesures acoustiques comparées entre énoncés sélectionnés en tant que représentants de catégories émotionnelles particulières mais de contenus segmentaux différents, l'exploitation conjointe par les auditeurs lors du processus d'attribution

émotionnelle des informations prosodiques et de qualité de voix d'une part, et lexicales et sémantiques d'autre part est bien établie (Ben-David et al., 2016).

Cette impossibilité d'assurer le principe de la comparaison « toutes choses égales par ailleurs », essentiel en sciences expérimentales et particulièrement sensible en phonétique, a conduit de nombreux auteurs à recourir à des expressions simulées par des acteurs sur un contenu lexical et sémantique fixé. Toutefois le caractère naturel des expressions émotionnelles simulées dans de tels contextes peu propices à l'expression du jeu d'acteur a fait l'objet de nombreuses critiques (voir par exemple Batliner et al. (2000)). Ce constat, combiné à l'impossibilité de recourir à une variation systématique pour l'étude d'expressions qui par définition ne peuvent être observées qu'en contexte comme par exemple l'étude du charisme dans le voix et la parole (Signorello et al., 2020; Niebuhr et al., 2020), a suscité au cours des 25 dernières années un regain d'intérêt pour l'analyse d'expressions spontanées dont le contenu lexical et sémantique n'est pas contrôlé. Dans un tel contexte, l'évaluation de l'information affective portée par les variations prosodiques et spectrales nécessite alors de s'abstraire le contenu lexical en présentant aux auditeurs des stimuli délexicalisés, l'objectif étant de préserver les informations prosodiques tout en éliminant l'information lexicale. Dans le cadre du projet COMEDIA, une telle approche était nécessaire afin d'évaluer perceptivement les informations émotionnelles relatives aux expressions de colère portées par les variations prosodiques et spectrales.

Plusieurs méthodes de délexicalisation ont été proposées dans la littérature, qui en dépit de leur efficacité avérée pour éliminer les informations lexicales présentent l'inconvénient de dégrader fortement la qualité du signal de parole en le rendant peu naturel (voir par exemple Sonntag et Portele (1998) pour une comparaison entre plusieurs de ces méthodes). L'approche la plus couramment utilisée a été le recours à un filtrage passe-bas afin d'éliminer les informations segmentales présentes en moyennes et hautes fréquences. Un tel filtrage ne conserve que les modulations de fréquence fondamentale et une partie des modulations d'intensité, la fréquence de coupure utilisée dans la plupart des études étant le plus souvent fixée arbitrairement à des valeurs telles que 400 Hz (Magen, 1998) ou 600 Hz (Niebuhr et al., 2020), ce qui peut s'avérer problématique pour des expressions émotionnelles fortement activées et caractérisées par des excursions importantes de fréquence fondamentale. Obin et al. (2011) ont quant à eux opté pour une fréquence de coupure dynamique afin d'inclure la première harmonique, en accord avec les recommandations de MacCallum et al. (2011).

Le manque de naturalité des stimuli transformés a conduit Pagel et al. (1996) à proposer une méthode de délexicalisation fondée sur l'analyse-resynthèse à l'aide d'un système de synthèse par diphone, combiné à la substitution de phones afin de masquer le contenu lexical et à la copie prosodique pour préserver les variations de fréquence fondamentale et éventuellement celles d'intensité. Cette méthode a notamment été utilisée pour générer la condition *saltanaj* dans les travaux de Ramus et Mehler (1999) sur l'identification de la langue à partir d'indices suprasegmentaux. La principale limite de cette approche par analyse/resynthèse est que la qualité vocale et les autres indices spectraux présents dans les stimuli délexicalisés sont ceux de la parole lue hyperarticulée produite dans la base de diphones utilisée. Cela peut se révéler problématique pour l'étude de la parole expressive, en particulier les expressions émotionnelles pour lesquelles les indices de qualité vocale peuvent être particulièrement saillants (Gobl & Ní Chasaide, 2003). Par ailleurs, les rares méthodes de délexicalisation proposées dans la littérature visant à préserver les indices de qualité de voix s'appuient sur un processus de filtrage inverse sujet à erreur et de mise en œuvre complexe

(Vainio et al., 2009), ou reposent sur le postulat discutable pour les expressions émotionnelles d'une qualité de voix constante au cours du temps (Kain & van Santen, 2010). Afin de dépasser ces limites et de répondre au défi posé par la délexicalisation d'expressions de colère chaude fortement activée dans lesquelles les modifications spectrales véhiculent une quantité importante d'information affective, j'ai été amené à proposer et évaluer une nouvelle méthode de délexicalisation pouvant être mise en œuvre de façon entièrement automatisée.

7.2.4.2 Description de la méthode de délexicalisation proposée

La méthode proposée reprend le principe d'analyse-resynthèse via un système de synthèse par diphone avec substitution de phones et copie prosodique (Pagel et al., 1996), mais introduit une étape supplémentaire de conversion de voix à l'aide d'un outil de morphing vocal avant la copie prosodique. Le système MBROLA (Dutoit et al., 1996) est utilisé pour réaliser la synthèse par diphone en utilisant une voix française d'homme ou de femme en fonction du locuteur ayant produit l'énoncé original, en appliquant les règles de substitutions de phones par classe phonémique utilisées par Ramus et Mehler (1999) dans la condition *saltanaj*. La conversion de voix est effectuée à l'aide de STRAIGHT (Kawahara, 2006), afin de reconstruire partiellement les caractéristiques spectrales des stimuli originaux et leur évolution temporelle. En effet, une reconstruction complète des caractéristiques spectrales des stimuli originaux aurait pour conséquence de réinjecter dans les stimuli générés l'information segmentale et donc l'information lexicale et sémantique. Après des essais initiaux avec différentes valeurs de taux de morphing, ce dernier a été fixé à 0,5 comme compromis entre un taux inférieur qui aurait éliminé la plus grande partie de l'information sur la qualité de la voix et un taux supérieur qui aurait également reconstruit l'information segmentale des stimuli d'origine. Afin de conserver dans les stimuli délexicalisés l'évolution temporelle des variations spectrales présentes dans les stimuli d'origine, la conversion de voix avec STRAIGHT a été effectuée par trames de 25 ms appariées temporellement entre le stimulus d'origine et celui resynthétisé par MBROLA, les trames modifiées par STRAIGHT étant ensuite combinées au moyen d'une transformée de Fourier à court terme (STFT) afin d'obtenir le stimulus délexicalisé. Enfin, une copie prosodique a été effectuée par l'algorithme TD-PSOLA (Moulines & Charpentier, 1990) afin de transférer au stimulus délexicalisé les variations de fréquence fondamentale du stimulus d'origine, et le contour d'intensité du stimulus délexicalisé a été modifié pour reproduire celui du stimulus d'origine.

Les stimuli générés avec cette méthode sont désignés ci-après comme « salatanaj+morphing ». A des fins de comparaison avec la méthode plus classique du filtrage passe-bas, cette méthode de délexicalisation a également été appliquée aux mêmes stimuli, avec une fréquence de coupure fixée à deux demi-tons au-dessus de la fréquence fondamentale maximale mesurée dans le stimulus, soit une fréquence de coupure comprise entre 136 Hz et 678 Hz (373 Hz en moyenne) pour les 18 stimuli originaux ainsi traités et inclus dans l'évaluation. La Figure 76 illustre les deux versions délexicalisées comparées à la version originale pour une expression de colère chaude produite par un locuteur masculin.

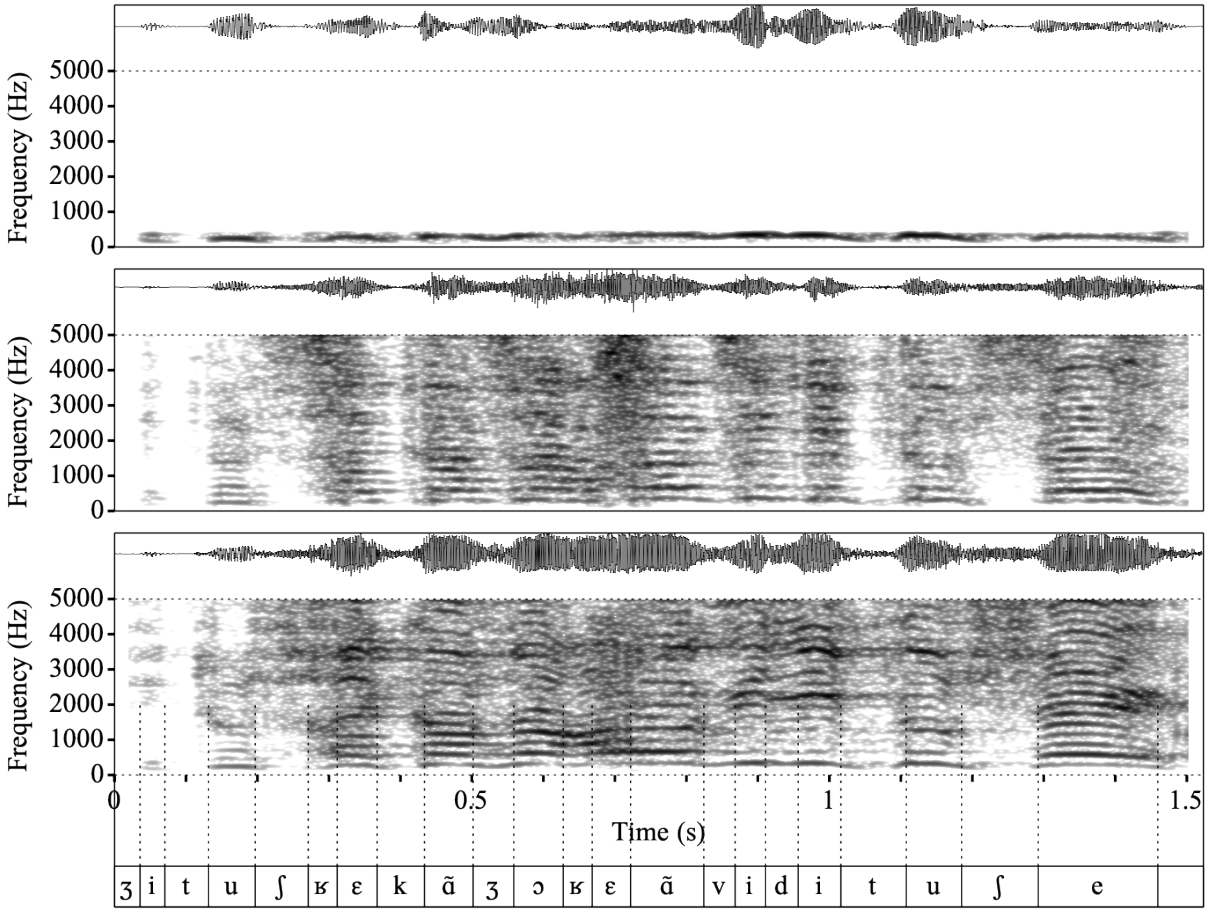


Figure 76 : Oscillogramme et spectrogramme à bande large des trois versions d'une expression de colère sur l'énoncé « J'y toucherai quand j'aurai envie d'y toucher ! » produit par un acteur français. En bas : version originale ; au milieu : version délexicalisée dans la condition « saltanaj_morphing » ; en haut : stimulus délexicalisé par filtrage passe-bas.

7.2.4.3 Evaluation de la délexicalisation, de l'attribution émotionnelle et de la naturalité

Neuf stimuli exprimant de la colère chaude ont été sélectionnés à partir d'un ensemble de départ de 73 stimuli extraits de films français. Chacun de ces stimuli a été apparié avec une expression neutre produite par le même acteur dans le même film, pour un total de 18 stimuli originaux sélectionnés.

Les stimuli synthétisés en condition « saltanaj+morphing » ont été validés par trois tâches perceptives. Une tâche d'évaluation de l'intelligibilité a tout d'abord été mise en place pour s'assurer que l'ajout d'informations spectrales à la condition de base « saltanaj » pour obtenir la condition « saltanaj+morphing » ne permettait pas aux auditeurs d'identifier le contenu lexical des stimulus. 47 auditeurs francophones natifs (38 femmes et 9 hommes, âgés en moyenne de 24,7 ans) ont participé à cette tâche dans laquelle la consigne était de transcrire orthographiquement les stimuli présentés en ordre aléatoire. Les transcriptions proposées par les auditeurs ont ensuite été analysées manuellement par des étudiantes orthophonistes afin de comptabiliser le nombre de mots reconnus par rapport à la transcription originale de référence, en ignorant les éventuelles modifications ou erreurs orthographiques dans les transcriptions. La proportion de mots identifiés était inférieure à 30% dans l'ensemble des 18 stimuli synthétisés en condition « saltanaj+morphing » (10% de mots reconnus en moyenne),

validant ainsi l'efficacité du processus de délexicalisation. Dans une tâche équivalente sur les stimuli délexicalisés par filtrage passe-bas auprès de 12 auditeurs francophones natifs (7 femmes et 5 hommes, âgés en moyenne de 21,9 ans), les auditeurs n'ont jamais été en mesure de proposer une transcription liée au stimulus original, d'où un taux d'intelligibilité de 0%.

Une seconde tâche de validation perceptive portant sur les stimuli délexicalisés en condition « saltanaj+morphing », à laquelle ont participé 39 auditeurs francophones natifs n'ayant pas pris part à la tâche d'évaluation de l'intelligibilité (21 femmes et 18 hommes, âgés en moyenne de 30,3 ans), a consisté en une tâche d'attribution émotionnelle en choix forcé dans laquelle les auditeurs devaient indiquer si l'extrait audio présenté correspondait à une expression de colère ou à une expression neutre. L'ensemble des 18 stimuli a été identifié par plus de 70% des auditeurs.

Enfin la troisième et dernière tâche de validation perceptive, à laquelle ont participé 39 auditeurs francophones natifs (34 femmes et 5 hommes, âgés en moyenne de 21,7 ans), a consisté en des jugements de naturalité des voix présentées sur une échelle de Likert à 5 points allant de « pas du tout naturelle » à « parfaitement naturelle ». Cette dernière tâche incluait les stimuli originaux en tant que référence suivant le protocole couramment utilisé pour l'évaluation de la qualité des systèmes de synthèse vocale, ainsi que les stimuli délexicalisés à l'aide du filtrage passe-bas et en condition « saltanaj+morphing », soit 54 stimuli au total. Comme l'illustre la Figure 77, bien que les stimuli délexicalisés à l'aide de la méthode « saltanaj+morphing » aient été jugés comme significativement moins naturels que les énoncés originaux, ils ont été évalués comme significativement plus naturels que ceux délexicalisés à l'aide d'un filtrage passe-bas.

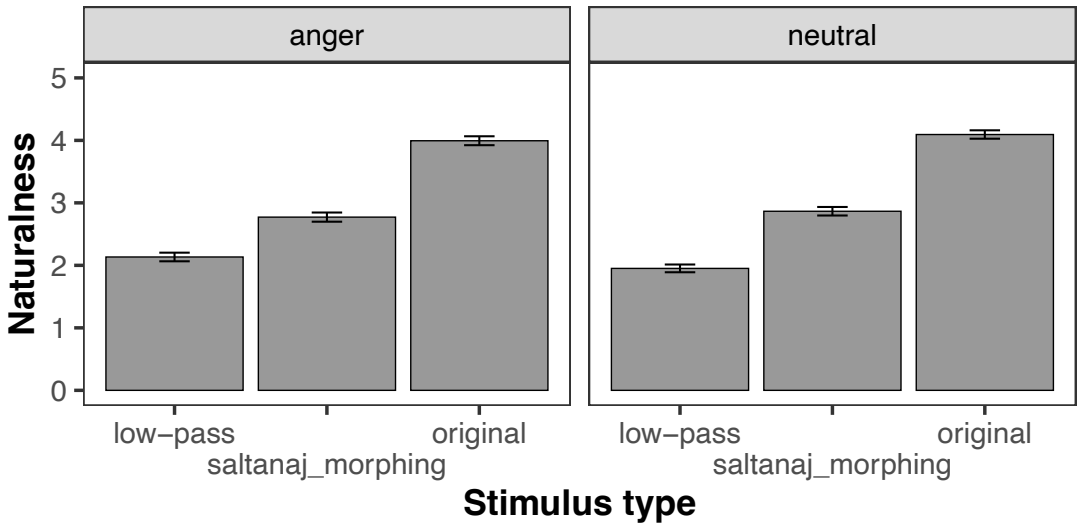


Figure 77 : Jugements moyens de naturalité recueillis auprès de 39 auditeurs francophones natifs à partir des neuf stimuli exprimant de la colère et des neuf stimuli neutres sélectionnés : chacun déclinés en trois versions : la version délexicalisée par filtrage passe-bas à gauche, celle délexicalisée à l'aide de la méthode proposée « saltanaj+morphing » d'analyse-resynthèse avec permutation de phonèmes eu centre, et la version originale non-modifiée à droite. Les barres d'erreur représentent l'erreur-type. D'après Audibert et al. (2023, [ACTI9]).

7.2.4.4 Implications pour la recherche sur la parole expressive et perspectives

L'un des avantages de la nouvelle méthode proposée par rapport aux méthodes de délexicalisation classiques telles que le filtrage passe-bas qui reposent sur un codage minimal de l'information suprasegmentale est qu'elle combine une délexicalisation efficace avec la préservation de l'information affective. Elle permet également de générer des stimuli synthétiques plus similaires à des énoncés naturels que les méthodes classiques de délexicalisation. En effet, le degré plus élevé de naturalité obtenu minimise la probabilité que les extraits délexicalisés soient traités comme des stimuli non-vocaux par les auditeurs, ce qui implique un traitement cérébral différent de celui des sons de parole (voir par exemple Palva et al. (2002)). Au-delà des perspectives prometteuses qu'ouvre cette première évaluation de la méthode proposée pour l'évaluation perceptive des émotions et autres affects, l'extension de son application aux études sur la perception de la prosodie linguistique en interaction avec la variation spectrale peut également être envisagée.

Contrairement à la méthode proposée par Vainio et al. (2009) qui s'appuie sur une modélisation de l'onde de débit glottique dans les stimuli originaux pour reproduire les variations de qualité de voix, la méthode « saltanaj+morphing » proposée ne nécessite pas de référence neutre et est donc également applicable aux extraits pour lesquels une comparaison directe avec d'autres productions du même locuteur n'est pas possible. De plus, elle permet la restitution partielle des changements spectraux qui ne sont pas directement liés à la forme d'onde du flux glottique mais qui sont pris en compte par certains auteurs dans une définition plus large de la qualité de la voix (Laver, 1980).

La principale limite en revanche est que, selon le niveau de morphing appliqué, certaines informations segmentales peuvent subsister, raison pour laquelle nous avons choisi de n'appliquer qu'un morphing partiel. Il est donc recommandé dans les utilisations futures de cette méthode d'inclure dans la procédure une vérification que le processus de délexicalisation entraîne effectivement une perte d'intelligibilité.

Avec la méthode que nous proposons, le degré de naturalité pourrait être amélioré en exploitant pleinement le potentiel d'un outil de morphing de la parole tel que STRAIGHT (Kawahara, 2006), ce qui n'est pas le cas avec cette première version. Dans cette étude, nous avons choisi un processus qui peut être entièrement automatisé à condition de disposer d'un alignement en phones des stimuli originaux. Cela induit quelques distorsions spectrales audibles dans les stimuli synthétisés, qui pourraient être réduites en ajustant manuellement les points d'ancrage temporels afin d'obtenir des stimuli délexicalisés plus proches des originaux en termes de degré de naturalité. Une autre piste, complémentaire, pour améliorer le degré de naturalité des stimuli générés serait d'exploiter, lorsque cela est possible, des productions neutres du même locuteur en remplacement de la resynthèse MBROLA comme appliqué dans un autre cadre par Dubeda (2006), ou en tant que base de diphones alternative.

Néanmoins on peut supposer un biais lexical dans les jugements de naturalité, le choix d'une délexicalisation de type « saltanaj » impliquant de générer des énoncés composés de pseudo-mots susceptibles d'être jugés comme moins naturels que des énoncés interprétables quel que soit le degré de naturalité de la voix proprement dite. Pour certaines applications, il pourrait ainsi être intéressant d'opter pour des énoncés sémantiquement neutres en tant que cible de la délexicalisation plutôt que pour l'application systématique des règles de permutation de phones.

7.3 Recueil et documentation de données

Comme j'ai pu le développer par ailleurs et quelle que soit leur thématique parmi celles sur lesquelles j'ai travaillé, mes recherches s'appuient sur l'analyse de données. Ainsi, mes travaux présentés dans les autres sections de ce document de synthèse et qui ne reposent pas sur des données déjà disponibles et annotées incluent une part de recueil et/ou de documentation de données. Je fais le choix de ne présenter de façon plus détaillée dans cette section que certains corpus auxquels j'ai directement contribué et susceptibles de faire l'objet d'analyses plus larges que celles déjà publiées.

7.3.1 Parole dysarthrique

Publication associée :

[ACTI55] Fougeron, C., Crevier Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., **Audibert, N.**, Bonastre, J.-F., Colazo Simon, A., Delooze, C., Duez, D., Gendrot, C., Legou, T., Lévêque, N., Pillot-Loiseau, C., Pinto, S., Pouchoulin, G., Robert, D., Vaissière, J., Viallet, F., & Vincent, C. (2010). The *DesPho-APaDy* Project: Developing an acoustic-phonetic characterization of dysarthric speech in French. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte, pp. 2831-2838.

Dans le cadre du projet ANR DesPho-APaDy auquel j'ai contribué, l'un des objectifs était le recueil structuré de données de de parole dysarthrique afin de permettre la caractérisation acoustico-phonétique de ces productions pathologiques éminemment variables du fait de la diversité des pathologies associées et de la manifestation des symptômes. Ayant rejoint le projet après son démarrage à l'occasion de mon séjour post-doctoral au Laboratoire d'Informatique d'Avignon, je n'ai pas directement participé à la collecte des données issues de sources multiples et incluant des signaux physiologiques, ni à leur organisation dans une base de données. Je ne reviens donc ici que sur les conventions de transcription orthographique adoptées pour l'annotation des données issues des tâches de lecture et des productions spontanées, les données collectées incluant également par ailleurs des voyelles tenues, diadococinésies et autres tâches destinées à évaluer les performances des 1850 patients enregistrés, comparés à 240 sujets contrôles.

Les transcriptions produites en suivant ces conventions qui ont servi de base à l'alignement forcé que j'ai été amené à utiliser et évaluer (Audibert et al. (2010, [ACTI54]) et Fougeron et al. (2010, [ACTI56]), présentés en section 7.2.1) ainsi qu'aux autres analyses auxquelles j'ai contribué ensuite. Bien que l'essentiel de ces conventions aient été définies en amont afin de cadrer autant que possible la tâche confiée aux transcripteurs, de telles conventions, aussi précises soient elles, impliquent toujours une marge d'interprétation. Je reviens donc ici brièvement sur les implications du choix de ces conventions de transcription sur le fonctionnement du système d'alignement, mais aussi sur la variabilité que j'ai pu observer dans mes travaux exploitant ces transcriptions ainsi que sur certains choix que j'ai été amené à faire en effectuant moi-même certaines transcriptions, en discutant l'incidence que cette variabilité dans l'interprétation des conventions peut avoir sur les mesures.

Les conventions de transcription ont été élaborées comme un compromis entre la qualité de l'alignement phonétique visée d'une part, et les contraintes imposées par les troubles de la parole dus à la dysarthrie d'autre part. Ainsi, pour tenir compte des divergences possibles entre la phonétisation associée à une forme orthographique dans le dictionnaire de

prononciation intégré dans le système d'alignement et la réalisation du mot correspondant, ces conventions de transcription incluent la possibilité de spécifier en alphabet SAMPA la prononciation effective par le locuteur en l'associant à la forme orthographique. Cela a pour effet d'ajouter au dictionnaire de prononciation cette nouvelle forme phonétisée en tant que variante de prononciation, permettant ainsi au système d'alignement de considérer cette prononciation comme l'une des phonétisations possibles du mot (Adda-Decker & Lamel, 2000). Si l'ajout de variantes de prononciation pour une même forme orthographique permet de mieux rendre compte des réalisations produites par les locuteurs, la multiplication de ces variantes peut aussi dégrader les performances du système d'alignement en complexifiant l'espace de probabilité dans lequel la correspondance optimale entre le signal acoustique et la segmentation en phones doit être trouvé. Afin d'éviter d'introduire du bruit à travers un nombre excessif de variantes de prononciation, le choix s'est donc porté ici sur l'ajout dynamique au cas par cas de variantes de prononciation pour chaque locuteur en fonction des formes phonétisées non-standard présentes dans la transcription plutôt que sur une approche cumulative.

Toutefois, afin de ne pas trop complexifier la tâche des transcripteurs, un seuil de degré de divergence à partir duquel la prononciation par le locuteur fait l'objet d'une transcription SAMPA a été défini au niveau de la syllabe. Ainsi, les transcripteurs avaient pour consigne de n'utiliser la notation spécifique réservée à l'introduction d'une prononciation non-standard uniquement lorsqu'une syllabe ou plus était omise, ajoutée ou substituée dans la réalisation d'un mot en comparaison de sa prononciation canonique déjà incluse dans le dictionnaire de prononciation utilisé par défaut. Des notations spécifiques ont été introduites pour indiquer les cas de délétion, d'ajout ou de substitution pour permettre de systématiser l'analyse des occurrences des réalisations non-standard dans la parole dysarthrique au-delà de l'objectif d'assurer le bon fonctionnement du système d'alignement forcé pour lequel la délétion d'un mot complet pourrait faire l'objet d'une simple suppression dans la séquence de mots fournie en entrée.

Si cette règle de ne pas coder les écarts à la norme inférieurs au niveau de la syllabe s'est révélée indispensable pour rendre réalisable la tâche des transcripteurs, déjà particulièrement complexe sur des productions parfois très altérées dans le cas des patients sévèrement dysarthriques, elle implique nécessairement une certaine imprécision de l'alignement en phones obtenu. En effet si certaines formes de variations à la fois fréquentes et prédictibles à partir de la structure syllabique comme la réalisation ou non des schwas ou encore la délétion des consonnes liquides finales situées après une obstruante (voir par exemple Dell (1995), ou Avanzi (2023) pour une étude sur corpus) peuvent être directement intégrées dans les dictionnaires de prononciation de même que certaines formes de réduction de mots ou séquence de mots parmi les plus courantes (Adda-Decker & Snoeren, 2011), la dysarthrie peut aboutir à d'autres déviations par rapport aux formes standard qui n'affectent pas toujours une syllabe complète comme par exemple pour les consonnes l'occlusion incomplète des occlusives ou le dévoisement des sonores (Antolík & Fougeron, 2013). Dans certains cas ces distorsions de l'articulation dues à la dysarthrie auraient été suffisamment importantes pour justifier une adaptation de la transcription SAMPA, mais trop localisées pour être codées comme telles. Si ces déviations localisées n'ont qu'une incidence limitée sur la précision temporelle des frontières segmentales attribuée par l'alignement forcé, elles peuvent remettre en question la validité de certaines mesures acoustiques qui ne sont interprétables que lorsqu'elles sont appliquées à certains types de segments. En effet contrairement aux cas dans

lesquels la transcription fait l'objet d'une correction via la transcription SAMPA, ces consonnes sont alors étiquetées suivant la prononciation standard intégrée dans le dictionnaire de prononciation.

Par ailleurs et bien que les conventions aient été définies avec un niveau élevé de précision, j'ai pu observer également quelques divergences d'interprétation entre transcripteurs qui ont fait l'objet de discussions entre participants au projet sans toujours aboutir à un consensus clair au-delà de décisions au cas par cas. Une part importante de ces incertitudes concernait la délimitation des parties devant faire l'objet d'une transcription SAMPA, en particulier dans le cas des distorsions affectant les segments adjacents aux syllabes omises, insérées ou substituées, ou localisées sur une syllabe répartie entre deux mots consécutifs. Dans ce cas également l'incidence sur les frontières temporelles était limitée, mais de telles divergence d'interprétation ont un impact direct sur l'inventaire des segments pris en compte dans les analyses acoustiques. De façon moins fréquente mais avec une incidence plus importante sur l'alignement, j'ai pu également observer des divergences d'interprétation portant sur les continuums jugés impossibles à segmenter. Si la présence d'un tel continuum ne faisait généralement pas de doute en dépit du biais lexical dû à la connaissance par les transcripteurs des textes lus, leur délimitation précise pouvait en effet s'avérer plus complexe.

7.3.2 Expression d'attitudes hostiles dans la parole politique

Publication associée :

[ACTI31] Kouklia, C., & **Audibert, N.** (2017). Relationships Between Speech Timing and Perceived Hostility in a French Corpus of Political Debates. *Proceedings of Interspeech 2017*, Stockholm, Suède, pp. 899-903.

Le recueil et l'annotation d'un corpus d'expressions d'attitudes hostiles dans la parole politique s'est inscrit dans le cadre de la thèse de Charlotte Kouklia (2019) que j'ai codirigée avec Jacqueline Vaissière et dont certains parties dans lesquelles je suis intervenu plus directement et qui ont donné lieu à des publications sont présentées en section 2. Si j'ai contribué de façon plus importante à l'élaboration et la mise en place de la procédure d'analyse des données, le mérite de la conception et du recueil de ce corpus original revient à la doctorante, mon rôle en la matière s'étant limité à des conseils méthodologiques pour l'annotation des données et leur validation perceptive. Je ne ferai donc ici qu'un bref résumé des principales caractéristiques de ce corpus.

Les données de ce corpus sont issues de séances du conseil municipal de Montreuil en 2013, enregistrées et diffusées sur une télévision locale. Le choix de cette source de données a été fait pour plusieurs raisons, la première étant sa densité en expressions d'hostilité par des personnes politiques. En effet, suite au basculement de la majorité municipale lors des élections précédente, le climat était particulièrement conflictuel au point que des échanges à la limite de l'agression physique soient relatés par la presse. Ce climat conflictuel avait mené à l'installation d'un système de régulation de tours de parole, qui empêchait de fait la superposition de tours de parole dans les données enregistrées et diffusées, tout en maintenant un contexte écologique de production. Les données traitées étaient issues de 24h d'archives, annotées et croisées avec les notes prises par la doctorante pendant les séances, pour une présélection de 200 extraits produits par cinq locuteurs appartenant à deux groupes opposés, sélectionnés en raison de la récurrence et de la spontanéité de leurs interventions.

Suite à une première étape de validation perceptive, une sélection de 125 stimuli a été retenue.

L'originalité de ce corpus tient en partie à la condition de relecture par les locuteurs et à leur auto-évaluation de l'hostilité exprimée, réalisées sur ces 125 stimuli sélectionnés. La condition de relecture a permis ensuite une comparaison directe avec les productions spontanées afin de pallier l'absence de contrôle du contenu lexical et la variabilité des structures prosodiques. Pour cette relecture, afin de limiter la variabilité de l'interprétation entre locuteurs et entre stimuli, la consigne donnée aux locuteurs a été de lire à la manière d'une dictée, de façon « neutre » et la plus intelligible possible les transcriptions des extraits.

En complément, une segmentation en phones corrigée manuellement a été effectuée à partir d'une première passe automatique. Les stimuli ont également été annotés en constituants syntaxiques afin de pouvoir croiser ces informations avec celles issues de l'analyse prosodique.

7.3.3 Documentation de la variation intra-locuteur

Publication associée :

[ACTI16] Fougeron, C., **Audibert, N.**, Gendrot, C., Chardenon, E., Wohmann, L. (2022). PATATRA and PATAFreq: two French databases for the documentation of within-speaker variability in speech. *Proceedings of the 13th Conference on Language Resources and Evaluation* (LREC 2022), pp. 1939–1944.

7.3.3.1 Contexte et objectifs des deux corpus

Ces deux corpus dédiés à l'étude de la variabilité entre sessions plus ou moins espacées dans le temps des productions des mêmes locuteurs, et qui font l'objet d'autres travaux présentés dans la section 2 de ce volume, s'inscrivent dans le cadre d'un projet de longue haleine. Les enregistrements du corpus PATATRA (Parole AdulTe A TRavers les Ages) auquel je participe également en tant que locuteur ont en effet débuté en 2013 au Laboratoire de Phonétique et Phonologie et se poursuivent d'année en année. L'objectif de ce corpus et de la variante PATAFreq qui inclut les mêmes locuteurs et se concentre sur des répétitions nombreuses à intervalles de temps plus rapprochés est ainsi de documenter l'évolution dans le temps des productions de voix et de parole d'un même locuteur sur les mêmes tâches, ainsi que leur fluctuation à plus court terme. Contrairement à d'autres corpus qui se focalisent sur l'étude de la variabilité intra-locuteur à partir d'un grand nombre de locuteurs mais un nombre de sessions limité, comme par exemple celui de Keating et al. (2019) avec 200 locuteurs et 12 tâches de parole pour trois sessions d'enregistrement, les corpus PATATRA et PATAFreq se concentrent sur un nombre plus restreint de locuteurs (respectivement 11 et 9) mais incluent un nombre de sessions plus important (7 à 18 selon les locuteurs) avec des délais variables entre enregistrements. On peut noter qu'il ne s'agit pas de locuteurs naïfs puisque tous sont des chercheurs en phonétique, ce qui pourrait avoir pour conséquence de minimiser la variabilité dans certaines tâches, avec des productions potentiellement plus contrôlées que celles de locuteurs naïfs.

7.3.3.2 Le corpus PATATRA : enregistrements annuels en chambre sourde

Lors de la soumission de l'article de présentation des corpus au début de l'année 2022, le corpus PATATRA comptait un total de 92 sessions d'enregistrement, produites par 11 locuteurs

(7 femmes et 4 hommes) âgés de 36 à 68 ans lors de la première année d'enregistrement, enregistrés chaque année sauf exception. Parmi les 11 locuteurs, un homme et une femme ne pas francophones natifs. Tous les locuteurs n'ayant pas débuté les enregistrements la même année et certains ayant manqué une session, l'une des locutrices ne comptait que 5 sessions d'enregistrement tandis que les autres locuteurs en comptaient 8 ou 9. Deux autres sessions se sont depuis ajoutées en 2022 et 2023, et une troisième est en cours d'enregistrement.

Hormis les sessions de 2016 et 2020 qui n'incluent que l'acoustique, pour toutes les sessions le signal électroglottographique a été enregistré de façon synchrone en complément du signal acoustique, les enregistrements ayant été réalisés en chambre sourde à l'exception de la session de 2020 remplacée par la première session de PATAFreq enregistrée à domicile en raison de la pandémie de Covid-19 (avec donc une unique lecture du texte et une possibilité de comparaison directe limitée à certaines tâches, cf. la section suivante pour les détails du protocole utilisé dans le corpus PATAFreq). Chaque session d'enregistrement est complétée par une auto-évaluation de la qualité de la voix du locuteur le jour de l'enregistrement à l'aide d'une version française du questionnaire Voice Handicap Index (Woisard et al., 2004) ainsi que par des informations sur l'usage habituel de la voix, la pratique sportive, les éventuelles habitudes de consommation de tabac et d'alcool et sur les infections ORL survenues au cours de l'année écoulée.

Chaque session inclut trois lectures avec un rythme, une intensité et une hauteur confortables de la fable « La bise et le soleil », dont la version dans différentes langue est d'un usage courant en phonétique, avec la consigne de répéter la phrase complète en cas de disfluences. Elle comprend également trois lectures d'une liste de 56 mots monosyllabiques français avec une pause silencieuse entre mots consécutifs, le premier et le dernier étant des distracteurs destinés à éviter la production des mots-cibles avec une prosodie différente des autres mots de la liste. Les 54 mots restants constituent des triplets minimaux de forme CVC avec la voyelle /a/, /i/ ou /u/. Cette tâche est conçue pour capturer l'espace vocalique /a,i,u/ et divers effets coarticulatoires entre voyelles et consonnes. Enfin, la dernière tâche de production de parole consiste en la production d'environ cinq minutes de discours spontané avec la personne chargée de l'enregistrement (cette tâche ayant été successivement à une orthophoniste, un post-doctorant puis deux doctorantes), dont le rôle est d'animer la conversation en posant des questions sur des sujets libres, par exemple les dernières vacances du locuteur.

Ces trois tâches de production de parole sont complétées par deux tâches destinées à l'analyse de la voix. Deux glissandi montants à faible et forte intensité sont recueillis afin de capturer certains aspects du registre dynamique de la voix. Le locuteur est invité à produire une sirène de la fréquence la plus basse à la plus haute sur une voyelle /a/, d'abord avec l'intensité la plus basse possible, puis avec l'intensité la plus haute possible. Enfin, une tâche destinée à mesurer le temps de phonation maximal du locuteur est effectuée. Le locuteur est invité à produire, après une inspiration profonde, une voyelle /a/ aussi longue que possible, à une hauteur confortable et une intensité confortable. Trois essais consécutifs sont enregistrés, le /a/ le plus long parmi les trois produits par le locuteur étant conservé comme estimation du contrôle pneumo-phonatoire.

7.3.3.3 Le corpus PATAFreq : enregistrements fréquents à domicile

Le corpus PATAFreq a été créé lors du premier confinement en 2020 par les neuf locuteurs natifs du corpus PATATRA qui se sont enregistrés régulièrement, avec un objectif d'environ dix

enregistrements sur une période de deux mois. Le corpus inclut un total de 80 enregistrements, avec 8 à 10 sessions par locuteur, à l'exception d'une locutrice qui n'a pu réaliser que 6 sessions en raison de difficultés techniques. Les locuteurs ont été invités à s'enregistrer régulièrement, avec un minimum de 24 heures entre les sessions et sur une période maximale de deux mois. En pratique, tous les locuteurs ont réalisé l'ensemble de leurs enregistrements dans un délai d'un mois et demi au maximum. Les locuteurs ont été encouragés à varier le moment de la journée pour les enregistrements, idéalement en alternant entre matin et après-midi, mais cette recommandation n'a pas pu être toujours suivie en raison de contraintes personnelles et familiales. Chaque session d'enregistrement a été complétée par une auto-évaluation sur une échelle de Likert à quatre points de l'état de fatigue générale du locuteur, de son état émotionnel, de sa fatigue vocale et du degré d'utilisation de la voix le jour de l'enregistrement.

Bien que les enregistrements aient été effectués au domicile des locuteurs, ce qui a induit un bruit ambiant variable, tous ont utilisé un modèle équivalent de carte de son et de microphone afin de limiter les facteurs de variation externes. Les locuteurs ont reçu pour instruction de s'enregistrer dans un environnement calme, toujours dans la même pièce si possible, et avec les mêmes réglages. Tout changement ou bruit ambiant inattendu devait être signalé dans un questionnaire complété après chaque enregistrement. Au début et à la fin de chaque session d'enregistrement, cinq secondes de silence étaient enregistrées afin de pouvoir contrôler un éventuel changement de bruit ambiant d'une session à l'autre ou en cours de session.

La plupart des locuteurs ont piloté leurs enregistrements à l'aide d'une version modifiée de l'application développée dans le cadre du protocole MonPaGe (Laganaro et al., 2021) pour la présentation des consignes et la gestion des enregistrements. Deux locutrices n'ont pas pu utiliser cette application et ont réalisé leurs enregistrements avec Praat, en s'appuyant sur un diaporama pour la présentation des consignes.

Les tâches de production étaient conçues pour être similaires à celles du corpus PATATRA, raccourci pour être compatible avec des enregistrements réalisés à domicile dans un délai réduit, certaines tâches étant adaptées pour répondre à des questionnements spécifiques qui n'étaient initialement pas ciblés lors de la mise en place du protocole du corpus PATATRA. Ainsi, le texte « La bise et le soleil » n'était lu qu'une fois, mais complété par deux autres textes courts conçus pour assurer une meilleure distribution des voyelles du français dans des contextes comparables. La liste de mots utilisée dans PATATRA était ici remplacée par une liste de 26 mots, chacune des 13 voyelles du français standard étant insérée dans deux contextes consonantiques antérieurs et postérieurs, afin de décrire plus largement la réalisation vocalique en incluant les voyelles, avec une variation des effets coarticulatoires. Une phrase entièrement voisée issue du protocole MonPaGe, permettant notamment de mesurer la fréquence fondamentale dans une tâche de production de parole, devait également être produite. Les locuteurs s'enregistrant seuls et à intervalles réduits, la tâche de production spontanée a également été adaptée, avec deux tâches distinctes. La première était de répondre à la question « Comment faites-vous une omelette ? » en donnant autant de détails que possible, les locuteurs étant invités à varier les recettes d'une session à l'autre. La seconde consistait en un récit de leurs activités depuis la session d'enregistrement précédente. Pour ces deux tâches de production spontanée, deux à trois minutes de parole devaient être produites.

De même que pour le corpus PATATRA, des tâches plus spécifiques étaient également incluses, dont le glissando à faible intensité et le temps de phonation maximale (avec uniquement deux essais) pour lesquels les consignes étaient les mêmes que celles de PATATRA. En complément, un /a/ tenu de deux à trois secondes, destiné aux mesures de qualité de voix, était également produit. Enfin, une tâche diadococinétique de performance maximale issue du protocole MonPaGe a été ajoutée, dans laquelle il était demandé au locuteur de produire le plus rapidement et le plus clairement possible une succession de syllabes. Cette tâche était déclinée en trois versions avec deux répétitions de la même syllabe (respectivement /ba/ et /go/) et une alternance de trois syllabes différentes avec la répétition de la séquence /badego/.

7.3.4 Comparaison entre parole spontanée et lecture en chinois mandarin

Publication associée :

[ACTI5] Sun, J., Wu, Y., **Audibert, N.**, & Adda-Decker, M. (2024). Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 291-300.

L'élaboration et le recueil de ce corpus s'inscrivent dans le cadre de la thèse de Jingyi Sun dirigée par Martine Adda-Decker depuis la fin de l'année 2022, qui vise à caractériser la réduction segmentale et tonale en chinois mandarin spontané. Yaru Wu (Université de Caen) et moi avons contribué à ce travail à travers des conseils méthodologiques, notamment concernant la conception et l'enregistrement d'un corpus conséquent de parole spontanée complété par une condition de contrôle qui a fait l'objet d'une première publication dans les actes de l'édition 2024 des Journées d'Études sur la Parole.

Le protocole retenu pour la partie conversationnelle de ce corpus est inspiré de celui utilisé pour le recueil du corpus francophone de parole conversationnelle NCCFr (Torreira et al., 2010), avec l'enregistrement des interactions verbales de binômes de locuteurs entretenant une relation amicale. En revanche pour des raisons éthiques, contrairement au corpus NCCFr pour lequel l'accord des locuteurs pour l'enregistrement de la partie conversationnel du corpus avait été recueilli a posteriori, dans ce corpus de chinois mandarin les locuteurs ont été pleinement informés des objectifs en amont de la réalisation des enregistrements. De plus, un post-traitement est réalisé afin d'éliminer les informations personnelles mentionnées par les locuteurs pendant l'interaction. Un ensemble de thématiques susceptibles de donner lieu à des interactions verbales telles que la nourriture ou l'hébergement sont proposées en début de session aux locuteurs afin d'amorcer l'interaction, l'expérimentatrice étant présente pour assurer le bon fonctionnement du matériel mais n'intervenant pas directement dans ces échanges. Cette session d'enregistrement de parole conversationnelle représente une durée de 40 à 70 minutes par binôme de locuteurs.

L'originalité du protocole retenu repose tout particulièrement sur la condition de production recueillie en complément de la parole conversationnelle, qui consiste en une relecture à l'issue de la session d'enregistrement d'un sous-ensemble leurs interactions spontanées. Cette relecture, destinée à permettre ensuite une comparaison directe entre parole conversationnelle et relecture sur le même matériel linguistique, est rendue possible par l'utilisation de l'outil multilingue de reconnaissance automatique de la parole Whisper (Radford et al., 2023) qui permet d'obtenir très peu de temps après la fin de la session d'enregistrement de parole conversationnelle une transcription automatique d'une qualité

acceptable. Des extraits sélectionnés de cette transcription sont alors révisés en interaction avec les locuteurs afin de corriger les erreurs résiduelles et de gommer certaines disfluences caractéristiques de la parole spontanée mais inadaptées dans une tâche de lecture. Enfin, chaque locuteur produit deux fois une lecture aussi neutre que possible des transcriptions sélectionnées et corrigées de ses propres productions, pour une tâche de lecture d'une durée de 10 à 15 minutes par locuteur.

La dernière étape de traitement, plus classique et réalisée a posteriori, consiste en une correction des transcriptions automatiques généralisée à l'ensemble des productions des locuteurs dans les deux conditions, avant d'utiliser ces transcriptions pour effectuer un alignement forcé à l'aide de Montreal Forced Aligner (McAuliffe et al., 2017) et de vérifier et corriger cet alignement.

Un total de 40 locuteurs sinophones natifs âgés de 20 à 32 ans, originaires de 19 provinces chinoises parmi les plus peuplées, ont été enregistrés en suivant ce protocole. L'ensemble de ces locuteurs étaient des étudiants s'exprimant sans accent régional perceptible en mandarin standard. Afin d'assurer la qualité des signaux acoustique et leur comparabilité, chacun des deux locuteurs en interaction était équipé d'un micro-casque, le même modèle étant utilisé pour les deux participants dans chacune des sessions d'enregistrement.

7.4 Développement d'outils en ligne

7.4.1 Motivations pour le développement de ce type d'outils

Au cours de ma carrière de chercheur, j'ai à de nombreuses reprises été amené à développer des outils dédiés au recueil, à l'analyse et/ou à la visualisation de données. Au-delà de scripts individuels ad-hoc développés pour mon propre usage et dont la réutilisation en l'état par d'autres personnes est difficilement envisageable, la majeure partie a consisté en des ensembles de scripts développés avec Matlab, Praat ou R, souvent adaptés à une tâche trop spécifique ou insuffisamment documentés pour faire l'objet d'une diffusion au-delà du cercle de mes collègues directement amenés à les utiliser dans le cadre de projets communs. Dans un certain nombre de cas, je me suis toutefois efforcé de concevoir ces scripts dans une optique centrée sur l'utilisateur, en faisant en sorte de les rendre aussi simples à prendre en main et paramétrables que possible. Cela m'a permis sans me placer dans une position de prestataire de services techniques de répondre régulièrement de façon positive aux sollicitations de collègues me demandant « Tu n'aurais pas un script qui fait … ? », les points de suspension pouvant correspondre à des tâches variées d'analyse acoustique ou statistique, d'annotation de données ou encore de visualisation. Récemment, j'ai entrepris de rassembler certains de ces scripts parmi ceux les plus génériques et réutilisables pour les rendre plus largement accessibles via ma page GitHub : https://github.com/nicolasaudibert.

Néanmoins, j'ai pu observer à la fois dans le cadre de mes activités d'enseignant amené à dispenser des cours de programmation appliquée à l'analyse acoustique de la parole ou encore de statistiques via l'utilisation de R, mais aussi dans mes interactions avec les masterants, doctorants et collègues que quelques soient les précautions prises pour les rendre aussi accessibles que possible, l'utilisation de tels outils nécessite malgré tout un certain niveau d'aisance technique, ne serait-ce que dans la gestion de l'organisation et de la structuration des fichiers. Ainsi et en dépit de la culture technique plus développée parmi les phonéticiens que dans de nombreux autres champs de la linguistique, il reste fréquent que des collègues pourtant extrêmement compétents dans leur domaine de spécialité et rompus par ailleurs à

l'analyse de données quantitatives se contentent d'utiliser les paramètres par défaut d'outils permettant une utilisation plus avancée, renoncent à l'utilisation d'outils en ligne de commande qui seraient pourtant les plus adaptés à l'analyse de leurs données, ou encore reprennent tels-quels des exemples de code dédié à l'analyse statistique ou des procédures d'analyse acoustique ou articulatoire faute de savoir les adapter à leurs besoins.

A mon sens, cette attitude répandue vis-à-vis des outils techniques peut également contribuer à une tendance d'évolution des sciences de la parole, déplorée par de nombreux collègues, dans laquelle la modélisation statistique de mesures quantitatives prend une telle importance qu'elle tend à se déconnecter de ce que ces mesures quantitatives sont supposées refléter dans les données. Ce point de vue peut sembler paradoxal alors que cette perte du lien entre la réalité des données et leur représentation constitue une critique de longue date de certaines pratiques en traitement automatique de la parole dans lesquelles les données de parole ne sont parfois considérées que comme des vecteurs de valeurs numériques utilisés comme point de départ pour l'entraînement de modèles de classification avancés. On peut mentionner également de nombreuses études en traitement automatique de la parole dans lesquelles les données de parole font l'objet d'extraction automatisée à grande échelle de mesures supposées être motivées phonétiquement avec des outils tels qu'OpenSMILE (Eyben et al., 2010) mais qui ne peuvent être interprétées que des contextes bien contrôlés.

Toutefois, j'estime que la phonétique et la phonologie de laboratoire sont également touchées par cette tendance d'évolution de notre domaine scientifique, avec le recours devenu incontournable à des modèles statistiques avancés qui parfois tendent à prendre le pas sur l'analyse fine des données et leur interprétation. Cette tendance s'est en effet accentuée dans les dix dernières années avec la généralisation de l'usage des modèles linéaires mixtes (voir par exemple Kirby & Sonderegger (2018) sur leur application en phonétique), et plus récemment la démocratisation de l'usage des modèles GAM (Wood, 2017) et des modèles de régression bayésienne (Vasishth et al., 2018). Si l'usage de modèles statistiques avancés permettant la prise en compte de facteurs aléatoire de variation a permis d'indéniables avancées pour l'analyse de données de parole complexes par nature, en dépit du développement de bibliothèques logicielles visant à simplifier leur usage, leur maîtrise nécessite une prise en main qui peut s'avérer chronophage, tout particulièrement pour les étudiants issus de formations en linguistique. Ainsi, j'ai pu observer que l'investissement en temps et en énergie nécessaire pour appréhender ces outils complexes s'accompagnait trop fréquemment d'une tendance regrettable à délaisser certains aspects plus qualitatifs de l'analyse phonétique, ainsi qu'à négliger l'inspection des données.

Ces observations ont contribué à me convaincre de la nécessité de mettre l'accent dans la formation des étudiants sur le lien entre mesures quantitatives et observations qualitatives issues des données de parole, typiquement l'analyse visuelle des informations spectrographiques pour les données acoustiques, ainsi que la maîtrise des statistiques descriptives et notamment de l'inspection des distributions comme préalable aux statistiques inférentielles. J'ai ainsi développé ces notions dans les cours à visée méthodologique que je dispense au niveau master, dans lesquels j'accueille régulièrement des doctorants dont les travaux nécessitent un complément de formation dans ces domaines. Ces réflexions m'ont aussi incité à contribuer à la mise à disposition des étudiants et collègues d'outils dont la prise en main soit aisée afin que les utilisateurs puissent se concentrer sur les questionnements scientifiques en lien avec l'interprétation des données plutôt que sur des aspects purement techniques.

De tels outils répondent à un réel besoin, et sont beaucoup plus susceptibles d'être adoptés par les utilisateurs issus de formations en linguistique que les méthodes d'analyse qui foisonnent dans la littérature mais dont la mise en œuvre nécessite a minima une aisance technique suffisante pour configurer et utiliser des outils en ligne de commande, et dans de nombreux cas des compétences en programmation. Parmi les premiers outils en ligne diffusés auprès de la communauté scientifique en sciences de la parole, on peut ainsi mentionner l'outil NORM dédié aux méthodes de normalisation des valeurs formantiques (Thomas et al., 2007), ou encore le calculateur en ligne Correlatore (Mairano & Romano, 2010) dédié aux métriques rythmiques. Plus récemment, le succès rencontré par l'outil Visible Vowels (Heeringa & Van de Velde, 2018) a bien illustré les besoins auxquels de tels outils dédiés à la visualisation des données permettent de répondre.

Au cours des dernières années, j'ai été amené à développer plusieurs outils en ligne dédiés à l'enseignement et/ou à la recherche. Parmi ces outils en ligne, je ne présente dans ce document de synthèse que ceux dédiés au traitement ou à la visualisation de données ayant donné lieu à des publications, spécifiquement axées sur la présentation de l'outil ou combinée à la présentation de résultats scientifiques, mes autres réalisations étant accessibles via la rubrique Ressources de ma page Web personnelle. Ces outils en ligne, développés à l'aide de R et du package Shiny (Chang et al., 2024) dédié au développement d'applications Web, sont actuellement hébergés par le serveur du Laboratoire de Phonétique et Phonologie qui dispose d'une capacité limitée de traitement de fichiers volumineux. Le code correspondant à ces applications est toutefois diffusé via GitHub sous licence libre GPL pour permettre son utilisation locale ou sur un autre serveur pour le traitement de jeux de données plus volumineux.

7.4.2 Exploration interactive de données : iHist et iScatter

Publication associée :

[ACTI2] **Audibert, N.** (2024). iHist et iScatter, outils en ligne d'exploration interactive de données : application aux valeurs aberrantes de f0 et de formants. *Actes des 35èmes Journées d'Études sur la Parole*, Toulouse, France, pp. 598-607.

7.4.2.1 Les défis posés par l'inspection et le « nettoyage » des données de parole

Qu'il s'agisse de mesures spécifiques aux données de parole ou plus généralement d'analyse de données quantitatives susceptibles d'inclure des erreurs de mesure ou tout autre mesure qui ne serait pas pertinente pour l'analyse, un processus de filtrage des valeurs erronées, qui peut être qualifié de processus de « nettoyage », est souvent nécessaire. Cela l'est d'autant plus qu'avec l'augmentation du volume de données traitées, en particulier dans les approches qui relèvent de la phonétique de corpus (M. Y. Liberman, 2019), le recours à des méthodes d'extraction au moins partiellement automatisées est souvent indispensable. On ne saurait trop répéter à quelle point la vérification des mesures obtenues automatiquement est cruciale, et cela est d'autant plus vrai pour les mesures acoustiques appliquées à la parole. En effet les méthodes d'analyse applicables peuvent être biaisées par un grand nombre de facteurs parmi lesquels on peut mentionner le bruit de fond, la superposition de tours de parole, l'imprécision de l'alignement automatique utilisé comme base pour les analyses, le dévoisement de segments supposés voisés, l'utilisation de modes de phonations autres que la

voix modale, ou encore l'application à des voix d'enfants de méthodes d'analyse évaluées uniquement sur des voix d'adultes.

Toutefois, dès lors que la taille des jeux de données analysés atteint un niveau critique, une vérification systématique de la validité de l'intégralité des mesures devient irréaliste, le processus de vérification ne pouvant être appliqué qu'à un sous-ensemble, sélectionné pour être représentatif du jeu de données complet lorsque les connaissances le permettent ou à défaut par échantillonnage aléatoire. En complément de cette vérification partielle, l'identification et éventuellement l'élimination des valeurs considérées comme déviante est alors nécessaire.

La méthode la plus courante consiste en la définition de seuils au-delà ou en deçà desquels les valeurs sont considérées comme erronées et donc éliminées. Lorsqu'il s'agit d'erreurs de mesure et autres biais identifiables, l'intérêt de l'élimination de ces mesures est évident, bien qu'il convienne de s'assurer que seule une proportion limitée de l'ensemble des données est concernée par ces erreurs, et que la distribution des erreurs reste homogène entre catégories destinées à faire l'objet de comparaisons. Ainsi sur les données acoustiques de parole, on peut mentionner le cas courant de l'analyse formantique pour laquelle les erreurs de détection ne sont généralement pas réparties de façon homogène entre catégories vocaliques (voir par exemple Gendrot & Adda-Decker (2005) sur le cas de la voyelle /u/ en français, beaucoup plus sujette à erreurs que les autres voyelles orales du français). Pour autant, hormis les cas d'erreurs de mesure grossières résultant en des valeurs aberrantes, la définition de seuils robustes permettant d'identifier ce qui constituerait ou non une erreur de mesure est particulièrement complexe, notamment en raison de la difficulté à définir une norme en matière de production de parole et les limites de ce qui constituerait la norme. Lorsque de telles normes existent pour des données de parole et incluent une caractérisation de la variabilité en complément des valeurs moyennes, une difficulté supplémentaire est qu'elles sont généralement établies dans des conditions particulières et que la possibilité de les transposer à d'autres conditions de production n'est jamais garantie. Ainsi, pour reprendre l'exemple des formants des voyelles orales du français, l'utilisation comme définition de la norme des valeurs formantiques moyennes et de leur variabilité, relevées sur des productions de voyelles isolées (voir par exemple Georgeton et al. (2012)), conduirait à considérer comme erronées de nombreuses mesures formantiques sur des voyelles produites en parole spontanée comme celles analysées par Meunier & Espesser (2011).

Une approche largement utilisée en psycholinguistique est de définir des seuils au-dessus ou sous lesquels les valeurs sont éliminées à partir de critères relatifs à la distribution des valeurs, le plus souvent en éliminant les valeurs situées à plus de deux écarts-types de la moyenne. En supposant que les données suivent une distribution normale, ce qui constitue le postulat à la base de cette approche, cela revient à conserver 95,4% des données situées autour de la moyenne en éliminant les valeurs les plus faibles et celles les plus élevées. Bien que cette méthode qui consiste en l'élimination systématique des valeurs qui dévient de la tendance majoritaire à partir de seuils fondés sur la distribution des valeurs reste couramment utilisée, le débat est vif parmi les psychologues entre les défenseurs de cette méthode qui considèrent qu'elle n'impacterait pas les résultats (André, 2022) tandis que d'autres concluent que cette pratique conduit à l'augmentation de l'erreur de type I, c'est-à-dire le fait de considérer à tort un effet comme significatif, et à biaiser les mesures d'intervalles de confiance (Karch, 2023). Par ailleurs des stratégies alternatives à l'élimination des valeurs extrêmes ont été proposées, comme par exemple leur remplacement par les valeurs extrêmes selon la

méthode dit du *winsorizing*. Ces deux stratégies peuvent toutefois être considérées comme équivalente du point de vue de leur impact sur les résultats d'après l'étude menée par Nicklin & Plonsky (2020) dans le cadre de tâches de lecture appliquées à l'évaluation de l'acquisition en langue seconde. Osborne & Overbay (2019) rappellent pour leur part qu'en dépit des avantages que peuvent conférer de telles stratégies d'élimination ou de remplacement des valeurs extrêmes pour la caractérisation statistique des effets étudiés, le fait de considérer comme aberrantes toutes les valeurs extrêmes peut mener à des biais importants, tout comme le fait de considérer comme valides toutes les valeurs plus proches de la tendance majoritaire. On peut par ailleurs souligner que même en s'inscrivant dans une approche centrée sur les données et en ne considérant pas le cas des erreurs de mesure, les valeurs extrêmes ne constituent qu'un type parmi d'autre d'anomalies susceptibles d'être présentes dans les données quantitatives (Foorthuis, 2021).

Au-delà de l'identification d'une tendance majoritaire dans un groupe de sujets ou dans une condition expérimentale, les valeurs extrêmes et plus généralement les valeurs pouvant être considérées comme statistiquement déviantes peuvent se révéler particulièrement informatives, notamment dans l'optique de l'étude de la variabilité inter et intra-individuelle qui me tient particulièrement à cœur. Outre leur usage pour l'enseignement des statistiques, la principale motivation pour le développement des deux applications en ligne dédiées à l'exploration interactive des données que je présente ci-dessous a été de fournir aux étudiants et collègues plus habitués à l'utilisation de tableurs des outils permettant de simplifier l'exploration de données quantitatives à partir de leur distribution, afin d'identifier des valeurs singulières (extrêmes ou non) et le cas échéant d'exporter des sous-ensembles dans des fichiers structurés en vue d'une inspection qualitative ultérieure plus détaillée.

Les deux applications présentées ci-dessous ont été conçues afin que leur interface utilisateur puisse être aisément traduite. Leur code source ainsi que les fichiers de localisation en français et en anglais sont distribués sous licence libre GPL via GitHub.

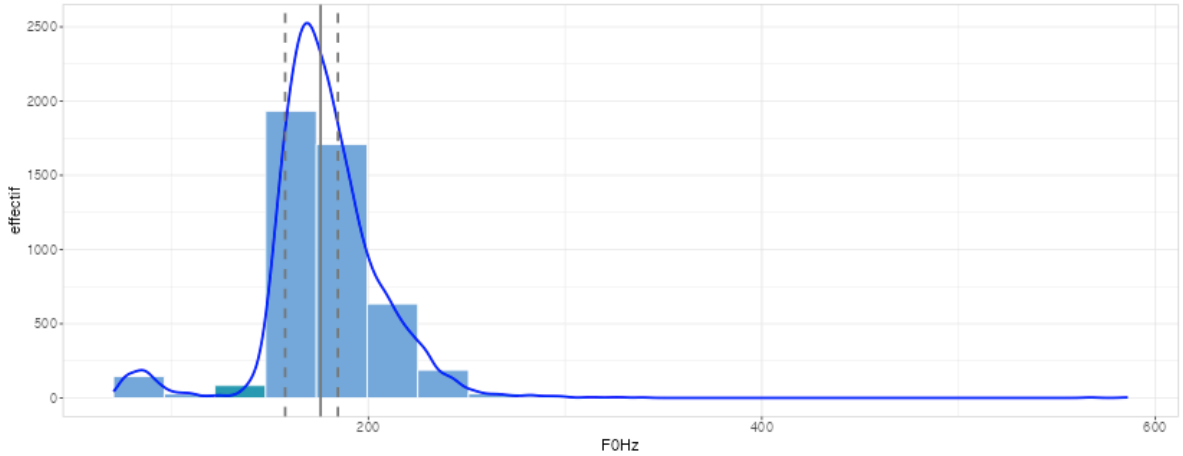
7.4.2.2 Exploration de distributions univariées : l'application iHist

Comme son nom l'indique, l'application iHist s'appuie sur la représentation sous la forme d'un histogramme de la distribution d'une variable numérique. L'affichage de l'histogramme est paramétrable à travers le choix du nombre de classes afin de proposer une sélection plus ou moins fine du sous-ensemble à explorer, et le choix des valeurs minimum et maximum affichées afin de permettre la comparaison directe entre jeux de données ne présentant pas la même étendue de valeurs. La sélection interactive d'un sous-ensemble de valeurs affiché dans un tableau et pouvant être exporté dans un format tabulaire se fait au moyen d'un clic sur la barre de l'histogramme correspondant à la classe à sélectionner.

Cette application permet à l'utilisateur d'importer un fichier de données au format tabulaire de son choix, typiquement au format Excel qui est celui le plus fréquemment utilisé par les étudiants et chercheurs en phonétique, et de prévisualiser le résultat afin d'ajuster si nécessaire l'encodage utilisé, ce qui peut s'avérer nécessaire lorsque les données incluent des symboles API ou autres caractères non-standard. La variable quantitative dont l'utilisateur souhaite visualiser la distribution peut ensuite être sélectionnée parmi les différentes variables numériques présentes dans le jeu de données, avec la possibilité d'ignorer d'éventuelles valeurs non-numériques comme par exemple les valeurs indéfinies attribuées par Praat dans un jeu de données issu d'une analyse acoustique. De façon optionnelle, les données affichées dans l'histogramme peuvent être filtrées en fonction des valeurs prises par d'autres variables

quantitatives ou catégorielles. La figure obtenue peut en outre être exportée avec une résolution élevée en vue de son inclusion dans une publication.

Initialement développée pour l'enseignement des statistiques afin de permettre aux étudiants de mieux appréhender le lien entre les données et la représentation de leur distribution, cette application inclut également le calcul et l'affichage de statistiques descriptives, et l'affichage optionnel de courbes et de lignes superposées. Outre la densité de distribution et l'affichage de la médiane et de quantiles illustrées par la Figure 78, ces courbes et lignes superposées peuvent inclure également à des fins de comparaison dans l'optique de l'application de tests d'hypothèse paramétriques le tracé de la distribution normale de même moyenne et de même écart-type, ainsi que l'affichage de la moyenne et de l'intervalle de confiance (paramétrique ou obtenu à partir des données par une méthode de *bootstrap*) de la moyenne avec une probabilité sélectionnée par l'utilisateur.



Exploration interactive des valeurs représentées dans l'histogramme

Cliquez sur une barre de l'histogramme pour afficher la plage de valeurs correspondantes ainsi que l'effectif de cette classe.

Classe n° 3/20 sélectionnée (122 - 147.8), 84 valeur(s) = 1.76% du total

☑ Affichage du détail des valeurs de la classe sélectionnée dans un tableau

typeProd	numSession	API	positionPt	F0Hz	lieu	aperture	nasalite	labialite	nomfichBase	idV
S	1.00	œ	50.00	125.04	anterieure	moyenne	orale	labiale	PTSVOX_LG011_F_session1_mic_S_1	v106575_PTSVOX_LG0
S	1.00	a	50.00	146.88	anterieure	ouverte	orale	non-labiale	PTSVOX_LG011_F_session1_mic_S_1	v106584_PTSVOX_LG0
S	1.00	i	50.00	145.60	anterieure	fermee	orale	non-labiale	PTSVOX_LG011_F_session1_mic_S_1	v106614_PTSVOX_LG0
S	1.00	a	50.00	143.99	anterieure	ouverte	orale	non-labiale	PTSVOX_LG011_F_session1_mic_S_1	v106712_PTSVOX_LG0
S	1.00	a	50.00	140.44	anterieure	ouverte	orale	non-labiale	PTSVOX_LG011_F_session1_mic_S_1	v106749_PTSVOX_LG0

Format de fichier

Exporter le sous-ensemble sélectionné | XLSX | Download

Figure 78 : Illustration de la partie principale de l'application *iHist* dédiée à l'affichage de l'histogramme interactif, et du tableau correspondant à la classe sélectionnée sur l'histogramme (en vert sur la figure). En complément de l'affichage de l'histogramme et de son paramétrage pour afficher 20 classes au lieu de 15 par défaut, deux éléments optionnels sont ici affichés : la densité de la distribution (courbe bleue), et celui de la médiane (ligne grise continue) et de quantiles sélectionnés (lignes grises pointillées, ici quantiles à 15% et 65%). Les données affichées consistent en des mesures de fréquence fondamentale obtenues par Praat avec une plage de détection de 60 Hz à 600 Hz, au milieu de chacune des voyelles produites par l'une des locutrices du corpus PTSVox (Chanclu et al., 2020). D'après Audibert (2024, [ACTI2]).

Parmi les usages possibles de cette application pour la recherche en phonétique, on peut mentionner la méthode d'affinage des valeurs limites pour la détection de la fréquence fondamentale présentée en section 7.2.2. La Figure 78 illustre l'affichage à l'aide de

l'application de la distribution des valeurs de fréquence fondamentale utilisées dans l'article de présentation de l'application (Audibert, 2024, [ACTI2]). Ces valeurs de fréquence fondamentale ont été obtenues par Praat à l'issue de la première passe de détection dans laquelle une plage large de détection allant de 60 Hz à 600 Hz est utilisée, au milieu de chacune des voyelles produites par l'une des locutrices du corpus PTSVox (Chanclu et al., 2020) en condition de lecture et de parole spontanée. Les données proches des valeurs limites candidates situées dans les « creux » de la distribution peuvent ainsi être identifiées et exportées pour donner lieu à une inspection qualitative des signaux correspondants, en l'occurrence à l'aide d'un script Praat.

La version française de cette application est disponible en ligne à l'URL suivant : https://shiny.laboratoirephonetiquephonologie.fr/iHist_fr/. La page principale de l'application inclut un lien de téléchargement des données de fréquence fondamentale mesurées automatiquement sur 4972 voyelles et utilisées dans l'article dans lequel l'application a été présentée (Audibert, 2024, [ACTI2]).

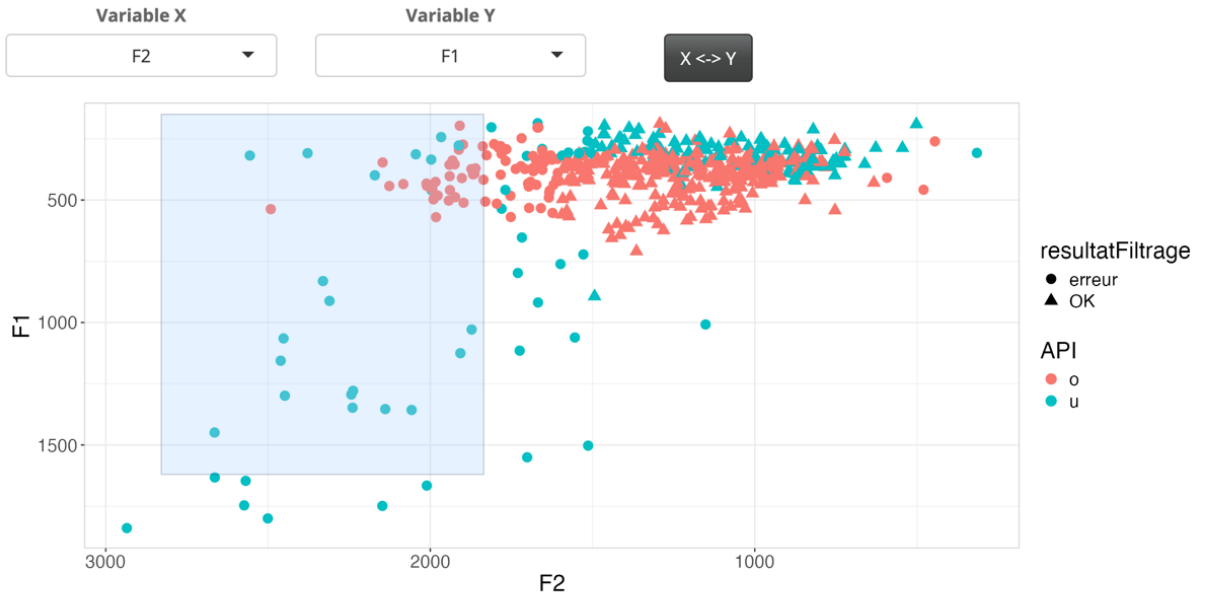
7.4.2.3 Exploration de distributions bivariées : l'application iScatter

L'application iScatter s'appuie sur la représentation sous la forme de nuage de points des liens entre deux variables numériques. La sélection interactive d'un sous-ensemble de valeurs affiché dans un tableau et pouvant être exporté dans un format tabulaire peut être effectuée soit par la sélection dans le nuage de points d'une zone rectangulaire incluant un ensemble de points, soit par un clic pour sélectionner le point le plus proche. De même que dans l'application iHist, les valeurs limites affichées sont paramétrables pour permettre une comparaison directe entre nuages de points, et les figures peuvent être exportées. De plus, certains paramètres graphiques sont ajustables, notamment la taille des points pour assurer la lisibilité de la figure dans le cas de jeux de données de taille conséquente. L'orientation des axes peut être inversée par rapport à l'orientation par défaut du nuage de points, typiquement pour afficher des valeurs des deux premiers formants en adoptant l'orientation habituelle pour la représentation des formants dans le plan F1*F2.

L'importation par l'utilisateur d'un fichier de données dans un format tabulaire repose sur le même principe que celui mis en œuvre dans l'application iHist. Après importation des données, les deux variables numériques peuvent être sélectionnées (avec la possibilité de permuter ensuite entre abscisses et ordonnées), avec également la possibilité d'ignorer d'éventuelles valeurs non-numériques et de filtrer les valeurs affichées en fonction des valeurs prises par d'autres variables quantitatives ou catégorielles. Pour répondre au cas courant dans lequel les données se subdivisent en sous-ensembles définis par les modalités d'une ou plusieurs variables dépendantes, une ou deux variables catégorielles peuvent être sélectionnées pour définir la couleur des points correspondant à chaque paire de valeurs, ainsi que leur forme.

En complément de l'affichage du nuage des points, des statistiques descriptives relatives à chacune des deux variables sélectionnées peuvent être affichées. De plus, les corrélations de Bravais-Pearson et de Spearman sont calculées pour l'ensemble des valeurs et pour chacune des modalités des variables indépendantes sélectionnées, et peuvent être exportées dans un tableau. De façon optionnelles, les droites de régression peuvent également être affichées pour chacune des modalités de la première variable indépendante.

La Figure 79 illustre une utilisation possible de l'application iScatter sur des valeurs de formants détectées automatiquement par Praat sur les mêmes voyelles que celles ayant fait l'objet d'une analyse automatique de fréquence fondamentale pour illustrer l'utilisation de iHist. Dans ce cas, seules les voyelles orales produites par la locutrice ont été conservées, pour un total de 4255 voyelles. Selon la méthode de Gendrot & Adda-Decker (2005), un crible a été appliqué afin d'étiqueter les mesures formantiques suspectes car trop éloignées des valeurs attendues connaissant la catégorie phonologique de la voyelle. La principale difficulté de cette approche dont les limites ont par exemple été soulignées par Lancien, Adda-Decker, et al. (2023) est de définir un crible à le fois suffisamment permissif pour ne pas considérer comme erronées des valeurs qui s'éloignent de celles considérées comme typiquement en raison de la variabilité de la parole, notamment dans le cas de la parole conversationnelle, tout en détectant les erreurs grossières telles que la non-détection du second formant. Dans le cas illustré par la Figure 79, un filtre a été appliqué pour n'afficher que les occurrences de /o/ et de /u/, et l'inspection des valeurs suspectes se concentre sur les occurrences de /u/ pour lesquelles la valeur de F2 détectée automatiquement est supérieure à 1500 Hz.



Affichage du détail des valeurs sélectionnées dans un tableau

loc	sexeLoc	typeProd	duree	API	ctxG_API	ctxD_API	F1	F2	F3	F4	resultatFiltrage
LG011	F	S	0.02	o	j	n	481.46	1977.24	3302.63	4323.47	erreur
LG011	F	S	0.02	u	t	ʒ	911.89	2310.52	3511.31	4428.10	erreur
LG011	F	S	0.04	u	t	ʃ	1348.10	2239.53	3840.36	4665.96	erreur
LG011	F	S	0.04	o	s	m	389.39	1864.96	2868.84	4674.40	erreur
LG011	F	S	0.06	o	n	k	392.51	1863.79	3104.94	4076.96	erreur

Figure 79 : Illustration de la partie principale de l'application *iScatter* dédiée à la sélection des variables à afficher en abscisses et en ordonnées, et à l'affichage du nuage de points interactif et du tableau correspondant au sous-ensemble sélectionné dans le nuage de points (rectangle bleu sur la figure). Dans cet exemple l'application est utilisée pour inspecter les valeurs étiquetées comme potentiellement erronées par l'application d'un crible dépendant de la voyelle inspiré de Gendrot & Adda-Decker (2005) dont le résultat est indiqué par la variable *resultatFiltrage*. Les deux variables indépendantes sont utilisées respectivement pour identifier les catégories vocaliques (couleurs) et le résultat de l'application du crible (formes des points). D'après Audibert (2024, [ACTI2]).

De même que dans l'exemple utilisé pour illustrer le fonctionnement de l'application iHist, le sous-ensemble de données sélectionné visuellement est exporté dans un format tabulaire pour servir de base à une inspection des signaux de parole correspondant afin d'identifier les exemplaires pouvant faire l'objet de corrections manuelles des valeurs de formants et éventuellement ceux devant être éliminés de l'analyse.

La version française de cette application est disponible en ligne à l'URL suivant : https://shiny.laboratoirephonetiquephonologie.fr/iScatter_fr/. De même que pour l'application iHist, la page principale de l'application inclut un lien de téléchargement des données formantiques utilisées dans l'article dans lequel l'application a été présentée (Audibert, 2024, [ACTI2]).

7.4.3 Calculateur de métriques relatives à l'espace vocalique

Publications associées :

[ACTI34] **Audibert, N.**, Fougeron, C., Gendrot, C., & Adda-Decker, M. (2015). Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS'15)*, Glasgow, Royaume-Uni, paper 0753 (actes en ligne).

[ACTI10] Hermes, A., **Audibert, N.**, & Bourbon, A. (2023). Age-related vowel variation in French. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic. pp. 2045-2049.

Ce calculateur en ligne, mentionné dans notre article sur les variations de l'espace vocalique liées à l'âge (Hermes et al. (2023, [ACTI10])) est en quelque sorte l'aboutissement d'un projet de longue haleine, débuté lors de discussions avec des collègues travaillant également sur la caractérisation de l'espace vocalique suite à ma présentation à Glasgow de l'article dans lequel nous avions introduit les métriques calculées à l'échelle de l'exemplaire présentées en section 7.2.3. Toutefois le travail de la stagiaire que j'avais encadré suite à ces échanges n'avait alors pas pu aboutir à une application en ligne fonctionnelle, et je m'étais donc contenté dans mes travaux suivants sur la caractérisation de l'espace vocalique de continuer à calculer ces métriques au moyen de scripts ad-hoc développés avec Matlab puis avec R.

Toutefois les développements récents de Shiny qui rend plus aisé le développement d'applications Web interactives et l'usage que j'ai été amené à en faire pour la mise en place des applications en ligne présentées dans les sections précédentes m'ont permis de mettre en place ce calculateur afin de rendre ces métriques accessibles aux collègues moins à l'aise avec l'utilisation de R.

Comme l'illustre la Figure 80, l'interface de ce calculateur permet d'importer un fichier de données dans un format tabulaire, puis de choisir une ou plusieurs variables de regroupement qui définissent les sous-ensembles sur lesquelles les métriques sont calculées (typiquement le locuteur, éventuellement complété par la condition expérimentale lorsque le jeu de données en inclut plusieurs), ainsi que la variable qui identifie les différentes catégories de voyelles.

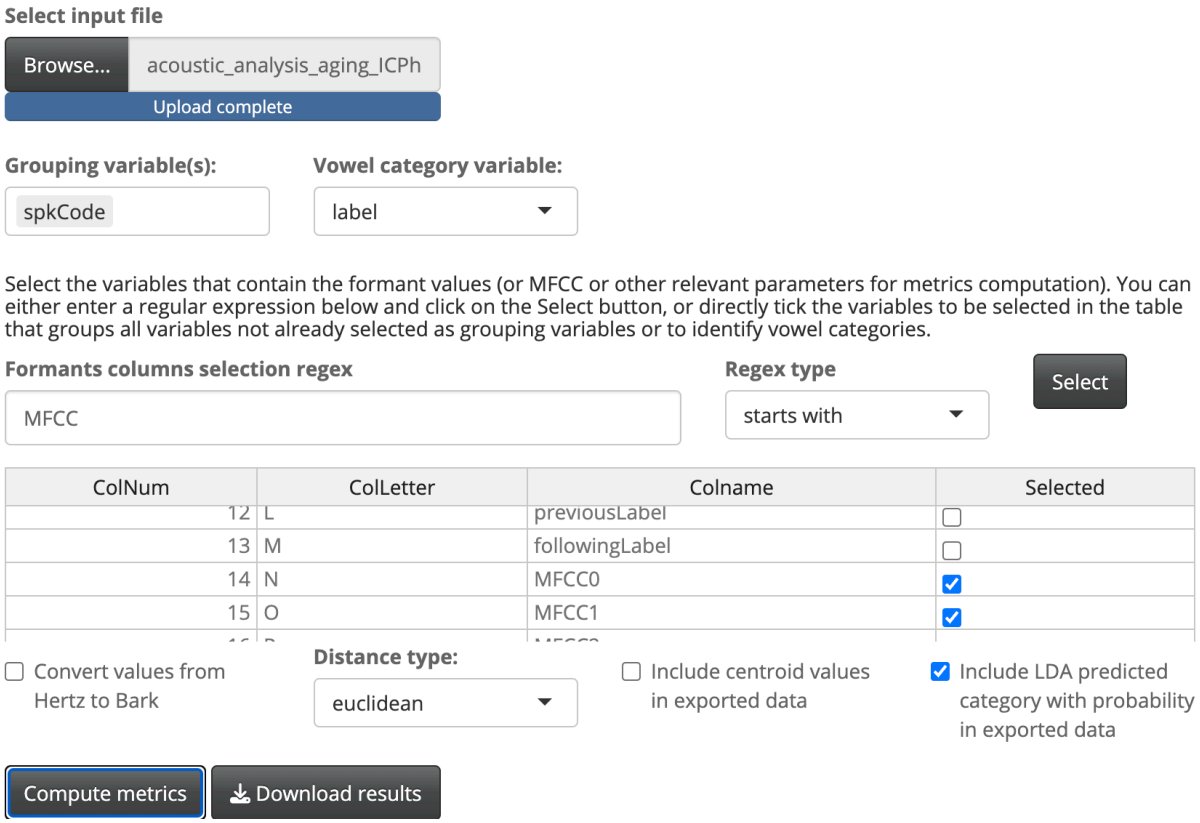


Figure 80 : Illustration de la partie principale du calculateur en ligne des métriques de caractérisation de l'espace vocalique DistCentroid, VDispersion et ContrastLoss définies dans Audibert et al. (2015, [ACTI34]), appliqué ici aux douze paramètres MFCC mesurés dans l'étude de Hermes et al. (2023, [ACTI10]).

L'utilisateur doit également sélectionner les variables qui correspondent aux valeurs formantiques, qui peuvent éventuellement être remplacés par des mesures alternatives pouvant se prêter à des calculs de distance et à une analyse discriminante linéaire, comme par exemple les douze coefficients MFCC utilisés dans Hermes et al. (2023, [ACTI10]). Cette sélection peut se faire directement dans la liste des variables présentes dans les données ou via une expression régulière, ce qui peut être utile dans le cas de paramétrisations via un nombre important de variables ou lorsque le jeu de données comporte un grand nombre de variables.

L'application permet également de façon optionnelle de convertir en Bark les valeurs de formants exprimées en Hertz, ce qui est bien entendu sans objet pour les valeurs de MFCC utilisées dans l'exemple illustré par la Figure 80, et/ou de sélectionner un autre type de distance que la distance euclidienne. La distance de Mahalanobis, qui pourrait s'avérer intéressante pour ces métriques appliquées à l'organisation du système vocalique (voir par exemple Lancien, Stuart-Smith et al. (2023)) sur son application au filtrage de valeurs formantiques), n'est pas encore intégrée à l'application mais pourra l'être prochainement.

Parmi les options disponibles, l'application permet de plus d'exporter les valeurs des centroïdes pour l'ensemble du système vocalique et par catégorie. Bien que ces métriques plus classiques ne soient pas directement intégrées dans l'application qui est dédiée à un niveau de granularité qui est celui de l'exemplaire, l'export des valeurs des centroïdes peut permettre de

compléter aisément l'analyse par des mesures plus globales telles que l'aire de l'espace vocalique à des fins de comparaison.

Enfin, en complément des valeurs de la métrique ContrastLoss qui correspond à la probabilité a posteriori d'appartenance à la catégorie vocalique de référence attribuée par l'analyse linéaire discriminante (LDA), l'application permet également d'exporter la catégorie vocalique prédite par la LDA et la probabilité associée. Cela peut par exemple permettre d'établir des matrices de confusion à partir des données acoustiques, ou simplement de comparer entre locuteurs, entre conditions ou entre catégories vocaliques la proportion d'exemplaires correctement catégorisés par la LDA.

Cette application en ligne étant a priori destinée à un public de chercheurs plus spécialisées que les applications de visualisation interactive de données, elle a été développée avec une interface utilisateur uniquement en anglais. Elle est accessible à l'URL suivant : https://shiny.laboratoirephonetiquephonologie.fr/vowel_space_metrics_computation/.

Bibliographie

Note : Afin de limiter les redondances, les entrées bibliographiques listées ci-dessous n'incluent pas mes propres publications, parmi lesquelles celles liées aux travaux présentés dans ce document de synthèse sont mentionnées au début des sections correspondantes, et qui sont par ailleurs listées en intégralité dans mon Curriculum Vitæ.

Abderrazek, S. (2023). *Évaluation de l'intelligibilité de la parole par apprentissage profond : Vers plus d'interprétabilité en phonétique clinique* [Thèse de doctorat]. Université d'Avignon.

Abry, C., & Boë, L.-J. (1980). A la recherche de corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français. In C. Abry, L.-J. Boë, P. Corsi, R. Descout, M. Gentil, & P. Graillot (Éds.), *Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiale.* (p. 217-237). Publications de l'Université des Langues et Lettres, Grenoble.

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *Actes des 26èmes Journées d'Études sur la Parole (JEP 2006)*, 389-400.

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., & Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, *39*(3), 261-270.

Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., & Kahn, J. (2016). Phonetic content impact on forensic voice comparison. *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, 210-217.

Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., & Kahn, J. (2017). Phonological content impact on wrongful convictions in Forensic Voice Comparison context. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2147-2151. https://doi.org/10.1109/ICASSP.2017.7952536

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*.

Al-Tamimi, J., & Khattab, G. (2015). Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *The Journal of the Acoustical Society of America*, *138*(1), 344-360. https://doi.org/10.1121/1.4922514

Alwan, M., & Paddle, P. M. (2022). Vocal cord paralysis: Pathophysiology, etiologies, and evaluation. *International Journal of Head and Neck Surgery*, *12*(4), 153-160.

Amerman, J. D., & Parnell, M. M. (1992). Speech timing strategies in elderly adults. *Journal of Phonetics*, *20*(1), 65-76. https://doi.org/10.1016/S0095-4470(19)30254-2

Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical science and technology*, *27*(4), 233-235.

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*(1), 213.

Anolli, L., Ciceri, R., & Infantino, M. G. (2002). From "blame by praise" to "praise by blame": Analysis of vocal patterns in ironic communication. *International Journal of Psychology*, *37*(5), 266-276. https://doi.org/10.1080/00207590244000106

Antolík, T. K., & Fougeron, C. (2013). Consonant distortions in dysarthria due to parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia. *Proceedings of Interspeech 2013*, 2152-2156. https://doi.org/10.21437/Interspeech.2013-509

Ardaillon, L., & Roebel, A. (2019). Fully-convolutional network for pitch estimation of speech signals. *Proceedings of Interspeech 2019*, 2005-2009. https://doi.org/10.21437/Interspeech.2019-2815

Aue, T., Cuny, C., Sander, D., & Grandjean, D. (2011). Peripheral responses to attended and unattended angry prosody: A dichotic listening paradigm. *Psychophysiology*, *48*(3), 385-392. https://doi.org/10.1111/j.1469-8986.2010.01064.x

Auzou, P., & Rolland-Monnoury, V. (2006). *BECD : batterie d'évaluation clinique de la dysarthrie*. Ortho édition.

Avanzi, M. (2023). Vot'artic'est formidab' : Une étude multifactorielle de la chute des liquides post-obstruantes finales de mot en français. *Journal of French Language Studies*, *33*(2), 137-167.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Bailles, F., Alazard-Guiu, C., & Prieto, P. (2022). Embodied prosodic training helps improve accentedness and suprasegmental accuracy. *Applied Linguistics*, *43*(4), 776-804.

Baken, R. J. (2005). The Aged Voice: A New Hypothesis. *Journal of Voice*, *19*(3), 317-325. https://doi.org/10.1016/j.jvoice.2004.07.005

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614-636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161.

Bänziger, T., & Scherer, K. R. (2007). Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. In A. C. R. Paiva, R. Prada, & R. W. Picard (Éds.), *Affective Computing and Intelligent Interaction* (p. 476-487). Springer Berlin Heidelberg.

Barkat-Defradas, M., Busseuil, C., Chauvy, O., Hirsch, F., Fauth, C., Révis, J., & Bretèque, B. A. de la. (2012). Dimension esthétique des voix normales et dysphoniques: Approches perceptive et acoustique. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, *28*.

Barrett, J., & Paus, T. (2002). Affect-induced changes in speech production. *Experimental Brain Research*, *146*(4), 531-537. https://doi.org/10.1007/s00221-002-1229-z

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. *Proceedings of the ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Béchet, M., Sandré, M., Hirsch, F., Richard, A., Marsac, F., & Sock, R. (2013). De l'utilisation de la pause silencieuse dans le débat politique télévisé. Le cas de François Hollande. *Mots. Les langages du politique*, *103*, 23-38.

Beckman, M., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Éds.), *Gesture, Segment, Prosody* (p. 68-89). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511519918.004

Ben Amor, I., & Bonastre, J.-F. (2022). BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison. *2022 International workshop on biometrics and forensics (IWBF)*, 1-6.

Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, *59*(1), 72-89.

Benjamin, B. J. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research*, *11*(2), 159-167. https://doi.org/10.1007/BF01068218

Bergeson, T. R., & Trehub, S. E. (2002). Absolute pitch and tempo in mothers' songs to infants. *Psychological science*, *13*(1), 72-75.

Bergounioux, G. (2016). How Statistics Entered Linguistics: Pierre Guiraud at Work. The Scientific Career of an Outsider. In J. Léon & S. Loiseau (Éds.), *History of Quantitative Linguistics in France* (p. 29-42). RAM-Verlag. https://hal.science/hal-01895613

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother–child speech. *Journal of child language*, *11*(3), 557-578.

Bigi, B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, *111*(ISSN: 0741-6164), 54-69.

Blatchford, H., & Foulkes, P. (2006). Idenfication of voices in shouting. *International Journal of Speech, Language and the Law*, *13*(2), 241-254. https://doi.org/10.1558/ijsll.2006.13.2.241

Blevins, J. (2006). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics*, *32*(2), 117-166. https://doi.org/doi:10.1515/TL.2006.009

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the institute of phonetic sciences*, *17*(1193), 97-110.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9), 341-345.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America*, *136*(2), EL116-EL121. https://doi.org/10.1121/1.4885482

Bonastre, J.-F. (2020). 1990-2020 : Retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire (1990-2020: A look back at 30 years of discussions on voice identification in the judicial system). In G. Adda, M. Amblard, & K. Fort (Éds.), *Actes de la Conférence conjointe JEP-TALN-RECITAL 2020* (p. 38-47). ATALA et AFCP. https://aclanthology.org/2020.jeptalnrecital-eternal.5

Bonastre, J.-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N., Fauve, B., & Mason, J. (2008). ALIZE/SpkDet: A state-of-the-art open source software for speaker recognition. *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2008)*, paper 20.

Boula de Mareüil, P. (2014). Qu'est-ce qu'un (phono)style ? *Cahiers de linguistique française*, *31*, 9-19.

Bourbon, A., & Hermes, A. (2020). Have a break: Aging effects on sentence production and structuring in French. *12th International Seminar on Speech Production*, 102-105.

Bradley, E. D. (2018). A comparison of the acoustic vowel spaces of speech and song. *Linguistic Research*, *35*(2).

Braun, A., & Cerrato, L. (1999). Estimating speaker age across languages. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, 1369-1372.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, *3*, 219-252.

Brugnara, F., Falavigna, D., & Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, *12*(4), 357-370.

Buck, R. (1999). The biological affects: A typology. *Psychological review*, *106*(2), 301.

Bürki, A. (2018). Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin & Review*, *25*(6), 1973-2004. https://doi.org/10.3758/s13423-017-1423-4

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335-359. https://doi.org/10.1007/s10579-008-9076-6

Calbi, M., Heimann, K., Barratt, D., Siri, F., Umiltà, M. A., & Gallese, V. (2017). How context influences our perception of emotional faces: A behavioral study on the Kuleshov effect. *Frontiers in psychology*, *8*, 1684.

Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, *124*(3), 1638-1652.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, *26*(2), 95-103.

Candea, M. (2000). Les euh et les allongements dits « d'hésitation » : Deux phénomènes soumis à certaines contraintes en français oral non lu. *Actes des XXIIIèmes Journées d'Etudes sur la Parole (JEP 2000)*, 73-76.

Carlson, R., Gustafson, K., & Strangert, E. (2006). Cues for hesitation in speech synthesis. *Proceedings of Interspeech 2006*, paper 1516. https://doi.org/10.21437/Interspeech.2006-382

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2023). On the speaker discriminatory power asymmetry regarding acoustic-phonetic parameters and the impact of speaking style. *Frontiers in psychology*, *14*, 1101187.

Caverlé, M. W. J., & Vogel, A. P. (2020). Stability, reliability, and sensitivity of acoustic measures of vowel space: A comparison of vowel space area, formant centralization ratio, and vowel articulation index. *The Journal of the Acoustical Society of America*, *148*(3), 1436-1444. https://doi.org/10.1121/10.0001931

Chanclu, A., Georgeton, L., Fredouille, C., & Bonastre, J.-F. (2020). PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire (PTSVOX: a Speech Database for Forensic Voice Comparison). *Actes de la Conférence conjointe JEP-TALN-RECITAL 2020*, 73-81.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Chevaillier, G., Sauvaget, E., Herman, P., & Tran Ba Huy, P. (2010). La cinématographie ultra rapide du larynx, ses apports en phoniatrie. *Revue de laryngologie, d'otologie et de rhinologie (1919)*, *131*(1), 23-29.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*(2), 141-176.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867-3878. https://doi.org/10.1121/1.1861893

Cho, T. (2016). Prosodic Boundary Strengthening in the Phonetics–Prosody Interface. *Language and Linguistics Compass*, *10*(3), 120-141. https://doi.org/10.1111/lnc3.12178

Cirelli, L. K., Jurewicz, Z. B., & Trehub, S. E. (2020). Effects of maternal singing style on mother–infant arousal and behavior. *Journal of cognitive neuroscience*, *32*(7), 1213-1220.

Conrad, N. J., Walsh, J., Allen, J. M., & Tsang, C. D. (2011). Examining infants' preferences for tempo in lullabies and playsongs. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *65*(3), 168.

Côté, M.-H. (2008). Empty elements in schwa, liaison and h-aspiré : The French Holy Trinity revisited. In J. Hartmann, V. Hegedűs, & H. C. van Riemsdijk (Éds.), *Sounds of Silence: Empty Elements in Syntax and Phonology* (p. 61-103). BRILL.

Crevier Buchman, L. (2001). Les dysphonies chroniques. *La lettre de l'Oto-Rhino-Laryngologie et de Chirurgie Cervico-Faciale*, *263*, 25-28.

Crevier Buchman, L., Tessier, C., Sauvignet, A., Arpin, S. B., & Monfrais-Pfauwadel, M.-C. (2005). Diagnostic d'une dysphonie non organique de l'adulte. *Revue de Laryngologie Otologie Rhinologie*, *126*, 353-360.

Crevier-Buchman, L., Laccourreye, O., Weinstein, G., Garcia, D., Jouffre, V., & Brasnu, D. (1995). Evolution of speech and voice following supracricoid partial laryngectomy. *The Journal of Laryngology & Otology*, *109*(5), 410-413. Cambridge Core. https://doi.org/10.1017/S0022215100130300

Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Éds.), *Ensemble Machine Learning: Methods and Applications* (p. 157-175). Springer New York. https://doi.org/10.1007/978-1-4419-9326-7_5

d'Alessandro, C., Rilliard, A., & Le Beux, S. (2011). Chironomic stylization of intonation. *The Journal of the Acoustical Society of America*, *129*(3), 1594-1604.

D'Alessandro, D. (2022). *Individual variations in anticipatory V-to-V coarticulation: Effects of Motor Speech Disorders, age, speech tempo changes and boundary type* [Thèse de doctorat]. Université Sorbonne Nouvelle.

D'Alessandro, D., Bourbon, A., & Fougeron, C. (2020). Effect of age on rate and coarticulation across different speech-tasks. *Proceedings of the 12th International Seminar on Speech Production*, 14-18.

D'Alessandro, D., & Fougeron, C. (2021). Changes in Anticipatory VtoV Coarticulation in French during Adulthood. *Languages*, *6*(4). https://doi.org/10.3390/languages6040181

Dalston, R. M., Warren, D. W., & Dalston, E. T. (1991). Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairment. *The Cleft Palate-Craniofacial Journal*, *28*(2), 184-189.

Daneš, F. (1994). Involvement with language and in language. *Journal of Pragmatics*, *22*(3), 251-264. https://doi.org/10.1016/0378-2166(94)90111-2

Dankovičová, J., & Nolan, F. (1999). Some acoustic effects of speaking style on utterances for automatic speaker verification. *Journal of the International Phonetic Association*, *29*(2), 115-128. Cambridge Core. https://doi.org/10.1017/S0025100300006496

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research*, *12*(2), 246-269. https://doi.org/10.1044/jshr.1202.246

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366. https://doi.org/10.1109/TASSP.1980.1163420

De Bodt, M. S., Hernández-Díaz Huici, M. E., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, *35*(3), 283-292. https://doi.org/10.1016/S0021-9924(02)00065-5

De Jong, G., McDougall, K., Hudson, T., & Nolan, F. (2007). The speaker discriminating power of sounds undergoing historical change: A formant-based study. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1813-1816.

De Jong, K. (1995). On the status of redundant features: The case of backing and rounding in American English. *Phonology and phonetic evidence: Papers in laboratory phonology IV*, 68-86.

De Looze, C. (2010). *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en Anglais* [Thèse de doctorat]. Université de Provence-Aix-Marseille I.

de Carvalho, J. B., Scheer, T., & Ségéral, P. (2008). *Lenition and Fortition*. Walter de Gruyter.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917-1930. https://doi.org/10.1121/1.1458024

de Cornulier, B. (1981). H-aspirée et la syllabation, Expressions disjonctives. In D. L. Goyvaerts (Éd.), *Phonology in the 1980's* (p. 183-230). John Benjamins.

de Gelder, B., Meeren, H. K. M., Righart, R., Stock, J. van den, van de Riet, W. A. C., & Tamietto, M. (2006). Chapter 3 Beyond the face: Exploring rapid influences of context on face processing. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso, & P. U. Tse (Éds.), *Progress in Brain Research* (Vol. 155, p. 37-48). Elsevier. https://doi.org/10.1016/S0079-6123(06)55003-4

Dell, F. (1973). *Les règles et les sons : Introduction à la phonologie générative*. HERMANN.

Dell, F. (1995). Consonant clusters and phonological syllables in French. *Lingua*, *95*(1-3), 5-26.

Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for deltaC. In P. Karnowski & I. Szigeti (Éds.), *Language and language-processing* (p. 231-241). Peter Lang. https://doi.org/10.5167/uzh-111789

Descout, R., Boë, L.-J., & Abry, C. (1980). Labialité vocalique et labialité consonantique. In C. Abry, L.-J. Boë, P. Corsi, R. Descout, M. Gentil, & P. Graillot (Éds.), *Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiale.* (p. 111-126). Publications de l'Université des Langues et Lettres, Grenoble.

de Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., & Bourlard, H. (2004). Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, *116*(3), 1781-1792. https://doi.org/10.1121/1.1781620

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of phonetics*, *24*(2), 187-208.

Dubeda, T. (2006). Prosodic boundaries in Czech: An experiment based on delexicalized speech. *Proceedings of the Ninth International Conference on Spoken Language Processing*.

Duez, D. (1991). *La pause dans la parole de l'homme politique*. CNRS.

Duez, D. (2001). Signification des hésitations dans la parole spontanée. *Revue parole*, *17*, 113-138.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, *3*, 1393-1396.

Eckert, P. (2017). Age as a Sociolinguistic Variable. In F. Coulmas (Éd.), *The Handbook of Sociolinguistics* (p. 151-167). https://doi.org/10.1002/9781405166256.ch9

El-Banna, M., & Youssef, G. (2015). Early Voice Therapy in Patients with Unilateral Vocal Fold Paralysis. *Folia Phoniatrica et Logopaedica*, *66*(6), 237-243. https://doi.org/10.1159/000369167

Enos, F., & Hirschberg, J. (2006). A framework for eliciting emotional speech: Capitalizing on the actors process. *Proceedings of the First international workshop on emotion: Corpora for research on emotion and affect (Satellite of the international conference on language resources and evaluation (LREC 2006))*, 6-10.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(3), 253-260. https://doi.org/10.1016/S0095-4470(11)00055-6

Esling, J. H., & Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL quarterly*, *17*(1), 89-95.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462.

Fagyal, Z., Hassa, S., & Ngom, F. (2002). L'opposition [e]-[E] en syllabes ouvertes de fin de mot en français parisien: Étude acoustique préliminaire. *Actes des 24èmes Journées d'Etudes sur la Parole (JEP 2002)*, 165-168.

Fagyal, Z., Nguyen, N., & de Mareüil, P. B. (2003). From dilation to coarticulation: Is there vowel harmony in French? *Studies in the Linguistic Sciences*, *32*(2).

Falk, S. (2011). Melodic versus intonational coding of communicative functions: A comparison of tonal contours in infant-directed song and speech. *Psychomusicology: Music, Mind and Brain*, *21*(1-2), 54.

Falk, S., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80-86.

Fang, C., Li, H., Ma, L., & Zhang, M. (2017). Intelligibility Evaluation of Pathological Speech through Multigranularity Feature Extraction and Optimization. *Computational and Mathematical Methods in Medicine*, *2017*(1), 2431573. https://doi.org/10.1155/2017/2431573

Fant, G. (1971). *Acoustic Theory of Speech Production*. De Gruyter Mouton. https://doi.org/10.1515/9783110873429

Fant, G. (1973). *Speech sounds and features.* The MIT Press.

Fant, G. (1975). Non-uniform vowel normalization. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *2*, 1-19.

Ferguson, S., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1241-1255.

Ferragne, E., & Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*, *38*(4), 526-539.

Ferré, G. (2005). Gesture, Intonation and the Pragmatic Structure of Narratives in British English Conversation. *York Papers in Linguistics*, *Series 2*(Issue 3), 55-90.

Feugère, L., d'Alessandro, C., Doval, B., & Perrotin, O. (2017). Cantor Digitalis: Chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, *2017*, 1-19.

Finitzo-Hieber, & Tillman, T. W. (1978). Room Acoustics Effects on Monosyllabic Word Discrimination Ability for Normal and Hearing-Impaired Children. *Journal of Speech and Hearing Research*, *21*(3), 440-458. https://doi.org/10.1044/jshr.2103.440

Fletcher, S. G. (1970). Theory and instrumentation for quantitative measurement for nasality. *The Cleft palate journal*, *7*(2), 601-609.

Flipsen, P., & Lee, S. (2012). Reference data for the American English acoustic vowel space. *Clinical linguistics & phonetics*, *26*(11-12), 926-933.

Fónagy, I. (1983). *La vive voix*. Payot.

Fónagy, I., & Bérard, E. (1972). «Il est huit heures» : Contribution à l'analyse sémantique de la vive voix. *Phonetica*, *26*(3), 157-192. https://doi.org/doi:10.1159/000259408

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050-1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Foorthuis, R. (2021). On the nature and types of anomalies: A review of deviations in data. *International journal of data science and analytics*, *12*(4), 297-331.

Fougeron, C., Delvaux, V., Menard, L., & Laganaro, M. (2018). The MonPaGe_HA database for the documentation of spoken French throughout adulthood. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Fougeron, C., Guitard-Ivent, F., & Delvaux, V. (2021). Multi-Dimensional Variation in Adult Speech as a Function of Age. *Languages*, *6*(4). https://doi.org/10.3390/languages6040176

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728-3740. https://doi.org/10.1121/1.418332

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Modelling Sociophonetic Variation*, *34*(4), 409-438. https://doi.org/10.1016/j.wocn.2005.08.002

Foulkes, P., Docherty, G., Khattab, G., & Yaeger-Dror, M. (2010). Chapter 14. Sound Judgments: Perception of Indexical Features in Children's Speech. In D. R. Preston & N. Niedzielski (Éds.), *A Reader in Sociophonetics* (p. 327-356). De Gruyter Mouton. https://doi.org/doi:10.1515/9781934078068.3.327

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical Society of America*, *90*(4), 1816-1827. https://doi.org/10.1121/1.401662

Fowler, C. A., & Brancazio, L. (2000). Coarticulation Resistance of American English Consonants and its Effects on Transconsonantal Vowel-to-Vowel Coarticulation. *Language and Speech*, *43*(1), 1-41. https://doi.org/10.1177/00238309000430010101

Frid, J., & Ambrazaitis, G. (2010). Automatic estimation of pitch range through distribution fitting. *XXIIIth Swedish Phonetics Conference, Fonetik 2010, Lund, June 2-4, 2010.*, 41-46.

Fromkin, V. (1964). Lip positions in American English vowels. *Language and speech*, *7*(4), 215-225.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*(1), 29-47. https://doi.org/10.1016/j.wocn.2012.08.007

Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Zygis (Éds.), *An Interdisciplinary Guide* (p. 281-302). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110226584.281

Fujimura, O. (1962). Analysis of nasal consonants. *The Journal of the Acoustical Society of America*, *34*(12), 1865-1875.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., & Choukri, K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 139-142.

Galliano, S., Gravier, G., & Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. *Proceedings of Interspeech 2009*, 2583-2586. https://doi.org/10.21437/Interspeech.2009-680

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*, *28*(2), 580-597. https://doi.org/10.1016/j.csl.2013.07.005

Garrido-Vásquez, P., Pell, M. D., Paulmann, S., & Kotz, S. A. (2018). Dynamic facial expressions prime the processing of emotional prosody. *Frontiers in human neuroscience*, *12*, 244.

Gauvain, J.-L., Lamel, L., & Adda, G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, *37*(1), 89-108. https://doi.org/10.1016/S0167-6393(01)00061-9

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, *63*(1), 223-230. https://doi.org/10.1121/1.381717

Gelfer, M. P., Andrews, M. L., & Schmidt, C. P. (1991). Effects of prolonged loud reading on selected measures of vocal function in trained and untrained singers. *Journal of Voice*, *5*(2), 158-167. https://doi.org/10.1016/S0892-1997(05)80179-1

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511790942

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. *Proceedings of Interspeech 2005*, 2453-2456. https://doi.org/10.21437/Interspeech.2005-753

Georgeton, L. (2014). *Renforcement des voyelles orales du français en position initiale de constituants prosodiques : Interaction avec les contrastes phonologiques.* [Thèse de doctorat]. Université Sorbonne Nouvelle - Paris 3.

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics*, *44*, 83-95.

Georgeton, L., Paillereau, N., Landron, S., Gao, J., & Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : À la recherche d'une référence pour les apprenants de FLE. *Actes de la Conférence conjointe JEP-TALN-RECITAL 2012*, 145-152.

Ghio, A., Lalain, M., Rebourg, M., Marczyk, A., Fredouille, C., & Woisard, V. (2021). Validation of an Intelligibility Test Based on Acoustic-Phonetic Decoding of Pseudo-Words: Overall Results from Patients with Cancer of the Oral Cavity and the Oropharynx. *Folia Phoniatrica et Logopaedica*, *74*(3), 209-222. https://doi.org/10.1159/000519427

Gili Fivela, B., Stella, A., D'Apolito, S., & Sigona, F. (2011). Coarticulation across prosodic domains in Italian: An ultrasound investigation. *Proceedings of Interspeech 2011*, 393-396. https://doi.org/10.21437/Interspeech.2011-158

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1), 189-212. https://doi.org/10.1016/S0167-6393(02)00082-1

Göhring, T. (2017). L'état Actuel Du H Disjonctif (H Aspiré) Une approche fondée sur la fréquence d'emploi. *Romanische Forschungen*, *129*(2), 147-168.

Goldman, J.-P., Prsir, T., & Auchlin, A. (2014). C-PhonoGenre: A 7-hours corpus of 7 speaking styles in French: Relations between situational features and prosodic properties. *Proceedings of LREC 2014*, 302-305.

Goy, H., Kathleen Pichora-Fuller, M., & van Lieshout, P. (2016). Effects of age on speech and voice quality ratingsa). *The Journal of the Acoustical Society of America*, *139*(4), 1648-1659. https://doi.org/10.1121/1.4945094

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Éds.), *Papers in Laboratory Phonology 7* (p. 515-546). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.515

Graillot, P., Boë, L.-J., Gentil, M., & Abry, C. (1980). Analyse des correspondances de paramètres descriptifs du jeu des lèvres en français. In C. Abry, L.-J. Boë, P. Corsi, R. Descout, M. Gentil, & P. Graillot (Éds.), *Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiale.* (p. 127-146). Publications de l'Université des Langues et Lettres, Grenoble.

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. *Proceedings of the Eighth international conference on Language Resources and Evaluation (LREC 2012)*. https://hal.science/hal-00712591

Greenberg, C. S., Martin, A. F., Brandschain, L., Campbell, J. P., Cieri, C., Doddington, G. R., & Godfrey, J. J. (2010). Human Assisted Speaker Recognition In NIST SRE10. *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2010)*, paper 32.

Grimaldi, M., & Cummins, F. (2009). Speech style and speaker recognition: A case study. *Proceedings of Interspeech 2009*, 920-923. https://doi.org/10.21437/Interspeech.2009-276

Hallé, P., & Adda-Decker, M. (2007). Voicing assimilation in journalistic speech. *Proceedings of the 16th International Congress of Phonetic Sciences*, *2007*, 493-496.

Harmegnies, B., & Poch-Olivé, D. (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech communication*, *11*(4-5), 429-437.

Harmegnies, B., & Poch-Olivé, D. (1994). Formants frequencies variability in French vowels under the effect of various speaking styles. *Le Journal de Physique IV*, *4*(C5), C5-509.

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Rothman, H., & Hollien, H. (2008). Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age. *Journal of Voice*, *22*(1), 58-69. https://doi.org/10.1016/j.jvoice.2006.07.004

Heald, S. L. M., & Nusbaum, H. C. (2015). Variability in Vowel Production within and between Days. *PLOS ONE*, *10*(9), e0136791. https://doi.org/10.1371/journal.pone.0136791

Heeringa, W., & Van de Velde, H. (2018). Visible Vowels: A tool for the visualization of vowel variation. *Proceedings of the CLARIN annual conference*, 8-10.

Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., Rubin, A., Eller, R., Chandran, S., Abaza, M., Lyons, K., Divi, V., Lott, J., Johnson, J., & Hillenbrand, J. (2014). Quantifying the Cepstral Peak Prominence, a Measure of Dysphonia. *Journal of Voice*, *28*(6), 783-788. https://doi.org/10.1016/j.jvoice.2014.05.005

Henrich, N., Sundin, G., Ambroise, D., d'Alessandro, C., Castellengo, M., & Doval, B. (2003). Just noticeable differences of open quotient and asymmetry coefficient in singing voice. *Journal of Voice*, *17*(4), 481-494. https://doi.org/10.1067/S0892-1997(03)00005-5

Herment, S., & Tortel, A. (2021). The intonation contour of non-finality revisited. In *English pronunciation instruction: Research-based insights* (Kirkova-Naskova, A.; Henderson, A.; Fouz-González, J., p. 176-195). John Benjamins Publishing.

Hermes, A., Mertens, J., & Mücke, D. (2018). Age-related Effects on Sensorimotor Control of Speech Production. *Proceedings of Interspeech 2018*, 1526-1530. https://doi.org/10.21437/Interspeech.2018-1233

Hirano, M. (1981). *Clinical examination of voice*. Springer-Verlag.

Hoge, H., Tropf, H. S., Winski, R., van den Heuvel, H., Haeb-Umbach, R., & Choukri, K. (1997). European speech databases for telephone applications. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *3*, 1771-1774 vol.3. https://doi.org/10.1109/ICASSP.1997.598873

Hoit, J. D., & Hixon, T. J. (1987). Age and Speech Breathing. *Journal of Speech, Language, and Hearing Research*, *30*(3), 351-366. https://doi.org/10.1044/jshr.3003.351

Hollien, H., Majewski, W., & Hollien, P. A. (1974). Perceptual identification of voices under normal, stress, and disguised speaking conditions. *The Journal of the Acoustical Society of America*, *56*(S1), S53-S53. https://doi.org/10.1121/1.1914230

Honda, K., & Maeda, S. (2008). Glottal-opening and airflow pattern during production of voiceless fricatives: A new non-invasive instrumentation. *The Journal of the Acoustical Society of America*, *123*(5_Supplement), 3738-3738.

Hoole, P. (2006). *Experimental studies of laryngeal articulation* [Thèse d'habilitation]. University of Munich.

Horii, Y. (1980). An accelerometric approach to nasality measurement: A preliminary report. *The Cleft Palate Journal*, *17*(3), 254-261.

Horii, Y. (1983). An accelerometric measure as a physical correlate of perceived hypernasality in speech. *Journal of Speech, Language, and Hearing Research*, *26*(3), 476-480.

Huet, K., & Harmegnies, B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *Actes des XXIIIèmes Journées d'Études sur la Parole (JEP 2000)*, *1*, 225-228.

Hummel, R., Chan, W.-Y., & Falk, T. H. (2011). Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech. *Proceedings of Interspeech 2011*, 3017-3020. https://doi.org/10.21437/Interspeech.2011-755

Hunter, E. J., Ferguson, S. H., & Newman, C. A. (2016). Listener estimations of talker age: A meta-analysis of the literature. *Logopedics Phoniatrics Vocology*, *41*(3), 101-105. https://doi.org/10.3109/14015439.2015.1009160

Hunter, L., Pring, T., & Martin, S. (1991). The use of strategies to increase speech intelligibility in cerebral palsy: An experimental evaluation. *International Journal of Language & Communication Disorders*, *26*(2), 163-174. https://doi.org/10.3109/13682829109012001

Huntley, R., Hollien, H., & Shipp, T. (1987). Influences of listener characteristics on perceived age estimations. *Journal of Voice*, *1*(1), 49-52. https://doi.org/10.1016/S0892-1997(87)80024-3

Illner, V., Sovka, P., & Rusz, J. (2020). Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. *Biomedical Signal Processing and Control*, *58*, 101831.

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019). Pathological Speech Intelligibility Assessment Based on the Short-time Objective Intelligibility Measure. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6405-6409. https://doi.org/10.1109/ICASSP.2019.8683741

Jessen, S., & Kotz, S. A. (2013). On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, *7*, 369.

Jesus, L. M. T., Martinez, J., Hall, A., & Ferreira, A. (2015). Acoustic Correlates of Compensatory Adjustments to the Glottic and Supraglottic Structures in Patients with Unilateral Vocal Fold Paralysis. *BioMed Research International*, *2015*(1), 704121. https://doi.org/10.1155/2015/704121

Johnson, K. (2002). *Acoustic and Auditory Phonetics 2e* (2nd edition). John Wiley &Sons.

Johnson, K. (2004). Massive reduction in conversational American English. *Proceedings of the 1st session of the 10th international symposium on Spontaneous speech: Data and analysis*, 29-54.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252-1263. https://doi.org/10.1121/1.1288413

Jouvet, D., & Laprie, Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. *Proceedings of the 25th European signal processing conference (eusipco)*, 1614-1618.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, *129*(5), 770.

Kahn, J. (2011). *Parole de locuteur : Performance et confiance en identification biométrique vocale* (Numéro 2011AVIG0187) [Thèse de doctorat, Université d'Avignon]. https://theses.hal.science/tel-00995071

Kain, A., & van Santen, J. P. H. (2010). Frequency-domain delexicalization using surrogate vowels. *Proceedings of Interspeech 2010*, 474-477. https://doi.org/10.21437/Interspeech.2010-201

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of child language*, *45*(5), 1035-1053.

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: Smiling, teaching or social convergence? *Royal Society open science*, *4*(8), 170306.

Karch, J. D. (2023). Outliers may not be automatically removed. *Journal of Experimental Psychology: General*, *152*(6), 1735.

Kavanagh, C. (2012). *New consonantal acoustic parameters for forensic speaker comparison* [Thèse de doctorat]. University of York.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, *27*(6), 349-353.

Keating, P., Kreiman, J., & Alwan, A. (2019). « A new speech database for within-and between-speaker variability », Paper in Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.). *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*.

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of Visual Information on the Intelligibility of Dysarthric Speech. *American Journal of Speech-Language Pathology*, *16*(3), 222-234. https://doi.org/10.1044/1058-0360(2007/027)

Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics*. Palgrave Macmillan.

Kent, R. D., Netsell, R., & Abbs, J. H. (1979). Acoustic Characteristics of Dysarthria Associated with Cerebellar Disease. *Journal of Speech, Language, and Hearing Research*, *22*(3), 627-648. https://doi.org/10.1044/jshr.2203.627

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders*, *74*, 74-97.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward Phonetic Intelligibility Testing in Dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482-499. https://doi.org/10.1044/jshd.5404.482

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of communication disorders*, *32*(3), 141-186.

Kewley-Port, D., & Zheng, Y. (1999). Vowel formant discrimination: Towards more ordinary listening conditions. *The Journal of the Acoustical Society of America*, *106*(5), 2945-2958. https://doi.org/10.1121/1.428134

Kharlamov, V., Brenner, D., & Tucker, B. V. (2023). Examining the effect of high-frequency information on the classification of conversationally produced English fricativesa). *The Journal of the Acoustical Society of America*, *154*(3), 1896-1902. https://doi.org/10.1121/10.0021067

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70-85.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326-347.

Kitzing, P., & Åkerlund, L. (2009). Long-Time Average Spectrograms of Dysphonic Voices before and after Therapy. *Folia Phoniatrica et Logopaedica*, *45*(2), 53-61. https://doi.org/10.1159/000266213

Klein, E., Brunner, J., & Hoole, P. (2019). The relevance of auditory feedback for consonant production: The case of fricatives. *Journal of Phonetics*, *77*, 100931. https://doi.org/10.1016/j.wocn.2019.100931

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward Improved Spectral Measures of /s/: Results From Adolescents. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1175-1189. https://doi.org/10.1044/1092-4388(2012/12-0038)

Kouklia, C. (2019). *Dominance, hostilité et expressivité vocale dans le débat politique: étude perceptive et acoustique du conseil municipal de Montreuil (93100)* [Thèse de doctorat]. Université Sorbonne Nouvelle - Paris 3.

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (1st edition). Wiley-Blackwell.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155-177. https://doi.org/10.3758/s13423-017-1272-1

Kubinec, R. (2023). Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds. *Political Analysis*, *31*(4), 519-536. Cambridge Core. https://doi.org/10.1017/pan.2022.20

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684-686.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1-13. https://doi.org/10.18637/jss.v036.i11

Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, *22*(1), 43-65. https://doi.org/10.1016/S0167-6393(97)00012-5

Laccourreye, O., Biacabe, B., Crevier-Buchmann, L., Laccourreye, H., Weinstein, G., & Brasnu, D. (1995). Duration and Frequency Characteristics of Speech and Voice following Supracricoid Partial Laryngectomy. *Annals of Otology, Rhinology & Laryngology*, *104*(7), 516-521. https://doi.org/10.1177/000348949510400703

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fournet, M., Catalano Chiuvé, S., Lopez, U., Trouville, R., Ménard, L., & others. (2021). Sensitivity and specificity of an acoustic-and perceptual-based tool for assessing motor speech disorders in French: The MonPaGe-screening protocol. *Clinical Linguistics & Phonetics*, *35*(11), 1060-1075.

Lamel, L. F., Gauvain, J.-L., & Eskénazi, M. (1991). BREF, a Large Vocabulary Spoken Corpus for French. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 505-508. http://www-tlp.limsi.fr/public/e91_0505.pdf

Lancien, M., Adda-Decker, M., & Stuart-Smith, J. (2023). Knowledge-Driven vs Data-Driven Methods for Filtering Acoustic Measures in Phonetics Corpora. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 3166-3170.

Lancien, M., & Côté, M.-H. (2018). Phonostyle et réduction vocalique en français laurentien. *SHS Web Conf.*, *46*. https://doi.org/10.1051/shsconf/20184609003

Lancien, M., Stuart-Smith, J., & Adda-Decker, M. (2023). Using Mahalanobis Distance to Filter Erroneous Vowel Features in Less-Resourced Languages: Application to Quebec French. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 3181-3185.

Landick, M. (1995). The Mid-Vowels in Figures: Hard Facts. *The French Review*, *69*(1), 88-102. JSTOR.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Laver, J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, *31*, 1-186.

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, *14*, 139-155.

Léon, P.-R. (1999). *Précis de phonostylistique: Parole et expressivité*. Nathan Université.

Lévêque, N., Laganaro, M., Fougeron, C., Delvaux, V., Pernon, M., Borel, S., & Catalano, S. (2016). MonPaGe : Un protocole informatisé d'évaluation de la parole pathologique en langue française. *Revue Neurologique*, *172*, A162-A163. https://doi.org/10.1016/j.neurol.2016.01.386

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, *5*(1), 91-107.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (p. 403-439). Springer.

Linville, S. E., & Korabic, E. W. (1986). Elderly listeners' estimates of vocal age in adult females. *The Journal of the Acoustical Society of America*, *80*(2), 692-694. https://doi.org/10.1121/1.394013

Linville, S. E., & Rens, J. (2001). Vocal Tract Resonance Analysis of Aging Voice Using Long-Term Average Spectra. *Journal of Voice*, *15*(3), 323-330. https://doi.org/10.1016/S0892-1997(01)00034-0

Lotto, A. J., Holt, L. L., & Kluender, K. R. (1997). *Effect of Voice Quality on Perceived Height of English Vowels*. *54*(2), 76-93. https://doi.org/10.1159/000262212

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska Directed Emotional Faces. *PsycTESTS Dataset*, *91*, 630.

MacCallum, J. K., Olszewski, A. E., Zhang, Y., & Jiang, J. J. (2011). Effects of low-pass filtering on acoustic analysis of voice. *Journal of Voice*, *25*(1), 15-20.

Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, *15*(1), 19-44. https://doi.org/10.1080/00437956.1959.11659682

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*(4), 381-400.

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., & Schuster, M. (2009). Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*(1), 926951. https://doi.org/10.1155/2010/926951

Mairano, P., & Romano, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. *La dimensione temporale del parlato*, 79-100.

Maisonneuve, M., Fredouille, C., Lalain, M., Ghio, A., & Woisard, V. (2024). Towards objective and interpretable speech disorder assessment: A comparative analysis of CNN and transformer-based models. *Proceedings of Interspeech 2024*, 1970-1974. https://doi.org/10.21437/Interspeech.2024-267

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962-3973. https://doi.org/10.1121/1.2990715

Martin, A. F., & Greenberg, C. S. (2009). NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels. *Proceedings of Interspeech 2009*, 2579-2582. https://doi.org/10.21437/Interspeech.2009-679

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52-59.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological science*, *26*(3), 341-347.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of Interspeech 2017*, *2017*, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, *13*(1), 89-126. https://doi.org/10.1558/ijsll.v13i1.89

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of/uː/in British English. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1825-1828.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362-378.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, *99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, *111*(1), 172.

Mennen, I. (2015). Beyond Segments: Towards a L2 Intonation Learning Theory. In E. Delais-Roussarie, M. Avanzi, & S. Herment (Éds.), *Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations* (p. 171-188). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45168-7_9

Meunier, C., & Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, *39*(3), 271-278.

Mixdorff, H., & Pfitzinger, H. R. (2005). Analysing fundamental frequency contours and local speech rate in map task dialogs. *Quantitative Prosody Modelling for Natural Speech Description and Generation*, *46*(3), 310-325. https://doi.org/10.1016/j.specom.2005.02.019

Moerman, M., Martens, J., Crevier-Buchman, L., Woisard, V., & Dejonckere, P. (2005). Perceptive evaluation of substitution voices: The I(I)NFVo rating scale. *Revue de Laryngologie-Otologie-Rhinologie*, *126*(5), 323-325.

Moon, S., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, *96*(1), 40-55. https://doi.org/10.1121/1.410492

Morlec, Y., Bailly, G., & Aubergé, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, *33*(4), 357-371.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453-467.

Mullennix, J., Barber, J., & Cory, T. (2019). An examination of the Kuleshov effect using still photographs. *PLOS ONE*, *14*(10), 1-13. https://doi.org/10.1371/journal.pone.0224623

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Myers, S. (2012). Final devoicing: Production and perception studies. *Prosody matters: Essays in honor of Elisabeth Selkirk*, 148-180.

Nadeu, M. (2014). Stress-and speech rate-induced vowel quality variation in Catalan and Spanish. *Journal of Phonetics*, *46*, 1-22.

Nagao, K. (2006). *Cross-language study of age perception* [Thèse de doctorat]. Indiana University.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, *22*(2), 171-184.

Narayan, C. R., & McDermott, L. C. (2016). Speech rate and pitch characteristics of infant-directed speech: Longitudinal and cross-linguistic observations. *The Journal of the Acoustical Society of America*, *139*(3), 1272-1281. https://doi.org/10.1121/1.4944634

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, *7*(3), 171.

Nefti, S. (2004). *Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance* [Thèse de doctorat]. Université Rennes 1.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, *28*(4), 661-677.

Nguyen, N., & Fagyal, Z. (2008). Acoustic aspects of vowel harmony in French. *Journal of Phonetics*, *36*(1), 1-27.

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 26-55.

Niebuhr, O., Brem, A., Michalsky, J., & Neitsch, J. (2020). What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and Mark Zuckerberg. *Cadernos de Linguistica e Teoria da Literatura*, *1*(1).

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. *Proceedings of the conference Law and Language: Prospect and Retrospect*, 12-15.

Obin, N., Lanchantin, P., Lacheret, A., & Rodet, X. (2011). Reformulating prosodic break model into segmental HMMs and information fusion. *Proceedings of Interspeech 2011*, 1829-1832. https://doi.org/10.21437/Interspeech.2011-40

Ohala, J. J. (1983a). Cross-Language Use of Pitch: An Ethological View. *Phonetica*, *40*(1), 1-18. https://doi.org/doi:10.1159/000261678

Ohala, J. J. (1983b). The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (Éd.), *The Production of Speech* (p. 189-216). Springer Science & Business Media.

Ohala, J. J. (1984). An Ethological Perspective on Common Cross-Language Utilization of F_o of Voice. *Phonetica*, *41*(1), 1-16. https://doi.org/10.1159/000261706

Ohala, J. J. (1997). Aerodynamics of phonology. *Proceedings of the Seoul International Conference on Linguistics*, *92*, 97.

Oliveira, G. M. G. F., Mrs., de Melo, D. C., Mrs., Serra, L. S. M., Dra., Granjeiro, R. C., Dr., & Sampaio, A. L. L., Dr. (2024). Dysphonia Interference in Schoolteachers' Speech Intelligibility in the Classroom. *Journal of Voice*, *38*(2), 316-324. https://doi.org/10.1016/j.jvoice.2021.09.004

Osborne, J. W., & Overbay, A. (2019). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, *49*(3), 197.

Pagel, V., Carbonell, N., & Laprie, Y. (1996). A new method for speech delexicalization, and its application to the perception of French prosody. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, *2*, 821-824.

Paja, M. S., & Falk, T. H. (2012). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. *Proceedings of Interspeech 2012*, 62-65. https://doi.org/10.21437/Interspeech.2012-26

Palva, S., Palva, J. M., Shtyrov, Y., Kujala, T., Ilmoniemi, R. J., Kaila, K., & Näätänen, R. (2002). Distinct gamma-band evoked responses to speech and non-speech sounds in humans. *The Journal of Neuroscience*, *22*(4), RC211.

Pätzold, M., & Simpson, A. P. (1997). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung Universität Kiel*, *32*, 215-247.

Pennington, M. C. (1999). Computer-aided pronunciation pedagogy: Promise, limitations, directions. *Computer assisted language learning*, *12*(5), 427-440.

Pereira, C. (2000). Dimensions of emotional meaning in speech. *Proceedings of the First ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Pettirossi, A. (2021). *La dysphonie chez les professeures des écoles: Perception et représentations* [Thèse de doctorat]. Université Sorbonne Nouvelle.

Picheny, M., Durlach, N., & Braida, L. (1986). Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, *29*(4), 434-446.

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.

Pompino-Marschall, B., Steriopolo, E., & Żygis, M. (2017). Ukrainian. *Journal of the International Phonetic Association*, *47*(3), 349-357.

Pon-Barry, H., & Shieber, S. M. (2011). Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, *2011*, 1-11.

Povey, D., Ghoshal, A., & Boulianne, G. (2011). The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning*, 28492-28518.

Ramig, L. A., & Ringel, R. L. (1983). Effects of Physiological Aging on Selected Acoustic Characteristics of Voice. *Journal of Speech, Language, and Hearing Research*, *26*(1), 22-30. https://doi.org/10.1044/jshr.2601.22

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, *105*(1), 512-521.

Rastatter, M. P., McGuire, R. A., Kalinowski, J., & Stuart, A. (2009). Formant Frequency Characteristics of Elderly Speakers in Contextual Speech. *Folia Phoniatrica et Logopaedica*, *49*(1), 1-8. https://doi.org/10.1159/000266431

Rathod, S., Charola, M., Vora, A., Jogi, Y., & Patil, H. A. (2023). Whisper Features for Dysarthric Severity-Level Classification. *Proceedings of Interspeech 2023*, 1523-1527. https://doi.org/10.21437/Interspeech.2023-1891

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *The Journal of the Acoustical Society of America*, *76*(6), 1624-1635. https://doi.org/10.1121/1.391609

Redenbaugh, M. A., & Reich, A. R. (1985). Correspondence between an accelerometric nasal/voice amplitude ratio and listeners' direct magnitude estimations of hypernasality. *Journal of Speech, Language, and Hearing Research*, *28*(2), 273-281.

Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, Production, Acoustics, and Perception, Second Edition*. John Wiley & Sons.

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America*, *66*(4), 1023-1028. https://doi.org/10.1121/1.383321

Ridouane, R. (2003). *Suites de consonnes en berbère : Phonétique et phonologie* [Thèse de doctorat]. Université de la Sorbonne Nouvelle-Paris III.

Robert, V., Bonneau, A., Wrobel-Dautcourt, B., & Laprie, Y. (2007). Prédiction phonétique de la coarticulation labiale. *Perturbations et réajustements: langue et langage*, 155-167.

Rouas, J.-L., Beppu, M., & Adda-Decker, M. (2010). Comparison of Spectral Properties of Read, Prepared and Casual Speech in French. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Roy, N. (2003). Functional dysphonia. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *11*(3), 144-148.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

Ryalls, J., Cliche, A., Fortier-Blanc, J., Coulombe, A., & Prud'hommeaux, A. (1997). Voice-onset time in younger and older French-speaking Canadians. *Clinical Linguistics & Phonetics*, *11*(3), 205-212. https://doi.org/10.3109/02699209708985191

Sander, D., & Scherer, K. (2009). *Traité de psychologie des émotions*. Dunod.

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, *134*(5), EL477-EL483. https://doi.org/10.1121/1.4826150

Sapir, S., Ramig, L. O., Spielman, J., & Fox, C. (2011). Acoustic metrics of vowel articulation in Parkinson's disease: Vowel space area (VSA) vs. Vowel articulation index (VAI). *Proceedings of the First Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2011)*, 173-175.

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research*, *53*(1), 114-125.

Sataloff, R. T., & Kost, K. M. (2020a). The Effects of Age on the Voice, Part 1. *Journal of Singing*, *77*(1), 63+. Gale Academic OneFile.

Sataloff, R. T., & Kost, K. M. (2020b). The Effects of Age on the Voice, Part 2. *Journal of Singing*, *77*(2), 205+. Gale Academic OneFile.

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*, *134*(5), 3793-3807. https://doi.org/10.1121/1.4824120

Scheer, T. (2024). Glottal stop insertion and production planning domains in French. *The Linguistic Review*, *41*(2), 339-379. https://doi.org/doi:10.1515/tlr-2024-2011

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227-256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695-729. https://doi.org/10.1177/0539018405058216

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123-148. https://doi.org/10.1007/BF00995674

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proceedings of the XIVth International Congress of Phonetic Sciences*.

Schiller, I. S., Morsomme, D., Kob, M., & Remacle, A. (2020). Noise and a Speaker's Impaired Voice Quality Disrupt Spoken Language Processing in School-Aged Children: Evidence From Performance and Response Time Measures. *Journal of Speech, Language, and Hearing Research*, *63*(7), 2115-2131. https://doi.org/10.1044/2020_JSLHR-19-00348

Schiller, N. O., & Köster, O. (1998). The ability of expert witnesses to identify voices: A comparison between trained and untrained listeners. *Forensic Linguistics*, *5*, 1-9.

Schindler, A., Favero, E., Nudo, S., Spadola-Bisetti, M., Ottaviani, F., & Schindler, O. (2005). Voice after supracricoid laryngectomy: Subjective, objective and self-assessment data. *Logopedics Phoniatrics Vocology*, *30*(3-4), 114-119. https://doi.org/10.1080/14015430500256592

Schuster, M., Noth, E., Haderlein, T., Steidl, S., Batliner, A., & Rosanowski, F. (2005). Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, *1*, I/61-I/64 Vol. 1. https://doi.org/10.1109/ICASSP.2005.1415050

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, *25*(3), 255-286.

Seidler, R. D., Alberts, J. L., & Stelmach, G. E. (2002). Changes in Multi-Joint Performance with Age. *Motor Control*, *6*(1), 19-31. https://doi.org/10.1123/mcj.6.1.19

Shadle, C. H., Chen, W.-R., Koenig, L. L., & Preston, J. L. (2023). Refining and extending measures for fricative spectra, with special attention to the high-frequency rangea). *The Journal of the Acoustical Society of America*, *154*(3), 1932-1944. https://doi.org/10.1121/10.0021075

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6907-6911.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., & others. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779-4783.

Shield, B., & Dockrell, J. E. (2004). External and internal noise surveys of London primary schools. *The Journal of the Acoustical Society of America*, *115*(2), 730-738. https://doi.org/10.1121/1.1635837

Shochi, T., Aubergé, V., & Rilliard, A. (2006). How prosodic attitudes can be false friends: Japanese vs. French social affects. *Speech Prosody 2006*, paper 249. https://doi.org/10.21437/SpeechProsody.2006-156

Signorello, R., Demolin, D., Bernardoni, N. H., Gerratt, B. R., Zhang, Z., & Kreiman, J. (2020). Vocal fundamental frequency and sound pressure level in charismatic speech: A cross-gender and-language study. *Journal of Voice*, *34*(5), 808-e1.

Simon, A. C., Auchlin, A., Goldman, J.-P., & Christodoulides, G. (2013). Tendances prosodiques de la parole radiophonique. *Cahiers de praxématique*, *61*.

Snoeren, N. D., Hallé, P. A., & Segui, J. (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, *34*(2), 241-268. https://doi.org/10.1016/j.wocn.2005.06.001

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Sonntag, G. P., & Portele, T. (1998). PURR—a method for prosody evaluation and investigation. *Computer Speech & Language*, *12*(4), 437-451.

Stevens, K. N., & House, A. S. (1955). Development of a Quantitative Description of Vowel Articulation. *The Journal of the Acoustical Society of America*, *27*(3), 484-493. https://doi.org/10.1121/1.1907943

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24-31. https://doi.org/10.1177/0098628316677643

Stuart-Smith, J., Foulkes, P., & Docherty, G. (1999). Glasgow: Accent and voice quality. *Urban voices: Accent studies in the British Isles*, 203-222.

Stuart-Smith, J., Timmins, C., & Tweedie, F. (2007). 'Talkin' Jockney'? Variation and change in Glaswegian accent. *Journal of Sociolinguistics*, *11*(2), 221-260. https://doi.org/10.1111/j.1467-9841.2007.00319.x

Sundberg, J. (1975). Formant technique in a professional female singer. *Acta Acustica united with Acustica*, *32*(2), 89-96.

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. https://doi.org/10.1121/1.401923

Swets, J. A. (1964). *Signal Detection and Recognition in Human Observers: Contemporary Readings*. John Wiley and Sons.

Tabain, M., & Perrier, P. (2005). Articulation and acoustics of/i/in preboundary position in French. *Journal of Phonetics*, *33*(1), 77-100.

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. *Journal of Memory and Language*, *24*(3), 320-335. https://doi.org/10.1016/0749-596X(85)90031-2

Temple, R. A. M. (1999). Phonetic and sociophonetic conditioning of voicing patterns in the stop consonants of French. *Proceedings of the 14th International Congress of Phonetic Sciences*, 1409-1412.

Temple, R. A. M. (2000). Old wine into new wineskins. A variationist investigation into patterns of voicing in plosives in the Atlas Linguistique de la France. *Transactions of the Philological Society*, *98*(2), 353-394. https://doi.org/10.1111/1467-968X.00068

Thomas, E. R., Kendall, T., Yeager-Dror, M., & Kretzschmar, W. (2007). Two things sociolinguists should know: Software packages for vowel normalization, and accessing linguistic atlas data. *Workshop at New Ways of Analyzing Variation (NWAV)*, *36*.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326-344.

Tibshirani, R., Walther, G., & Hastie, T. (2002). Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Tirronen, S., Javanmardi, F., Kodali, M., Kadiri, S. R., & Alku, P. (2023). Utilizing wav2vec in database-independent voice disorder detection. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5.

Titze, I. R. (1999). Toward occupational safety criteria for vocalization. *Logopedics Phoniatrics Vocology*, *24*(2), 49-54. https://doi.org/10.1080/140154399435110

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, *52*(3), 201-212.

Trainor, L. J., Clark, E. D., Huntley, A., & Adams, B. A. (1997). The acoustic basis of preferences for infant-directed singing. *Infant Behavior and Development*, *20*(3), 383-396.

Tranel, B. (1987). *The Sounds of French: An Introduction*. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511620645

Tranel, B., & Del Gobbo, F. (2002). Local conjunction in Italian and French phonology. In J. Camps & C. R. Wiltshire (Éds.), *Romance Phonology and Variation* (p. 191-218). John Benjamins.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97-100. https://doi.org/10.1121/1.399849

Trehub, S. E., & Gudmundsdottir, H. R. (2015). *Mothers as singing mentors for infants*. online publication. Oxford University Press.

Tsang, C. D., & Conrad, N. J. (2010). Does the message matter? The effect of song type on infants' pitch preferences for lullabies and playsongs. *Infant Behavior and Development*, *33*(1), 96-100.

Tsang, C. D., Falk, S., & Hessel, A. (2017). Infants prefer infant-directed song over speech. *Child development*, *88*(4), 1207-1215.

Vainio, M., Suni, A., Raitio, T., Nurminen, J., Järvikivi, J., & Alku, P. (2009). New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. *Proceedings of Interspeech 2009*, 1703-1706. https://doi.org/10.21437/Interspeech.2009-514

Vaissière, J. (2008). On acoustic salience of vowels and consonants predicted from articulatory models. *Keynote paper, 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers*.

Vaissière, J., Honda, K., Amelot, A., Maeda, S., & Crevier-Buchman, L. (2010). Multisensor Platform for Speech Physiology Research in a Phonetics Laboratory (< Feature Article> Methodology for Speech Physiology Research). *Journal of the Phonetic Society of Japan*, *14*(2), 65-77.

Van Dommelen, W. A. (1987). The Contribution of Speech Rhythm and Pitch to Speaker Recognition. *Language and Speech*, *30*(4), 325-338. https://doi.org/10.1177/002383098703000403

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19-38. https://doi.org/10.1016/S0095-4470(19)30723-5

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, *12*(1), 1-23. https://doi.org/10.1016/0167-6393(93)90015-D

Vasilescu, I., Adda-Decker, M., & Lamel, L. (2012). Cross-lingual studies of ASR errors: Paradigms for perceptual evaluations. *Proceedings of LREC 2012*, 3511-3518.

Vasilescu, I., Candea, M., & Adda-Decker, M. (2004). Hésitations autonomes dans 8 langues : Une étude acoustique et perceptive. *Actes du Workshop MIDL04*.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147-161.

Vaysse, R., Astésano, C., & Farinas, J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *The Journal of the Acoustical Society of America*, *152*(5), 3091-3101.

Walter, H. (1977). *La Phonologie du français*. Presses universitaires de France. Paris; BNF381856103.

Warner, N. (2012). Methods for studying spontaneous speech. In *A. Cohn, C. Fougeron, & M. K. Huffman (Eds.), The Oxford handbook of laboratory phonology* (p. 621-633).

Weirich, M., & Simpson, A. (2013). Investigating the relationship between average speaker fundamental frequency and acoustic vowel space size. *The Journal of the Acoustical Society of America*, *134*(4), 2965-2974. https://doi.org/10.1121/1.4818891

Weirich, M., Simpson, A. P., Öjbro, J., & Ericsdotter Nordgren, C. (2019). The phonetics of gender in Swedish and German. *Proceedings of Fonetik 2019*, 49-53.

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*, *53*(1), 1-18.

Westbury, J. R., & Keating, P. A. (1986). On the naturalness of stop consonant voicing. *Journal of Linguistics*, *22*(1), 145-166. Cambridge Core. https://doi.org/10.1017/S0022226700010598

Whalen, D., & Chen, W.-R. (2019). Variability and central tendencies in speech production. *Frontiers in Communication*, *4*, 49.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501-522. https://doi.org/10.1016/j.wocn.2007.02.003

Whiteside, S. P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *The Journal of the Acoustical Society of America*, *110*(1), 464-478. https://doi.org/10.1121/1.1379087

Woisard, V., Bodin, S., & Puech, M. (2004). The Voice Handicap Index: Impact of the translation in French on the validation. *Revue de laryngologie-otologie-rhinologie*, *125*(5), 307-312.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.

Wright, R. (2006). Intra-speaker variation and units in human speech perception and ASR. *Proceedings of the Speech Recognition and Intrinsic Variation Workshop*.

Wright, R., Local, J., Ogden, R., & Temple, R. (2004). Factors of lexical competition in vowel articulation. *Papers in laboratory phonology VI*, 75-87.

Yarmey, A. D. (2001). Earwitness descriptions and speaker identification. *International Journal of Speech, Language and the Law*, *8*(1), 113-122.

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608-54649. https://doi.org/10.1109/ACCESS.2024.3389497

Yeni–Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*(4), 966-975. https://doi.org/10.1121/1.387031

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., Narayanan, S., & Busso, C. (2004). An acoustic study of emotions expressed in speech. *Proceedings of Interspeech 2004*, 2193-2196. https://doi.org/10.21437/Interspeech.2004-242

Yoo, K., & Nolan, F. (2020). Sampling the progression of domain-initial denasalization in Seoul Korean. *Laboratory Phonology*, *11*(1).

Yoon, D. (2024). *Effets de la morphologie corporelle et de la langue parlée sur la parole d'hommes et de femmes* [Thèse de doctorat]. Université Sorbonne Nouvelle.

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders*, *11*(6), 499-512. https://doi.org/10.1016/0021-9924(78)90024-2

Yoshimura, T., Fujimoto, T., Oura, K., & Tokuda, K. (2023). SPTK4: An open-source software toolkit for speech signal processing. *12th Speech Synthesis Workshop (SSW) 2023*.

Yuan, C., Gonzalez-Fuente, S., Baills, F., & Prieto, P. (2019). Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition*, *41*(1), 5-32.

Zellou, G., Kim, L., & Gendrot, C. (2024). Comparing human and machine's use of coarticulatory vowel nasalization for linguistic classification. *The Journal of the Acoustical Society of America*, *156*(1), 489-502. https://doi.org/10.1121/10.0027932

Zhang, Y., Chen, L., Chen, J., Mi, X., & Zhu, S. (2022). Redefining Womanhood in Generation Z: An Analysis of Gender Representation and Awareness in K-pop Culture. *Proceedings of the 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022)*, 2868-2877. https://doi.org/10.2991/978-2-494069-89-3_329

Zuraw, K., & Hayes, B. (2017). Intersecting constraint families: An argument for Harmonic Grammar. *Language*, *93*(3), 497-548. JSTOR.

Table des illustrations