La parole : du traitement automatique à la mesure de l'intelligibilité

Jérôme Farinas

MÉMOIRE

En vue de l'obtention de l'

HABILITATION Á DIRIGER LES RECHERCHES

Délivré par : *l'Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *21/12/2023* **par :**

Jérôme FARINAS

La parole : du traitement automatique à la mesure de l'intelligibilité

JURY

Rapporteurs :

Christine MEUNIER	Université Aix Marseille Laboratoire Parole et Langage
Martine ADDA-DECKER	Université Sorbonne Nouvelle Laboratoire de Phonétique et Phonologie
Hervé GLOTIN	Université de Toulon Laboratoire d'Informatique et Systèmes

Examinateurs :

Corinne FREDOUILLE	Avignon Université Laboratoire Informatique d'Avignon
Virginie WOISARD-BASSOLS	Centre Hospitalier Universitaire de Toulouse Université Toulouse Jean Jaurès Laboratoire de NeuroPsychoLinguistique
Julien PINQUIER	Université Toulouse III Paul Sabatier Institut de Recherche en Informatique de Toulouse

Remerciements

Un grand merci à Christine Meunier, Martine Adda-Decker et Hervé Glotin d'avoir accepté de rapporter sur cette Habilitation à Diriger les Recherches. Il m'a paru évident de croiser vos compétences afin de pouvoir avoir un regard différent et complémentaire sur mes travaux.

Un remerciement spécial à Julien Pinquier, pour avoir su être très patient avant de pouvoir avoir un exemplaire entre les mains de ce mémoire ! Merci pour ton travail remarquable à la direction de l'équipe de recherche et pour m'avoir supporté dans le même bureau depuis tant d'années ! Je garde un très bon souvenir de toutes les projets que nous avons développés et j'espère que nous continuerons encore longtemps ensemble !

Merci aux membres de l'équipe SAMOVA, présents et passés, et en particulier Isabelle Ferrané, Christine Sénac, Thomas Pellegrini, Hervé Bredin, Philippe Joly, Julie Mauclair pour vos soutiens et encouragements dans ma rédaction. Un grand merci aux stagiaires, doctorants et post-doctorants : c'est grâce à vous que tout a été possible ! Un grand merci à José, Sébastien, Mathieu, Timothy, Vincent, Robin, Lila pour leur collaboration, mais aussi pour tous les bons moments passés en déplacement ou en dehors du laboratoire !

Un grand merci à Virginie Woisard et Pascal Gaillard, qui ont chacun grandement influé sur mon parcours et avec qui j'ai passé de très bons moments.

Merci à Véronique Moriceau, Sandrine Mouysset, Maxime Le Coz et Corine Astésano pour votre amitié et les co-encadrements que nous avons réalisés ensemble !

Merci à Corinne Fredouille, Jean-François Bonastre, Muriel Lalain, Alain Ghio pour le temps passé à collaborer sur des projets dans la bonne humeur, mais également pour tous les bons moments lors des déplacements !

Merci à Jean-Marc Pierson, Dominique Longin et André-Luc Beylot, pour la bonne humeur et les réunions de travail constructives !

Un grand merci à Catherine Blanc, Charlotte Sicre, Stéphanie de la Hoz, Françoise Grelaud, Véronique Debats, Clémentine Roger, Chloé Bourbon, Léonor Araujo, Cyril Pouzac, Annie Planque, Thierry Bichot, Laurent Chauveau, Jean-François Gendet, Michaël Etesse, William Vincent, Romain Roure, Jacques Thomazeau, Julie Mballa, Philippe Manfré et toutes les personnes qui m'ont apporté leur support et font en sorte que la laboratoire fonctionne bien !

Et enfin merci à ma famille et mes enfants Eryn et Nolan de m'avoir permis de réaliser ce travail !

Je me rends compte qu'il va être difficile d'être exhaustif et citer toutes les personnes qui ont de près ou de long contribué à mes recherches ! Un grand merci à vous tous !

De nombreuses personnes m'ont grandement influencé tout au long de mon existence, et m'ont permis de réussir à m'épanouir professionnellement : merci à Ernest, Raymonde, Martine, Darvin, Paula, Alain, Agnès, Hélène, Jean-François, Isabelle et Jean-Philippe pour m'avoir donné un environnement familial riche et équilibré, vecteur de stabilité et de robustesse dans le travail.

Table des matières

Table des figures

Liste des tableaux

Liste des algorithmes

Glossaire

AM : Approche Métrique auto segmentale en prosodie
ASR : Automatic Speech Recognition (cf. TAP)
f_0 : fréquence fondamentale (fréquence de vibration des cordes vocales)
F1 : premier formant (fréquence de résonance du conduit vocal)
F2 : deuxième formant (fréquence de résonance du conduit vocal)
F3 : troisième formant (fréquence de résonance du conduit vocal)
CNN : Convolutional Neural Network ou ConvNet (acronyme anglais pour réseau de neurones convolutifs)
DNN : Deep Neural Network (acronyme anglais pour Réseau de Neurones Profond)
DTW : Dynamic Time Warping (acronyme anglais pour méthode de programmation dynamique pour l'alignement de séquences)
EMS : Envelope Modulation Spectrum (spectre d'enveloppe d'amplitude)
GMM : Gaussian Mixture Model (acronyme anglais pour modèle de mélange de Gaussiennes)
HC : Healthy Controls (acronyme anglais pour témoins sain)
HETD : Heure équivalent Travaux Dirigés
HINT : Hearing in Noise Test (évaluation de la perception auditive dans le bruit)
HNC : Head and Neck Cancer (cancer de la tête et du cou)
IA : Intelligence Artificielle
ORL : Oto-Rhino-Laryngologie (spécialité de médecine traitant les maladie de la tête et du cou)
RAP : Reconnaissance Automatique de la Parole
RNN : Recurrent Neural Network (acronyme anglais pour Réseau de Neurones Récurrent)
RNN : Support Vector Machines (acronyme anglais pour machine à vecteurs de support, ou séparateurs à vaste marge)
SHS : Sciences Humaines et Sociales
TAP : Traitement Automatique de la Parole
WER : Word Error Rate (acronyme anglais pour taux d'erreur mots)

Introduction générale

1 Contexte général

> « Mon art de maïeutique a les mêmes attributions générales que celui des sages-femmes. La différence est qu'il délivre les hommes et non les femmes et que c'est les âmes qu'il surveille en leur travail d'enfantement, non point les corps. »
>
> *Socrate dans le Théétète, Platon, 150B-150C, trad. A. Diès, 1926*

Le Théétète est un dialogue du IV^e^ siècle avant J.-C., écrit par Platon, sur la science et sa définition. Il met en scène Socrate et le jeune mathématicien Théétète. Socrate est en train de pratiquer son art de la maïeutique sur Théétète. Il lui rappelle que sa propre mère, Phénarète, était une sage-femme habile et renommée. Et il prétend que lui-même exerce le même art. La maïeutique se définit comme l'accouchement des esprits. Par le biais de questionnements, l'esprit du questionné parvient à trouver en lui-même les vérités. Elle permet à l'interlocuteur de « dire plus de choses que l'on n'en portait en soi ».

La maïeutique est donc l'art d'accoucher les esprits, de leur faire enfanter la vérité. Socrate, en philosophe, affirme que chacun porte en lui le savoir, sans en avoir conscience. Le questionnement vise à faire ressurgir ce savoir.

La lecture de ce texte m'a marqué : la description de cet art de la maïeutique de Socrate a fait naître beaucoup d'analogies avec le travail à réaliser pour l'écriture d'une habilitation à diriger les recherches, que je cherche à réaliser après deux décennies de recherches après mon doctorat. C'est une étape dans la vie du chercheur qui impose de se poser de nombreuses questions, de faire resurgir les expériences passées, de chercher le sens dans toutes les démarches scientifiques effectuées. C'est le moment où l'on s'arrête et l'on regarde derrière soi, c'est le moment où l'on pose par écrit l'organisation de notre expérience passée. C'est une période ou les questionnements et les remises en question foisonnent, mais qui permet d'avoir une conscience du travail réalisé sans pareil.

J'ai été très attaché, tout au long de ces années, à m'intéresser à la reconnaissance automatique de la parole, à la fois par une approche « bas niveau » (c'est-à-dire très proche du signal de parole), en me focalisant sur la modélisation acoustique des sons, mais également à une approche de plus « haut niveau » quand il s'agissait de travailler sur la prosodie. Ce positionnement a été salvateur pour éviter de s'éparpiller, mais également pour affirmer ma place dans mon équipe de recherche. C'est pour cela que la partie concernant la modélisation du langage m'a peu attiré, et cela m'a permis de me reposer sur ma collègue Isabelle Ferrané quand il fallait travailler sur la chaîne complète d'un système automatique de la parole. Je n'ai également pas cherché à m'intéresser à l'environnement du signal de parole, en me reposant sur l'expertise de Julien Pinquier dans ce domaine. Cette complémentarité avec mes collègues nous a permis de nous associer souvent dans des encadrements ou des projets de recherche, en profitant de nos compétences complémentaires.

Mon intérêt pour le traitement automatique de la parole provient du fait qu'il s'agit d'un domaine foncièrement multidisciplinaire : informatique, mathématique, linguistique, mais aussi physique et biologique. L'informatique pour son côté formalisation de la représentation et l'extraction de connaissances, traitement du signal, reconnaissance des formes, algorithmique, programmation, apprentissage automatique... Les mathématiques pour les modélisations, le cadre statistique, stochastique. La linguistique pour la phonétique, la phonologie, la syntaxe et la sémantique. La physique pour la mécanique et l'acoustique. La biologie pour tout ce qui a trait à l'être humain. Le domaine de l'étude du traitement de la parole est vaste et riche, ce qui le rend très attrayant et qui est très motivant pour partir sur des sujets de recherche. Motivation qui s'est d'autant plus renforcée que nous avons été approchés dans l'équipe de recherche par des médecins qui souhaitaient développer plusieurs approches pour généraliser les usages du traitement automatique de la parole au niveau hospitalier : l'objectif sociétal s'en est trouvé décuplé, en entrevoyant la possibilité d'influer sur les conditions de vie de nombreuses personnes.

Il est assez amusant de voir les évolutions des modes au niveau de la terminologie employée dans le monde de la recherche. Nous sommes actuellement en pleine période d'Intelligence Artificielle ! L'IA a envahi les médias grand public : par exemple les deux hors-séries de Libération « Voyage au cœur de l'IA » et « L'IA au cœur de l'humain » [Bloch and Joffrin, 2017, Bloch and Joffrin, 2018], mais également sous forme romancée : « M, le bord de l'abîme » de Bernard Minier [Minier, 2019]. Et depuis début 2022 nous assistons à un embrasement des usages : génération d'image, avec par exemple Stable Diffusion [Rombach et al., 2022], DALL-E [Ramesh et al., 2022] et Midjourney [1] [Holz, 2022] et génération de texte avec ChatGPT [OpenAI, 2022]. Ce dernier a mis en ébullition les médias : d'une part, car il peut produire des résultats que l'on ne pensait pas possibles pour un algorithme automatique, d'autre part car la quantité d'information qui l'alimente est gigantesque et enfin, car beaucoup voient cela comme une technologie disruptive, qui pourrait changer le paysage de la recherche sur internet et mettre en difficulté des entreprises géantes que l'ont pensait bien ancré dans le paysage [Lausson, 2023]. Que de chemin parcouru depuis Eliza [Weizenbaum, 1966] !

Depuis mon arrivée dans le monde de la recherche en 1997, le paysage des techniques a bien évolué, et a nécessité une constante remise en question des savoirs. Le point pivot le plus important dans mon parcours a eu lieu en 2010, lorsque l'IA par réseaux de neurones profonds (DNN) a révolutionné le traitement automatique de la parole. Des recherches éprouvées depuis les années 1990 ont laissé la place aux DNN, qui, bien que qualifiés de « boîtes noires », ont imposé un changement brutal de par les performances qu'ils arrivaient à obtenir. Nous avons maintenant dompté ces techniques, et nous arrivons même à un point ou l'explicabilité de ces systèmes est très étudiée et efface ce caractère sombre qu'elles pouvaient avoir au départ. L'évolution des ordinateurs et la disponibilité sur le réseau mondial des articles et codes des programmes, nous permettent maintenant une recherche beaucoup plus rapide et transparente qu'elle ne l'était au début de ma carrière.

2 Organisation du document

Chronologiquement, ce document a été rédigé en commençant par la partie III, le Curriculum Vitæ. Cette partie cherche à rendre compte, de la façon la plus complète possible, du déroulement de ma carrière d'enseignant-chercheur depuis la soutenance de mon doctorat en 2002 [Farinas, 2002]. Le développement est plus riche que la version succincte demandée par l'École Doctorale, mais elle m'a permis d'avoir une vue complète sur mes travaux. Je ferai des renvois vers cette partie quand je passerai en revue les recherches effectuées.

1. Vous trouverez d'ailleurs à la fin de chaque chapitre des illustrations que j'ai générées par Midjourney et qui vous donneront un aperçu du niveau actuel de ce genre d'application

Ensuite j'ai rédigé la partie I, qui m'a demandé de faire des choix dans les travaux réalisés. J'ai décidé de vous les présenter en suivant le fil conducteur de la parole :

- La parole : un signal temporel et des modélisations en constante évolution (chapitre 1). Ce chapitre me permet de détailler les différentes modélisations que j'ai utilisées pour représenter la parole. Il s'agit de considérer la parole d'un point de vue purement acoustique et de référencer les réalisations qui ont parsemé mon parcours.
- La parole : un signal perturbé… mais prévisible (chapitre 2). Ce chapitre rend compte des travaux qui ont été effectués afin d'essayer de prédire les résultats d'un système automatique de reconnaissance de la parole, mais avant de l'appliquer, en se basant sur des analyses rapides liées au signal lui-même.
- La parole : à la recherche de l'intelligibilité (chapitre 3). Ce chapitre traite plus particulièrement les problématiques qui ont émergé pour essayer de mesurer l'intelligibilité de la parole, en considérant le problème d'un point de vue de la perception de la parole, mais également au niveau de la production de la parole.

La partie II se place sur un point de vue plus prospectif, en éclairant en particulier deux aspects que je souhaite développer par la suite :

- Le traitement automatique de la prosodie (chapitre 1). Ce chapitre rassemble les différents travaux en cours pour représenter automatiquement la prosodie, à travers différents domaines d'application.
- La modélisation de la déglutition (chapitre 2). Ce chapitre détaille les travaux en cours en vue de traiter des signaux différents de la parole, mais néanmoins complémentaires : tout ce qui concerne la déglutition, la toux et les autres phénomènes qui sont produits au niveau de la gorge autre que la parole.

En annexe se trouvent les documents demandés par la commission d'habilitation, et en particulier les cinq articles principaux et le diplôme de doctorat.

FIGURE 1 – Un scientifique à la manière de Joan Miró (généré par Midjourney en février 2023, moteur de rendu n°4)

Première partie

Travaux de recherche

Introduction

> “ Rien à voir avec nos bons vieux rapports de chez nous pianotés sur ordinateur d'un doigt frileux par un pote qui a longtemps hésité entre la carrière de flic et celle de maréchal-ferrant. Chez les Yankees, tu causes et ça s'enregistre, propre et en ordre, sans rature, répétition ni impropriété de termes. N'importe quelle crapule, ayant appris à lire sur une machine à sous, te torche des aveux en comparaison desquels la Confession d'un enfant du siècle passerait pour le mode d'emploi d'une poudre insecticide traduit du romanche.
>
> Pour piloter ce machin, y avait pas besoin de sortir de Princeton. Tout ce qu'avait à branler l'opératrice, en dehors de son touffu joli, c'était de faire répéter un mot mal prononcé, et encore l'appareil suggérait-il une tripotée de synonymes concordant avec le sens de la phrase. En voyant fonctionner l'engin, je pensai à tous mes confrères trémulsés de la coiffe dont il rendrait la prose intelligible. ”
>
> *Frédéric Dard alias San Antonio, « Du sable dans la vaseline », sept. 1998, p. 40-41*

La reconnaissance de la parole qui impressionnait San Antonio en 1998 [Dard, 1998] a beaucoup évolué. Cela coïncide avec le démarrage de mon doctorat, et cela correspondant à l'utilisation de la programmation dynamique et la modélisation par modèles de Markov cachés et mélange de lois normales (HMM-GMM) avec la boîte à outils HTK [Young and Young, 1993], dont l'équipe Interface Homme Machine Parole Texte (IHMPT) de l'IRIT avait acquis une licence d'utilisation. Cette approche de la modélisation d'un système automatique de la parole a été balayée après les années 2010 : il est alors devenu incontournable d'utiliser les réseaux de neurones profonds qui ont presque divisé par deux les scores d'erreurs et ont par conséquent redonné une dynamique à la recherche en traitement de la parole, jusqu'à parvenir à égaler voire surpasser l'humain dans cette tâche (cf. figure 1.1). Le chapitre 1 synthétise cette évolution et présente un résumé des principes sous-jacents à la modélisation de la parole, un signal dont la composante temporelle n'est pas à négliger quand on souhaite le représenter.

Dans le chapitre 2 je m'intéresse plus particulièrement aux causes de variabilité du signal de parole. Cela a été fait pour répondre au verrou suivant : « est-on capable de prédire, avant d'appliquer un système automatique complet, la qualité de transcription atteignable ? ». Cela provient d'un verrou scientifique de l'entreprise Authôt, qui avait à sa disposition plusieurs systèmes de transcription et qui cherchait à optimiser leur usage, pour rationaliser les coûts, mais également pour informer le client sur la qualité *a priori* qui pourrait être obtenue sur une tâche de transcription. Cela nous a amené à détailler les différentes sources de variabilité des systèmes de reconnaissance, et d'estimer ces différentes perturbations afin de modéliser la difficulté de transcription.

Dans le chapitre 3, j'ai rassemblé les travaux entrepris depuis 2012 visant à simuler automatiquement l'intelligibilité perceptive d'une personne, afin de pouvoir régler des prothèses auditives, mais également à mesurer automatiquement l'intelligibilité de personnes pathologiques. La première est issue d'une collaboration avec la société Archean. La seconde concerne plusieurs projets en collaboration

avec l'unité Voix et Déglutition du CHU de Toulouse. Ces projets nous ont amenés à constituer différents corpus afin de valider nos expérimentations. Le détail de ces ressources, qui ont été constituées pour répondre à cette problématique, sera également décrit dans ce chapitre.

FIGURE 2 – Enregistrement audio sur ordinateur à la manière de Vincent Van Gogh (généré par Midjourney en mars 2023, moteur de rendu n°4)

Chapitre 1

La parole : un signal temporel et des modélisations en constante évolution

Sommaire

1.1 Introduction

Le signal de parole est un signal qui est très variable. Certains le considèrent comme une variable aléatoire. C'est une définition pratique pour pouvoir utiliser l'artillerie statistique, mais cela occulte la façon dont ce signal est formé : il est important, pour bien le modéliser, de prendre en compte l'enchaînement temporel de ce signal. L'importance de cette dynamique légitime les modélisations qui sont apparues au fil du temps : la programmation dynamique [Sakoe and Chiba, 1978], la modélisation par modèles de Markov cachés [Rabiner and Juang, 1986], les réseaux bayésiens [Deviren and Daoudi, 2002], les réseaux de neurones récurrents [Graves et al., 2006,Graves et al., 2013]. La figure 1.1 rassemble les principales évolutions des systèmes de Traitement Automatique de la Parole (TAP), en observant la difficulté croissante introduite dans les campagnes à chaque fois que le problème était considéré comme résolu, c'est-à-dire que les performances égalaient celles qu'un humain aurait pu obtenir. J'ai compilé sur ce schéma les principales campagnes d'évaluation nationales et internationales qui ont rythmé de manière continue les recherches en traitement automatique de la parole. Le NIST, institut de standardisation américain, a joué un grand rôle dans l'organisation de ces challenges. Au départ, ces évaluations avaient pour but de donner des indications sur les bons choix d'investissement du gouvernement américain. Au final, ces évaluations ont pris une dimension plus internationale et ont posé les jalons dans l'état de l'art de nombreux domaines. Cela a apporté de la transparence aux recherches menées, mais également une émulation internationale qui a permis de concentrer l'intérêt de nombreuses équipes de recherche sur la parole. En France, une initiative a été impulsée, par le Centre d'Expertise Parisien de la Délégation Générale de l'Armement (DGA) et l'Association Francophone de la Communication Parlée (AFCP), au début des années 2000 pour essayer de transposer cette situation américaine au traitement de la parole sur le français. Le pendant français du NIST, l'Association française de normalisation (AFNOR),

a été impliqué pour organiser plusieurs campagnes d'évaluation de systèmes de transcription enrichie d'émissions Radiophoniques (ESTER).

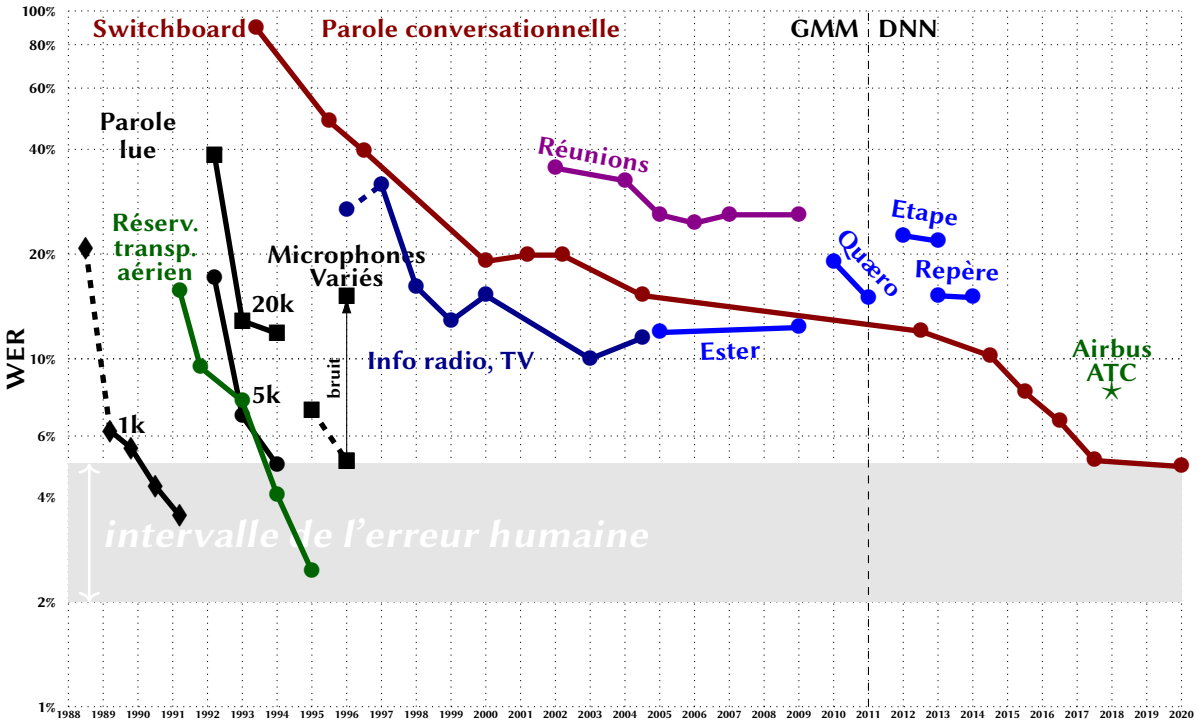


FIGURE 1.1 – Évolution des performances (taux d'erreur de mots en échelle logarithmique) des systèmes TAP de 1988 à 2020

On retrouve sur la figure 1.1 sur la partie gauche les premières campagnes américaines. Celles-ci ont assez rapidement dépassé les performances humaines : en effet, il s'agissait de tâches assez simples, avec un vocabulaire limité (1000, 5000 puis 20000 mots), sur de la parole lue. La complexité a donc été augmentée progressivement, jusqu'à intégrer des environnements d'enregistrements de plus en plus variables. Les modèles de Markov cachés associés à des modèles de mélanges de lois normales (HMM-GMM) ont régné en maître entre les années 1990 à 2010. Les innovations qui ont eu cours sur cette période ont principalement porté sur l'optimisation de la paramétrisation du signal de parole [Saon and Chien, 2012]. Or en 2010, nous avons vu débarquer des systèmes basés sur les réseaux de neurones, qui, après avoir apporté un vent frais sur les performances des systèmes de reconnaissance d'image, ont commencé à envahir le domaine de la reconnaissance de la parole, et permis une amélioration sensible des performances lorsque les données d'apprentissage avoisinaient le millier d'heures : là où les HMM-GMM ne permettaient pas de capturer toute la richesse des informations à modéliser [Hinton et al., 2012]. Les données de parole conversationnelles (représentées par le corpus Switchboard [Godfrey et al., 1992], ligne rouge sur la figure) ont débuté avec un taux d'erreur de mots (WER) de plus de 89% en 1993 pour arriver à 4,9% en 2020 [Wang et al., 2020]. Sur ce corpus, plusieurs études ont été menées pour estimer le taux d'erreur réalisé par des humains : celui-ci a été évalué à 5,9% en 2016 par Microsoft [Xiong et al., 2016], ce qui a été confirmé par IBM et Appen en 2017 [Saon et al., 2017] : la machine produit donc maintenant moins d'erreurs sur ce corpus de 500h de parole conversationnelle ! Les campagnes représentées en bleu clair sur la figure 1.1 sont issues des évaluations françaises menées à partir du programme de recherche Technolangues [2]. Les résultats sont en adéquation avec la tendance globale sur

2. http://www.technolangue.net

la parole conversationnelle. J'ai également situé les résultats obtenus par la campagne d'évaluation des systèmes sur la reconnaissance de la parole dans le cadre du transport aérien, ce qui reste cohérent avec l'augmentation de la difficulté liée à un modèle de langage très spécifique et des conditions d'enregistrement fortement dégradées. Je détaille dans la section 5.5 (page 87) mes expériences et mes participations dans le domaine des campagnes d'évaluation. Dans ce chapitre, je souhaite revenir sur la chronologie des techniques : de la programmation dynamique aux systèmes de reconnaissance automatique par réseaux profonds, en passant par les modèles de Markov cachés et apporter un regard critique à leurs égards, tout en situant mes contributions.

1.2 Programmation dynamique

La programmation dynamique (cf. annexe D pour plus de détails sur la méthode) est une des plus anciennes méthodes ayant permis la reconnaissance de mots clés de manière efficace. Elle est toujours utilisée quand la puissance de calcul disponible est limitée. C'est pour cela que c'est une technique toujours utilisée en embarqué. Elle prend bien en compte la variabilité temporelle du signal de parole, car elle ne pénalise pas énormément les contractions et allongements possibles sur le signal. Cette élasticité permet de correctement représenter les séquences acoustiques. L'objectif de la programmation dynamique est de produire une mesure de dissemblance entre deux séquences. On peut donc réaliser de la reconnaissance de la parole en identifiant les mots les plus proches parmi un dictionnaire constitué de mots de référence.

Cette méthode et sa simplicité en font une bonne introduction à l'apprentissage du traitement automatique de la parole. Je m'en sers au niveau de mes cours pour introduire la contrainte de la prise en charge de l'évolution temporelle de la parole, et elle se prête bien à une mise en œuvre lors de travaux dirigés. C'est également un bon moyen pou débuter sur l'échelle des difficultés éducatives au niveau de l'apprentissage des systèmes de reconnaissance de la parole. La mise en pratique est également intéressante, car l'algorithme utilisé permet d'arriver à une réalisation par les étudiants en très peu de temps. J'ai pu également superviser avec Sandrine Mouysset des étudiants en projets longs à l'ENSEEIHT pour qu'ils proposent une implémentation souple et modulaire destinée à proposer de piloter des drones volants par reconnaissance de la parole sur un smartphone. Nous avons mis en pratique ces programmes pour l'apprentissage de l'unité d'enseignement Modélisation et Calcul Scientifique pour les étudiants du Master 1 Informatique. Nous avons proposé un projet qui visait à ce que les étudiants injectent l'algorithme dans le projet et développent une démarche expérimentale visant à tester les différents enregistrements qui leur étaient proposés.

Mais j'ai eu à implémenter cette méthode de façon plus industrielle également (cf. §5.6.3 page 90). En effet, les contraintes matérielles du dispositif Wizzili qui doit être installé dans les domiciles et pouvoir répondre à des commandes vocales, ne permettaient pas l'intégration de systèmes plus évolués. Le dispositif est relié à internet, mais il y avait des contraintes au niveau du temps de réaction qui nécessitait que le dispositif sache détecter les mots clés déclencheurs en interne. Les entrepreneurs souhaitaient également qu'il n'y ait pas ou peu de phase d'apprentissage pour pouvoir utiliser le dispositif. J'ai donc proposé l'utilisation de programmation dynamique pour réaliser cette détection de parole. La méthode a été complètement réimplémentée en langage C par Wishinnov, et interfacée avec les bibliothèques standard d'extraction de paramètres [ETSI, 2007]. Un débruitage a été réalisé et le signal de parole a été représenté par une paramétrisation MFCC normalisée. L'algorithme a été adapté de manière à pouvoir fonctionner en parole continue [Sakoe, 1979]. De nombreuses optimisations ont été appliquées pour que cette recherche de mot clé soit la plus rapide possible et dégage du temps de traitement pour les autres fonctionnalités temps réel du dispositif. Cette détection permet de déclencher l'analyse de la

parole et du dialogue réalisée sur un serveur distant. J'ai été consulté très régulièrement en 2017–2018 afin de répondre aux demandes d'expertise des développeurs.

Cette technique reste intéressante quand elle ne met en jeu que quelques mots à reconnaître. Mais même pour des mots clés, il est plus intéressant actuellement de se tourner vers des modélisations plus performantes comme les réseaux de neurones appris sur de très gros corpus. Si bien sûr l'on dispose de suffisamment de capacité de calcul et que l'on n'est pas limité par la consommation énergétique.

1.3 Modèles de Markov Cachés

Les modèles de Markov cachés (HMM) ont été très utilisés en reconnaissance de la parole dès mon arrivée au laboratoire. L'IRIT disposait de licences de la plateforme logicielle Hidden Markov Model ToolKit (HTK) [Young and Young, 1993] diffusée par la société Entropic à Cambridge. C'est aussi un outil qui était très utilisé à l'Université Toulouse III pour enseigner les travaux pratiques de reconnaissance de la parole (mis en place par Christine Dours-Sénac).

La première réalisation en HMM sous HTK a consisté en la modélisation de brame du cerf. En effet, David Reby, chercheur en biologie, était venu nous trouver, car il souhaitait effectuer, sur les données issues de ses campagnes d'acquisition lors de période de brame, une segmentation et regroupement en animal ! J'avais avec Arnaud Galinier réalisé des modélisations qui avaient marché de manière correcte. Ces travaux, précurseur sur l'analyse de voix animale, ont été publiés dans un article de JASA [Reby et al., 2006].

J'ai ensuite été amené à travailler avec Isabelle Ferrané et Julien Pinquier sur la mise en place d'un système de reconnaissance grand vocabulaire de la parole en français. J'ai mis en place l'infrastructure globale du système et me suis concentré sur la modélisation acoustique avec des HMM. Isabelle travaillait sur la modélisation du langage et Julien sur les segmentations préalables à la reconnaissance de la parole (détection d'activité vocale, détection du locuteur). Nous avons utilisé JULIUS [Lee et al., 2001], une plateforme de reconnaissance de la parole qui permettait de travailler en temps réel et qui exploitait les paramétrisations et les modèles au format d'HTK. Nous avons participé à la campagne ESTER où nous avons pu soumettre un système qui a réussi à traiter tous les fichiers de l'évaluation. Les résultats étaient perfectibles [de Calmès et al., 2005], nous avions manqué de temps et de main d'œuvre pour optimiser le système. Cela nous a néanmoins permis de disposer d'une base qui nous a été utile pour les recherches qui ont suivi.

J'ai ensuite utilisé la modélisation acoustique pour extraire des segmentations phonétiques qui ont servi dans nos travaux sur la mesure de l'intelligibilité (cf. §3.4 page 32). J'ai été amené à utiliser également les modèles acoustiques développés par le LIUM sous la plateforme SPHINX. Cela a permis de mettre en place le système de mesure de l'intelligibilité permettant de faciliter le réglage de prothèses auditives (cf. §5.4.9 page 79).

J'ai également utilisé les HMM et HTK dans le cadre de la modélisation des événements liés à la déglutition (cf. §2.2 page 51). En effet, je disposais de peu d'exemples d'apprentissage pour modéliser la déglutition avec le jeu de données dont je disposais en 2019, et l'usage des HMM m'a permis de proposer une modélisation qui prenait en compte la variabilité temporelle des observations.

1.4 Modèles à base de réseaux de neurones profonds

Les réseaux neuronaux sont apparus comme une approche de modélisation acoustique attrayante pour la reconnaissance automatique de la parole à la fin des années 1980. Les réseaux neuronaux font moins d'hypothèses explicites sur les propriétés statistiques des caractéristiques que les HMM et présentent plusieurs qualités qui en font des modèles de reconnaissance attrayants pour la reconnaissance

de la parole. Lorsqu'ils sont utilisés pour estimer les probabilités d'un segment de la parole, les réseaux neuronaux permettent un apprentissage discriminatif de manière naturelle et efficace. Cependant, malgré leur efficacité dans la classification d'unités temporelles courtes telles que les phonèmes individuels et les mots isolés, les premiers réseaux neuronaux étaient rarement efficaces pour les tâches de reconnaissance continue en raison de leur capacité limitée à modéliser les dépendances temporelles.

Dans le cadre des projets sur la modélisation automatique de la langue (cf. §5.4.3 page 77), j'ai été amené à collaborer avec l'Institut des Sciences Cognitives de Lyon pour mettre en place des solutions basées sur des RNN. Nous n'avons pas réussi a obtenir de meilleurs résultats que les techniques que nous utilisions alors pour réaliser l'identification des langues.

En 2012, du fait de la présence de corpus de plus en plus importants en taille (par exemple Librispeech [Panayotov et al., 2015] avec plus de 1000 heures de données alignées), les réseaux de neurones ont surpassé les modélisations par mélange de lois normales (GMM) dans les approches hybrides avec des HMM. La plateforme KALDI [Povey et al., 2011b] s'est démarquée dans la communauté scientifique, en offrant une augmentation significative des performances grâce à des méthodes récentes. C'est le système que nous avons utilisé dans le doctorat de Sébastien Ferreira (cf. §6.2.2 page 98) pour réaliser les indices de performance de ses systèmes de reconnaissance de la parole grand vocabulaire.

Pour des tâches ne nécessitant pas des systèmes complets de reconnaissance de la parole, quand on cherche à modéliser des mots ou bien des séquences de taille plus ou moins fixe de signal audio, il m'arrive d'utiliser des réseaux de neurones convolutifs (CNN). En effet ces séquences temps-fréquences donnent de bons résultats lorsque les séquences à modéliser sont de l'ordre de la seconde. C'est ce qui a été réalisé avec Lila Gravellier (cf. §6.2.6 page 102) et Philippe Allet (cf. §26 page 97) dans leurs modélisations de séquences audio.

1.5 Modèles de « bout en bout » (E2E)

Depuis 2014, la recherche s'est beaucoup intéressée à la modélisation « de bout en bout » (E2E). Les approches traditionnelles basées sur la phonétique (c'est-à-dire tous les modèles basés sur les HMM) nécessitaient des composants et un apprentissage séparés pour la prononciation, le modèle acoustique et le modèle linguistique. Les modèles de bout en bout apprennent conjointement tous les composants de l'appareil de reconnaissance vocale.

Vincent Roger (cf. §6.2.4 page 100) et Jérôme Susgin (cf. §23 page 95) ont été amenés à utiliser ce genre de modélisations, mais dans un cadre plus contraint de la modélisation de la parole grand vocabulaire ; respectivement dans le cadre de la mesure de l'intelligibilité et de la modélisation des émotions.

1.6 Conclusion

Le paysage de la modélisation automatique de la parole a bien évolué depuis mon arrivée à l'IRIT. Les expertises sur les HMM acquises au fur et à mesure des enseignements et activités de recherche on du laisser la place aux modèles hybrides HMM-DNN puis aux approches par modèles de « bout-en-bout ». Les performances des systèmes de reconnaissance de la parole grand vocabulaire atteignent maintenant des performances remarquables dans un cadre généraliste, ce qui permet d'envisager de nouvelles applications dérivées et ouvre le domaine de la reconnaissance automatique de la parole à de nouvelles communautés. Nous avons également cherché à pérenniser et rendre accessible nos modèles de reconnaissance à travers le projet PATY (cf. §5.4.18 page 85).

Dans le chapitre suivant, je détaille les recherches que nous avions menées afin d'évaluer les différents systèmes et analyser leurs forces et faiblesses. Et l'on verra qu'il est possible directement en analysant le signal d'arriver à prédire le taux d'erreur de mots des systèmes avant leur application.

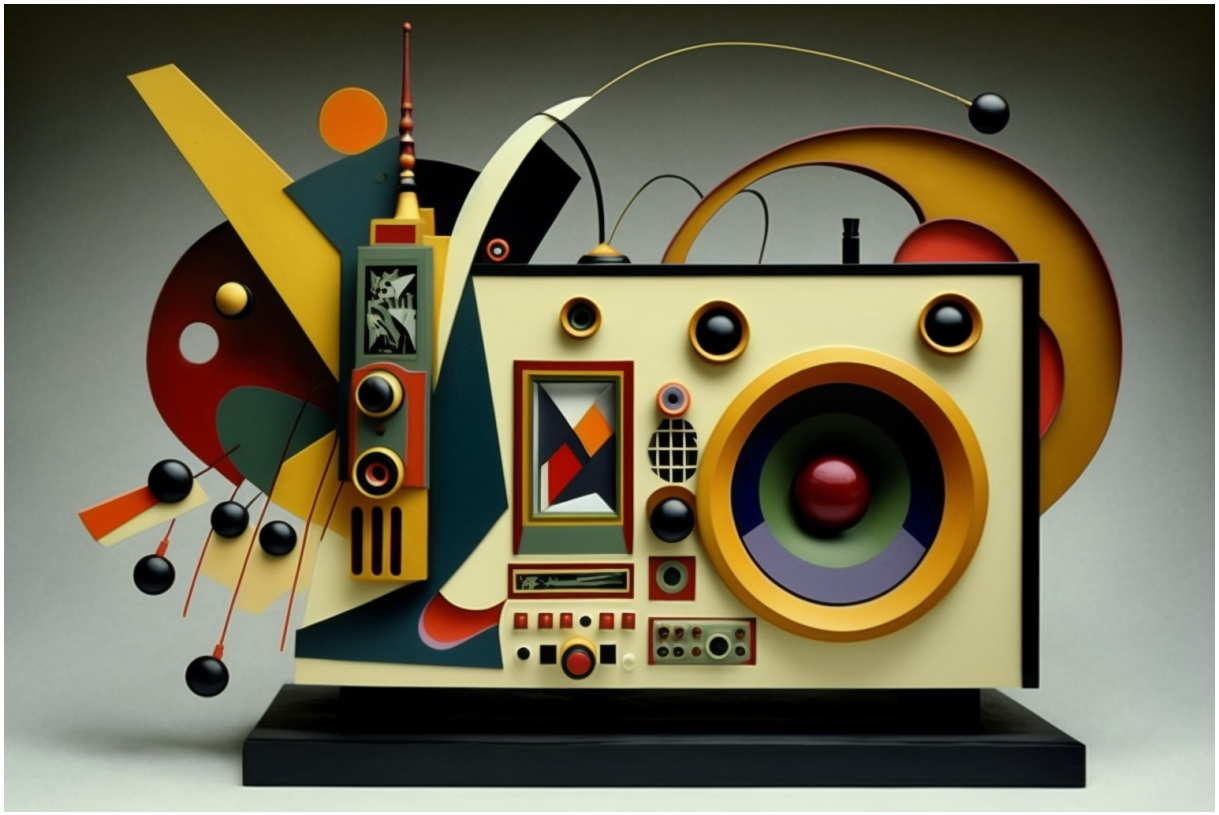


FIGURE 1.2 – Les réseaux de neurones pour le traitement automatique de la parole à la manière de Vassily Kandinsky (généré par Midjourney en mars 2023, moteur de rendu n°4)

Chapitre 2

La parole : un signal perturbé… mais prévisible !

Sommaire

2.1 Introduction

Les systèmes de transcription de la parole affichent actuellement de très bonnes performances : on avoisine les 3% d'erreur sur les mots, ce qui est en dessous des performances moyennes de l'humain. Ceci est atteint sur des corpus composés de parole lue ou peu difficile. Mais comment les systèmes évoluent-ils lorsque la tâche se complique ? La section 2.2 cherche à répondre plusieurs questions ci-après. Comment se comportent les différents systèmes de reconnaissance académiques et commerciaux sur de la parole atypique ? Quelles sont les conditions qui mettent le plus à mal les systèmes ? On a pu répondre à ces questions en 2018 grâce à une prestation de la SAT TTT. Cela nous a permis d'obtenir un panorama de la situation, et de nous indiquer les verrous et enjeux dans le domaine de la transcription de la parole. La section 2.3 présente une synthèse des sources de variabilité du signal de parole. Ces différentes formes prises par les signaux peuvent avoir des répercussions plus ou moins grandes sur les systèmes de reconnaissance de la parole. La section 2.4 correspond à des travaux menés lors du doctorat de Sébastien Ferreira (cf. section 6.2.2) entre 2016 et 2020. Il s'agit de répondre à la question : est-on capable de de prévoir les performances d'un système de transcription de la parole en analysant *a priori* le signal ? Il s'agit d'une problématique de recherche qui a émergé avec la collaboration avec la société Authôt. Celle-ci propose des services de transcription, effectués par des humains ou bien

effectués de manière automatique par des systèmes de reconnaissance de la parole. Une prédiction qui pourrait être effectuée avant l'utilisation de services de transcription permettrait de renseigner les utilisateurs des services Authôt, de la qualité à prévoir pour une transcription automatique, et donc mieux l'aiguiller vers les choix de services possibles. En dehors de ces considérations économiques et pratiques, la problématique constitue un vrai challenge, et permettrait de proposer des analyses à même d'identifier et de caractériser les signaux représentant la parole.

2.2 Variabilité du signal de parole et impact sur la reconnaissance

Pour évaluer les performances des différents systèmes de transcription automatique de la parole, une étude financée par TTT (cf. section 5.6.2) a été menée en 2018. L'objectif est de dresser un panorama des performances des systèmes de reconnaissance de la parole et de dégager les verrous et les enjeux. La prestation a duré trois mois, et a permis de constituer un corpus d'évaluation, qui permet de représenter une grande variété de situations.

2.2.1 Corpus de test

La langue visée par cette étude est le français. En effet, nous avons constaté qu'à la suite des grandes campagnes nationales ESTER [Gravier et al., 2004, Galliano et al., 2005], ESTER2 [Galliano et al., 2009] et ETAPE [Galibert et al., 2014], il était difficile de se rentre compte des performances sur les systèmes français. Il est beaucoup plus simple de trouver des études en anglais, du fait de la présence de corpus bien diffusé au niveau international. On trouve de nombreuses pages d'information qui classent les systèmes sur différents corpus anglais [Synnaeve, 2022, Szymański et al., 2020, Stojnic et al., 2022]. Le cahier des charges établi initialement devait permettre de répondre aux caractéristiques suivantes pour la parole traitée :

- disposer de mots clefs simples, afin de se rapprocher de tâches typiquement utilisées par des orthophonistes et neuropsychologues (fluence verbale, calcul de temps de réaction à la présentation d'un stimulus auditif ou visuel), qui rendent la reconnaissance assez complexe du fait du manque de contexte autour des mots à reconnaîtrez.
- proposer différents styles de parole : chroniques, journaux d'information radio (pour disposer d'un référentiel de parole préparée), mais également des productions plus spontanées.
- présenter différentes conditions de bruit environnant (bruit synthétique, mais également naturel).
- être issu de personnes présentant des difficultés au niveau de la production de la parole (parole pathologique).
- être produite par des personnes dont le français est une langue seconde.

L'objectif était d'obtenir sur une quantité restreinte de données (moins de 7h au total), une forte hétérogénéité des conditions d'acquisitions, de qualité de production et de contexte.

Le corpus est constitué de trois catégories :

- de la parole préparée, souvent lue :

  BREF80 extrait du corpus BREF80, composé de paroles lues (textes du journal Le Monde),enregistré dans un studio [Lamel et al., 1991].

  ESTER une tranche matinale de France Inter, composée de journaux d'informations et de billets d'intervenants [Gravier et al., 2004].

QUÆRO données radiophoniques issues de France Culture et France Inter, avec une alternance de parole préparée et de parole spontanée (débat, interventions téléphoniques …) [Lamel et al., 2011]

ACSYNT extraits du corpus oral du français contemporain ACSYNT, composé de trois types de données orales : de la lecture oralisée de texte, des monologues et des entretiens guidés [CLLE, 2013].

— de la parole en condition bruitée ou ambiante :

UPS-U4 81 itérations d'une diffusion de parole lue en conditions studio (BREF) à différentes distances source-microphone (pour analyser l'impact de la réverbération) [Decroix, 2017]

AIRBUS Enregistrement de conversations entre la tour de contrôle et le pilote : nombreux tours de parole et qualité d'enregistrement faible (transmission par radio) [Farinas and Pellegrini, 2018].

ARCHEAN Enregistrements d'une voix d'homme, de femme et d'enfant avec superposition d'un bruit conversationnel de type « babel noise » à différents rapports signal sur bruit [Fontan et al., 2017]

ESLO 2 Enregistrements avec des habitants de toute l'agglomération d'Orléans à partir de 2008 : parole naturelle, interview d'habitants dans la rue, dans les transports publics [Baude and Dugua, 2011]

ADREAM parole spontanée, distante au microphone, simulation de situations d'enseignement [Decroix, 2017]

NOISEX application de la banque de bruit NOISEX [Varga and Steeneken, 1993] sur de la parole à différents niveaux de SNR [Ferreira et al., 2019a]

— de la parole avec une production « atypique » :

PFC parole avec de nombreux accents régionaux, enregistrés en France, mais également à l'étranger (Québec, Burkina Faso…)

iPFC parole lue par des apprenants japonais [Detey and Kawaguchi, 2008, Racine et al., 2012]

ESTER 2 données de radios françaises internationales, avec de forts accents nord-africains [Galliano et al., 2009]

C2SI parole pathologique, de patients atteints de cancers de la tête et du cou [Woisard et al., 2022, Woisard et al., 2021]

Le tableau 2.1 détaille le nombre de fichiers et la durée de caque sous-corpus utilisé.

2.2.2 Systèmes évalués

Les systèmes évalués ont dû répondre à un cahier des charges assez drastique. En effet, nous disposions d'un budget contraint, ce qui ne nous a pas permis de faire tourner tous les systèmes commerciaux existants. Il faut noter que les systèmes évalués n'ont pas été optimisés pour les différents types de données présentes dans le corpus de test : nous souhaitions obtenir des résultats bruts sur les différents types de données. Donc aucune adaptation acoustique ou linguistique n'a été opérée. Au niveau académique nous n'avons pas été exhaustifs : nous n'avons pas eu le temps de demander à tous les laboratoires de produire des résultats.

Voici la liste des systèmes qui ont été évalués :

Authôt système utilisé en interne par la société Authôt au moment de l'étude (provenant de VOXOLAB)

TABLE 2.1 – Répartition des durées des corpus composant l'ensemble de test de la prestation TTT-ASR

Acronyme	Nombre de fichiers	Durée totale
BREF80	56	29,27 min
ESTER	1	29,38 min
QUÆRO	1	30,52 min
ACSYNT (journal lu)	9	13,22 min
ACSYNT (entretien)	5	9,44 min
ACSYNT (présentation)	5	7,14 min
UPS-U4	82	24,84 min
AIRBUS	4	10,26 min
ARCHEAN	90	62,29 min
ESLO 2	6	27,75 min
ADREAM	2	26,72 min
NOISEX	135	39,38 min
PFC	14	41,89 min
iPFC	26	8,73 min
ESTER 2	1	9,11 min
C2SI	27	15,75 min

Bing logiciel de reconnaissance utilisé pour la commande vocale sous Windows

Google transcription généraliste de Google, accessible sous forme d'API

IBM transcription généraliste d'IBM, accessible sous forme d'API

IRIT un système sous KALDI GMM-HMM (sans détection performante des zones de parole) [Heba, 2021]

LIA un système sous KALDI développé par le LIA

Nuance un logiciel de dictée vocale

SpeechMatics transcription généraliste de SpeechMatics, accessible sous forme d'API

Sphinx système de reconnaissance généraliste (version Pocket, travaillée pour l'embarqué)

Wit transcription généraliste de Facebook, accessible sous forme d'API

2.2.3 Résultats et analyse des résultats de la reconnaissance de la parole

Nous avons produit de nombreuses analyses. Vous trouverez dans l'annexe E page 197 une version exhaustive des résultats calculés.

La figure 2.1 détaille les résultats généraux. Il s'agit du score global : tous les résultats ont été regroupés, et il a été réalisé une pondération pour garder un équilibre sur tous les corpus.

Le meilleur résultat est obtenu par le système de Google, qui réalise 49% de bonnes reconnaissances. Le système du LIA suit avec 46% et Facebook avec 43%. Le système qui ferme la marche est celui d'IBM avec un peu moins de 5%. La variété des résultats est due à de mauvaises performances de certains systèmes sur certaines tâches, nous le verrons ci-dessous. Mais il est intéressant de noter de manière globale, qu'en fournissant à reconnaître des fichiers avec des niveaux de difficulté variable, on peut dans le meilleur des cas obtenir un mot sur deux de juste seulement ! Il s'agit certes d'un instantané réalisé en 2018, mais nous sommes loin des résultats classiques sur de la parole propre.

Les résultats suivants seront présentés tâche par tâche, avec un code couleur pour chaque système (du bleu foncé au jaune) :

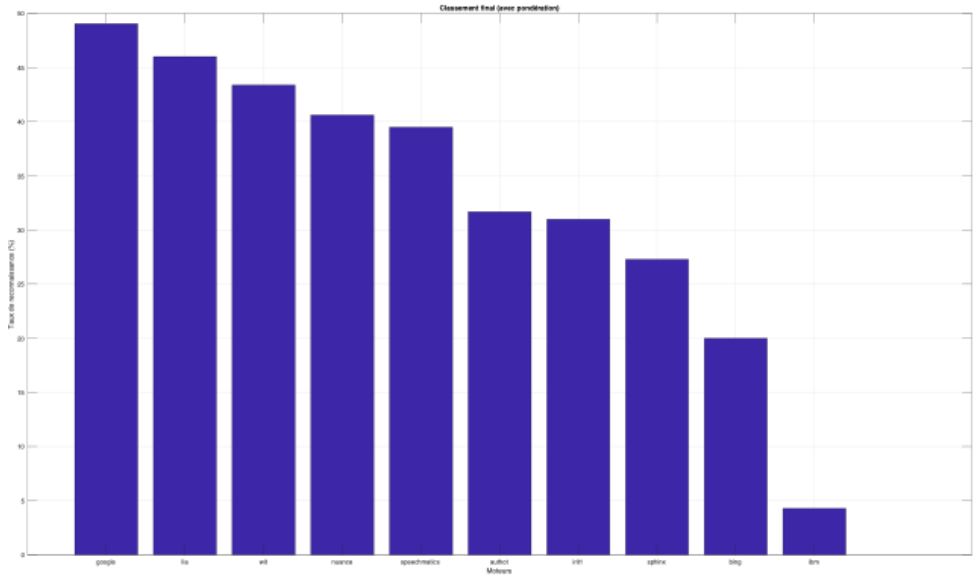


FIGURE 2.1 – Résultats pondérés de l'analyse de système ASR : résultats généraux pondérés

— Authôt (bleu foncé)
— Bing
— Google
— IBM
— IRIT
— LIA
— Nuance
— SpeechMatics
— Sphinx
— Facebook (jaune)

La figure 2.2 détaille les résultats avec de la parole préparée. C'est le cas de figure qui donne les meilleurs résultats. La figure présente un détail sur les 3 conditions du corpus ACSYNT (entretien, présentation et texte lu), puis BREF, ESTER et QUÆRO. Le dernier diagramme présente un score amalgamé. IBM et SPHINX présentent des résultats assez faibles, ils ne sont pas forcément appris pour traiter un grand vocabulaire. Cette fois-ci on obtient des résultats qui sont entre 70% et 80% sur les extraits de corpus qui ont été choisis.

La figure 2.3 détaille les résultats avec de la parole bruitée. On note tout de suite que le corpus Archean est redoutable : il s'agit de fichiers où l'on a appliqué une simulation de presbyacousie. Les systèmes ne sont pas faits pour résister à cette dégradation. Notons toutefois Bing qui s'en sort bien. Mais vu qu'il s'agit plutôt d'un système de commande vocale, cela reste somme toute logique. Google reprend le dessus dès que les enregistrements présentent des phrases. Les bruits urbains de ESLO2 et les dégradations avec du « babble noise » sont bien les tâches les plus difficiles. Les sons de machines de NOISEX semblent présenter moins de difficultés aux systèmes.

La figure 2.4 détaille les résultats avec de la parole « atypique ». Le premier groupe, le corpus C2SI, est la tâche la plus difficile pour les systèmes. Le taux de reconnaissance ne dépasse pas 40%. Les altérations situées au niveau de la production de la parole sont parfois très importantes, cela restera une tâche très difficile même après que l'on aura réalisé des adaptations aux systèmes. Les trois autres groupes représentent les variantes de prononciation : accent nord-africain, locuteurs japonais

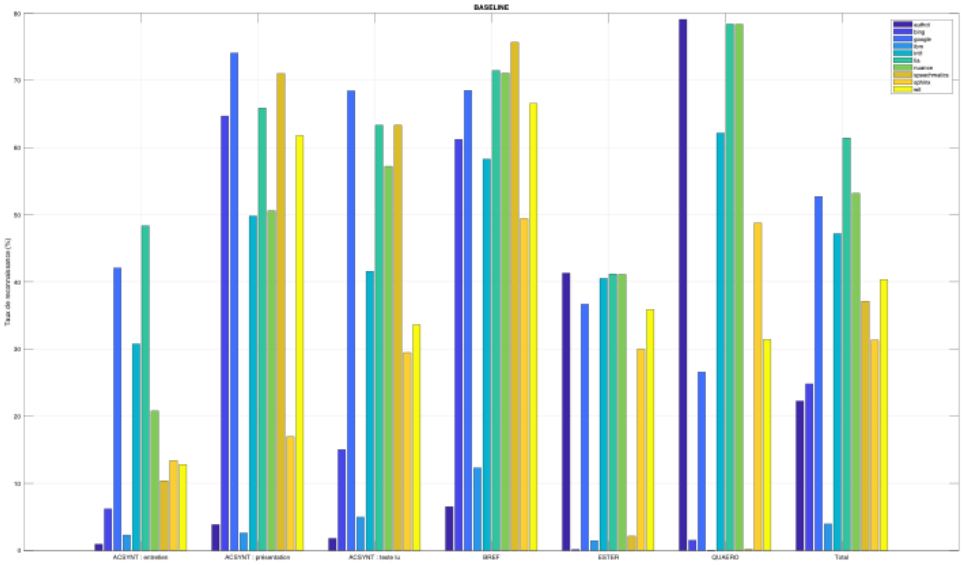


FIGURE 2.2 – Résultats pondérés de l'analyse de système ASR : parole préparée

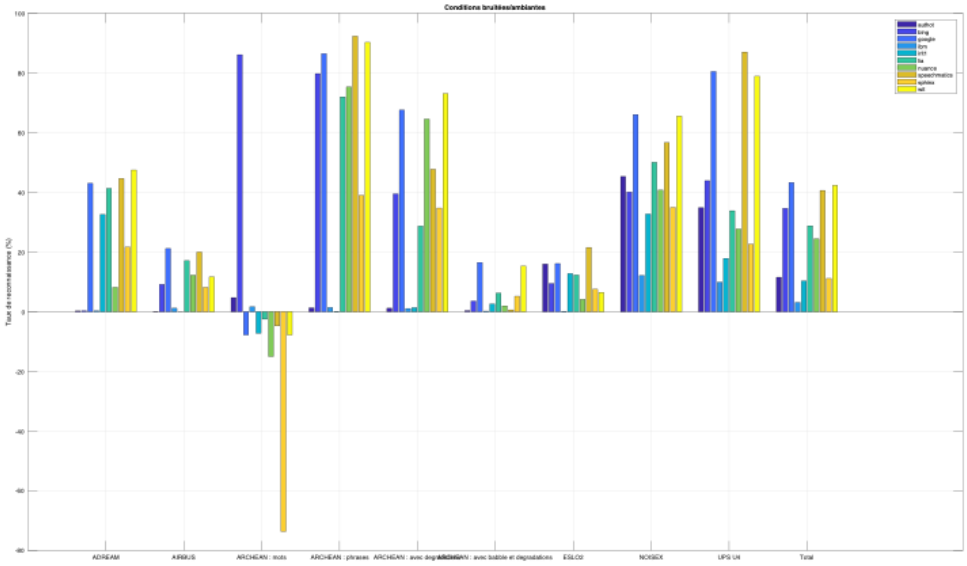


FIGURE 2.3 – Résultats de l'analyse de système ASR : parole en condition bruitée

apprenant le français, parlers régionaux. Les résultats sont globalement bons, mais Speechmatics sort du lot.

2.3 Sources de variabilité d'un système de RAP

Au commencement du doctorat de Sébastien Ferreira (§6.2.2 page 98), nous nous sommes interrogés, avec Authôt, sur les différentes sources de variabilité de la parole, qui auraient un impact important sur la reconnaissance de la parole. Ceci afin d'orienter les priorités sur les problèmes à traiter lors de son travail de thèse. La figure 2.5 détaille les différentes sources recensées.

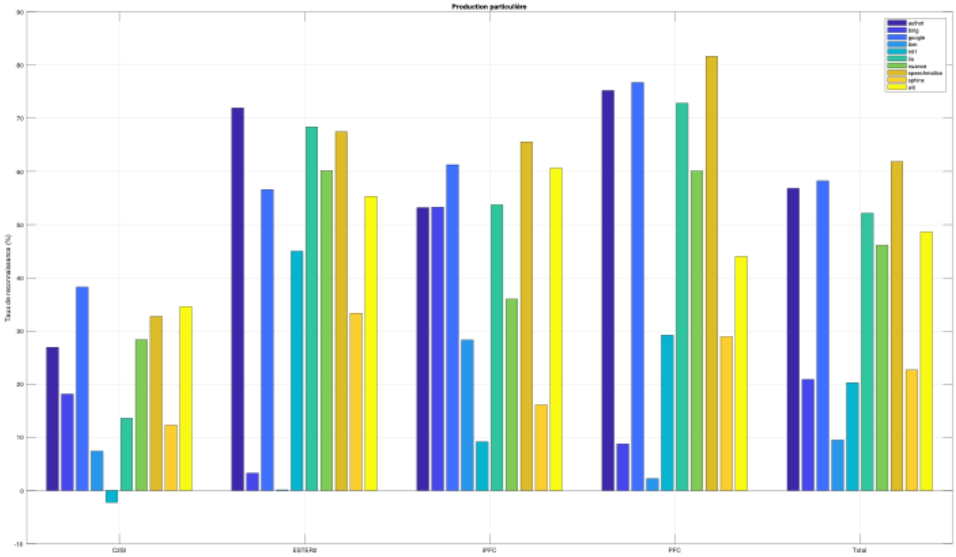


FIGURE 2.4 – Résultats de l'analyse de système ASR : parole « atypique »

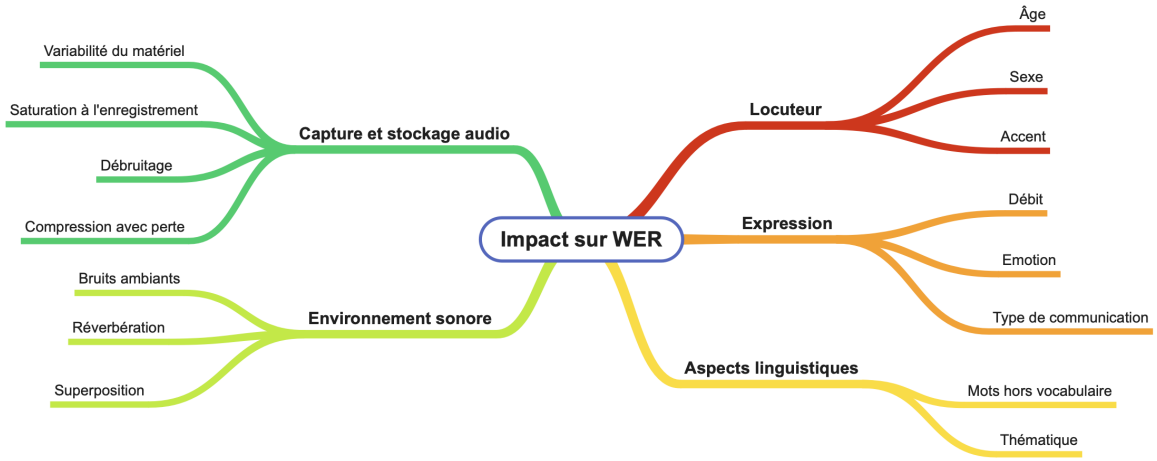


FIGURE 2.5 – Sources de variabilité du taux d'erreur sur les mots d'un système de reconnaissance automatique de la parole

2.4 Modélisation de la prédiction *a priori*

La problématique de thèse de Sébastien Ferreira était de pouvoir prédire *a priori* les performances d'un système de reconnaissance. Le but était d'utiliser un traitement rapide sur l'enregistrement à traiter afin de pouvoir aiguiller au mieux les clients de la société sur les solutions de transcription (et en particulier vers la solution complètement automatique ou bien vers le traitement manuel). Ce verrou nous a finalement mis devant des défis assez intéressants concernant l'analyse de signaux : nous ne devions pas utiliser de grands systèmes grand vocabulaire, et cela nous a conduits à proposer des analyses très proches du signal à traiter. Sébastien a réussi à traiter trois problématiques qui représentaient une bonne proportion dans les fichiers de tous les jours à traiter par l'entreprise de transcription.

2.4.1 Environnement bruité

Le bruit, et nous l'avons vu en section 2.2.3 représente une difficulté pour les systèmes et constitue la majeure part des dégradations que les systèmes ont à traiter. Un simple calcul du SNR (rapport signal à bruit) ne donne pas forcément d'information très corrélée avec le WER (taux d'erreur sur les mots qui sera produit par un système de reconnaissance de la parole). Nous avons développé deux indicateurs : S-SF (extraction d'indicateurs suite à l'application d'un masque binaire sur le spectrogramme [Ferreira et al., 2018, Ferreira, 2021]) et S-SNR (calcul d'un indicateur basé sur un calcul du SNR par bandes [Ferreira et al., 2019a, Ferreira, 2021]) qui ont permis d'obtenir des prédictions très proches du WER des fichiers à traiter. L'évaluation de ces algorithmes a été effectuée sur le corpus Wall Street Journal [Garofolo et al., 1993], bruité avec NOISEX (afin de pouvoir se comparer avec la littérature), mais également avec des fichiers issus du système en production chez Authôt. À la fin de la thèse, Sébastien a implémenté ces algorithmes pour être mis en production dans la chaîne de traitement d'Authôt.

2.4.2 Réverbération

La réverbération, qui est en général contrôlée par des ingénieurs du son, est donc peu présente dans les contenus édités : c'est un phénomène qui apparaît de manière plus fréquente lors de captations réalisées par des particuliers. En effet, les caractéristiques du microphone, les dimensions de la pièce utilisée, peuvent grandement altérer le signal de parole. La réverbération, du point de vue du signal, apparaît comme un signal fantôme, un peu moins énergétique que le principal, qui se retrouve copié sur le signal principal, avec un petit décalage avec l'original. Cette superposition peut induire des confusions du système de reconnaissance de la parole (voir section 5.2.1 de [Ferreira, 2021] pour la mise en évidence du problème).

Nous avons proposé un paramètre appelé « Excitation Behaviour » afin de représenter le degré de réverbération d'un signal de parole. Il est basé sur l'analyse des résidus de la prédiction linéaire sur les zones voisées du signal. Les résultats sont probants, car mieux corrélés que les indicateurs existants dans l'état de l'art [Ferreira et al., 2020b, Ferreira et al., 2019b, Ferreira, 2021].

2.4.3 Parole superposée

Nous allons considérer, pour la suite, que la parole peut être superposée à une autre parole ou bien à de la musique. Nous avons alors deux cas de figure :

- la parole superposée à de la musique. Par exemple, dans le domaine des médias, une musique d'ambiance est souvent ajoutée/mixée (films documentaires, reportages),
- la parole superposée à de la parole. Par exemple, lorsque des locuteurs se coupent la parole (superposition de courte durée), ou lors d'un bavardage durant une réunion (longue superposition), ou enfin, le « voice-over » [3]

Nous avons proposé d'extraire plusieurs indicateurs pour caractériser la parole superposée :

- la modulation de l'entropie (la modulation de la parole est plus importante de celle de la musique),
- une mesure issue du suivi de partiels (maxima locaux) dans le spectre,
- une mesure de densité de segments,
- un indicateur issu de l'intersection de partiels (lorsque le signal de parole est superposé avec un autre signal, les segments détectés sont souvent très proches, voire se coupent : il y a une

3. pour faire un doublage de qualité, il faut faire attention à ce que les phonèmes prononcés par le locuteur original et ceux du doublage soient cohérents (synchronisés). Ainsi le procédé de doublage est long et coûteux. Certaines productions optent pour le « voice-over » qui consiste à laisser le locuteur originel parler avec un volume moins important tout en ajoutant la traduction du texte prononcé, la parole est ajoutée/mixée en post-production.

intersection des segments ; pour chaque pic, appartenant à un segment, nous gardons dans la représentation les valeurs qui correspondent à une fenêtre Hamming centrée sur le segment),
— des mesures liées à la fréquence fondamentale (ratio de voisement, multimodalité de distribution de f_0, harmonicité),
— Excitation Behaviour (proposé ci-dessus pour la réverbération)

Ces paramètres sont donc issus de méthodes connues et d'autres ont été créés spécialement pour cette tâche). Tous les paramètres peuvent être extraits *a priori*. Les résultats de classification sont plutôt bons [Ferreira et al., 2020a,Ferreira, 2021], sachant que différents niveaux de SNR pour la superposition ont été utilisés. L'identification du type de superposition obtient d'excellents résultats, mais l'identification du SNR utilisé pour la superposition reste plus complexe.

2.5 Conclusion

Dans ce chapitre, j'ai mis en évidence la diversité des signaux qu'un système de reconnaissance automatique de parole a à traiter. J'ai pu montrer les conséquences sur les performances des systèmes. En cherchant à prédire *a priori* le taux de reconnaissance des systèmes, nous avons réussi à mettre en place des indicateurs, assez rapidement calculables, qui peuvent caractériser les passages des enregistrements audio à traiter. Cela permettrait d'enclencher le choix de certains modèles plus performants sur les caractéristiques détectées. Cela contribue également à fournir des informations pour des modèles de plus haut niveau (par exemple la segmentation en émissions pour un enregistrement radiophonique ou télévisuel), qui pourraient alors bénéficier de ces informations.

FIGURE 2.6 – La recherche en informatique à la manière de Joan Miró (généré par Midjourney en février 2023, moteur de rendu n°4)

Chapitre 3

La parole : à la recherche de l'intelligibilité

Sommaire

3.1 Introduction

En 2012, suite à une mise en relation de la Région avec la société Archean Technologies, j'ai été amené à travailler sur un verrou qui m'a mené sur une nouvelle thématique : la mesure automatique de l'intelligibilité. L'entrepreneur souhaitait que l'on puisse mesurer automatiquement l'intelligibilité et la compréhensibilité de la parole par ordinateur, afin de pouvoir procéder à des réglages plus rapides des prothèses auditives pour des personnes atteintes de presbyacousie.

Avec l'équipe de Pascal Gaillard, travaillant sur la perception humaine de l'acoustique à l'Université Jean Jaurès, nous avons dû mettre en place un protocole scientifique pour pouvoir avancer sur ce problème. Restait à concevoir la partie mesure avec l'ordinateur. Cela a donné lieu au projet « Équipement électronique intelligent de mesure de compréhension de la parole » et la solution trouvée sera expliquée en section 3.4.

Peu après, suite à des présentations scientifiques présentées lors de rencontres sur le thème du traitement de la parole dans la région toulousaine, nous avons été mis en contact avec Virginie Woisard

(médecin oto-rhino-laryngologiste, phoniatre, responsable de l'Unité de la Voix et de la Déglutition dans le service ORL du Pr Serrano du CHU de Rangueil-Larrey à Toulouse), dont l'objectif était d'arriver à améliorer la qualité de vie des patients. L'amélioration de la condition des patients pouvait passer par une automatisation de la mesure de l'intelligibilité. En effet, il est difficile pour un médecin d'arriver à porter un jugement objectif sur l'état de la qualité de la voix d'un patient, car à force de proximité, l'être humain s'habitue aux altérations de la voix et finit par deviner ce qui est prononcé. Les altérations étant très nombreuses dans le cas d'atteinte par des cancers de la sphère ORL, l'intelligibilité de ces patients est vraiment très basse, et cela peut conduire à un très gros problème pour leur communication en société, et au final dégrader sensiblement leur qualité de vie. Je vais détailler les différentes approches que j'ai suivies dans la section 3.5.

Mais pour pouvoir alimenter ces recherches sur la mesure de l'intelligibilité, et expérimenter pour valider les résultats, il est indispensable de pouvoir disposer de corpus constitué d'enregistrements et surtout d'annotations, d'un référentiel quantifié. Dans le cas de la production de la parole, ce référentiel devra provenir de plusieurs évaluations de spécialistes, afin de ne pas être biaisé dans la mesure. Je vais détailler le matériel qui a été utilisé tout au long de ces études dans la section 3.3.

Mais avant de commencer, il est une étape indispensable qu'il est nécessaire de traiter : la définition des termes d'intelligibilité et de compréhensibilité. En effet, on retrouve dans la littérature diverses acceptions, parfois contradictoires [Pommée et al., 2022]. Nous y avons longuement réfléchi au cours de ces recherches, et le point final placé suite à une étude publiée par Timothy Pommée (doctorant du projet TAPAS), cherchant un consensus au niveau des spécialistes de la voix. La section suivante résume ses travaux et détaille les définitions qui seront utilisées par la suite.

3.2 Définitions

Pour parvenir au consensus, Timothy Pommée a mis en place une méthodologie Delphi [Jones and Hunter, 1995], qui est composée de plusieurs phases d'enquêtes afin de permettre de converger vers la solution qui fait l'objet d'un accord commun [Pommée et al., 2022].

Voici les définitions qui ont résulté de cette étude (traduction de l'anglais) :

intelligibilité L'intelligibilité se réfère à la reconstruction d'un énoncé au niveau acoustico-phonétique, l'information liée à l'intelligibilité est donc véhiculée par le signal acoustique (c'est-à-dire que l'intelligibilité se concentre sur l'information dépendante du signal). Cette reconstruction est rendue possible à la fois par la capacité de production phonétique-acoustique du locuteur et par les compétences de décodage acoustico-phonétique de l'auditeur. D'un point de vue perceptif, l'intelligibilité est mieux analysée sur des stimuli peu prévisibles : phonèmes, syllabes, pseudo-mots, mais aussi mots (en paires minimales) et phrases imprévisibles pour une évaluation plus fonctionnelle prenant en compte la coarticulation et les symptômes au niveau de la phrase (par exemple, la respiration et la prosodie), tant que les processus de compensation cognitive descendante de l'auditeur sont évités (c'est-à-dire sans aide du contexte sémantique ou linguistique). Objectivement, l'intelligibilité peut être évaluée à l'aide de l'acoustique des consonnes, des voyelles et des semi-voyelles (y compris les transitions de formants entre phonèmes), qu'il s'agisse de phonèmes isolés ou intégrés dans des syllabes, dans des (pseudo-)mots ou dans des phrases. En outre, dans certains cas, la qualité de la voix contribue également à l'intelligibilité, car elle joue un rôle dans certains contrastes phonémiques. Les paramètres suprasegmentaux (par exemple, évalués objectivement par le débit de parole ou l'accentuation) contribuent également à l'intelligibilité.

compréhensibilité La compréhensibilité se réfère à la reconstruction d'un message au niveau sémantique-discursif, après la reconstruction acoustico-phonétique. L'intelligibilité est donc une

composante de la compréhensibilité. Outre le décodage acoustico-phonétique, elle comprend également des éléments contextuels indépendants du signal, tels que le contexte linguistique ou non verbal. Cependant, on peut être compréhensible sans que toutes les unités de bas niveau soient nécessairement décodées avec précision ; par conséquent, si l'intelligibilité influe sur la compréhensibilité, cette dernière n'en dépend toutefois pas entièrement. La compréhensibilité renvoie à la dimension plus fonctionnelle de la communication et est mieux évaluée sur le plan perceptif à l'aide d'évaluations liées au sens (c'est-à-dire en tenant compte des processus cognitifs descendants susceptibles de compenser les informations acoustiques et phonétiques dégradées). Aujourd'hui, aucune mesure instrumentale objective n'est encore adaptée pour évaluer la compréhensibilité en tant que telle (c'est-à-dire la transmission du sens global du message). Toutefois, certains paramètres suprasegmentaux contribuent à l'intelligibilité et peuvent être évalués objectivement (par exemple, les mesures du rythme et de l'intonation).

J'utilise par la suite cette terminologie dans ce document. On notera que la compréhensibilité est un objectif difficilement atteignable pour le moment à l'aide de mesures automatiques, bien que certaines composantes prosodiques sous-jacentes soient atteignables.

3.3 Ressources

Au cours des différents projets qui ont financé les recherches sur la mesure de l'intelligibilité, nous avons pu constituer des corpus pour alimenter la production des systèmes de reconnaissance et l'évaluation des performances.

3.3.1 Corpus Archean

Ce corpus a été créé pendant le projet AGILE IT « Équipement électronique intelligent de mesure de compréhension de la parole » (cf. §5.4.9 page 79) et utilisé au cours du projet CLE PHONICS (cf. §5.4.10 page 79).

L'objectif était de créer un corpus permettant l'évaluation de la compréhensibilité de patients en simulant des dégradations liées à la presbyacousie. Ceci a été fait de façon à obtenir un référentiel des performances perceptives et cognitives humaines.

Le corpus est destiné à évaluer l'intelligibilité et la compréhensibilité, et est constitué de :

1. mots : test d'intelligibilité, très utilisé par les audioprothésistes, consiste pour l'auditeur à répéter des mots dissyllabiques. Le matériel sonore du test a été sélectionné à partir d'un sous-ensemble des listes de Fournier [Fournier, 1951]. Le corpus est constitué de 10 mots d'entraînement et 60 mots de tests.
2. phrases : test d'intelligibilité. Les phrases utilisées sont issues de la version française du test HINT (pour Hearing in Noise Test, développé initialement en anglais par Nilsson et collègues [Nilsson et al., 1994], et adapté au français [Vaillancourt et al., 2005]). Ce test est également très utilisé par les audioprothésistes pour évaluer l'intelligibilité de la parole en milieu bruité. Le corpus est constitué de 10 phrases d'entraînement et de 60 phrases de test.
3. compréhension de phrases (idem test 2). Après avoir écouté la phrase, le participant doit choisir une image parmi quatre proposées. Une image et trois alternatives issues des ambiguïtés obtenues dans le test précédent.
4. exécution de commandes verbales : test de compréhension (cf. [Fontan, 2012,Fontan et al., 2013]) consistant à demander aux auditeurs de répondre à des consignes verbales leur demandant de déplacer des images sur un écran d'ordinateur (ex. : « Mettez <objet 1> <position> <objet 2> »).

Les phrases du test ont été élaborées afin que la proportion de déplacements à gauche/droite et au-dessus/au-dessous soit la plus équilibrée possible, et que chaque objet cible n'apparaisse qu'une fois dans le test. Au total, 10 phrases d'entraînement (ex : « Mettez l'hélicoptère au-dessus de la girafe ») et 30 phrases de test ont été définies. Pour chaque phrase, les deux images cibles sont accompagnées de quatre autres images pouvant induire l'auditeur en erreur.

Ce matériel oral a été enregistré par trois locuteurs : une femme, un homme et un enfant. Les enregistrements ont été réalisés dans la cabine audiométrique PETRA [4], avec un microphone omnidirectionnel Sennheiser MD46, une console de mixage TASCAM DM 3200 et un ordinateur Mac Pro équipé du logiciel Reaper. Les phrases ont ensuite été égalisées en sonie par un panel d'auditeurs puis mixées avec un bruit de fond vocal de type "brouhaha" et un rapport signal sur bruit de 5 dBA. Les phrases avec et sans bruit de fond ont enfin été dégradées par simulation de la presbyacousie suivant une variation en dix niveaux. Le premier niveau correspond au signal non dégradé et les neuf niveaux suivants vont d'un âge théorique de 60 ans à 110 ans. La procédure de traitement de signal permet de simuler les différentes atteintes de la presbyacousie [Moore, 2007] : l'augmentation du seuil d'audibilité, la réduction de la sélectivité fréquentielle et le recrutement de sonie (cf. [Fontan et al., 2014] pour plus de détails).

L'évaluation perceptive a été conduite avec trente participants. Le profil des participants répondait aux critères d'inclusion suivants : étudiants francophones natifs, âgés de 18 à 30 ans inclus, sans problème de vue non corrigé par des lentilles ou des lunettes. Le niveau d'audition de chaque participant a été vérifié par un audiogramme tonal (critère d'inclusion : moyenne des pertes entre 2 kHz et 8 kHz inférieur à 15 dB).

3.3.2 Corpus C2SI

Ce corpus a été constitué lors du projet C2SI (cf. §5.4.12 page 80) porté par Virginie Woisard et financé par l'Institut National du Cancer (INCa).

L'objectif de ce projet était de créer une base de données de voix françaises de personnes témoins et de patients atteints de cancers ORL et de valider des indices de mesure de sévérité et d'intelligibilité afin d'obtenir une évaluation objective de la qualité de vie des personnes atteintes de ce type de troubles de la parole.

Le corpus est composé d'un total de 113 locuteurs, dont 87 patients atteints de cancers et 26 témoins. Plusieurs critères ont été vérifiés avant d'inclure les patients, à savoir que les patients devaient avoir reçu un traitement par chimiothérapie et/ou radiothérapie et/ou chirurgie depuis plus de 6 mois. Cette durée minimale permet alors d'être sûr que la dégradation au niveau de la parole est stable et que ces troubles impactent la vie quotidienne des patients. Les patients présentant d'autres troubles connus de la parole indépendamment de leur pathologie n'ont pas été inclus. De même pour les personnes atteintes de troubles ne leur permettant pas de réaliser certaines tâches comme par exemple des troubles visuels qui ne permettraient pas de lire un texte convenablement.

L'âge moyen des locuteurs contrôles est de 55,7 ans ($\pm$ 12,8) tandis que celui des patients est de 65,8 ans ($\pm$ 9,5). L'âge entre ces deux groupes est significativement différent (p-valeur du test de Mann-Whitney inférieure à 0,02). Ceci est observable sur la figure 3.1 qui nous montre que le groupe contrôle est composé de deux groupes d'âges différents, un premier autour de 40 ans et un second autour de 60 ans.

Au niveau du sexe des personnes enregistrées, le corpus est composé de 46% de femmes au total. Cependant, la proportion hommes/femmes dans le groupe cancer n'est pas équivalente à celle dans le

4. http://petra.univ.tlse2.fr

FIGURE 3.1 – Répartition de l'âge et du sexe des locuteurs dans le corpus en fonction du groupe (contrôle ou cancer). À gauche, la distribution de l'âge des locuteurs est affichée en fonction du groupe de patients. À droite, la répartition hommes/femmes est également affichée en fonction du groupe des locuteurs.

groupe contrôle avec 41% de femmes dans le groupe cancer contre 62% dans le groupe contrôle. Cela est illustré dans la partie droite de la figure 3.1.

Les enregistrements ont été effectués au Centre Hospitalier Universitaire de Toulouse au sein d'une chambre anéchoïque. Un microphone Neumann TLM 102 a été utilisé et les fichiers sont échantillonnés à 48 000 Hz. Une dizaine de tâches comme de la lecture de texte ou une description d'image ont été réalisées par les patients, mais tous n'ont cependant pas réalisé l'ensemble des tâches. Une description plus détaillée de ce corpus, avec la description de chaque tâche ainsi que des informations médicales sur les locuteurs est disponible dans l'étude [Woisard et al., 2021] (disponible également ici : §A.3 page 144).

3.3.3 Corpus RUGBI

Ce corpus a été élabore au cours du projet RUGBI (cf. §5.4.15 page 83). Il est constitué de l'intégralité du corpus C2SI pour les patients atteins de cancer de la tête et du cou. A cet ensemble il a été procédé à l'ajout d'une vingtaine d'enregistrements de patients qui avait déjà procédé à un enregistrement dans C2SI. Ces ajouts permettent d'envisager une étude longitudinale et d'évaluer l'évolution de la parole d'un patient au cours du temps.

En plus du corpus cancer, un corpus de parole parkinsonienne a été intégré au projet ANR RUGBI. Ce corpus a été constitué dans le Service de Neurologie du Centre Hospitalier du Pays d'Aix à Aix-en-Provence et provient de la base de données AHN [Ghio et al., 2012]. Il contient au total 316 locuteurs, 205 patients parkinsoniens et 111 sujets témoins. La répartition en âge est davantage équilibrée entre les deux groupes avec une moyenne d'âge de 66,5 ans ($\pm$=9,6) pour les patients parkinsoniens et de 62,7 ans pour les témoins ($\pm$=11,0). La différence des moyennes est significative entre les deux groupes, mais seulement 4 ans les différencient ce qui ne devrait pas avoir un fort impact en terme de vieillissement de la voix et notamment au niveau prosodique [Hixon et al., 2008]. La répartition de l'âge des patients est illustrée sur la partie gauche de la figure 3.2.

La répartition hommes/femmes est assez inégale entre les groupes avec une majorité d'hommes chez les personnes atteintes de la maladie de Parkinson (67% d'hommes). Ce déséquilibre peut être justifié par le fait que la maladie de Parkinson touche en moyenne 1,5 fois plus les hommes que les femmes d'après sur les données nationales de surveillance de la fréquence de la MDP entre 2010 et

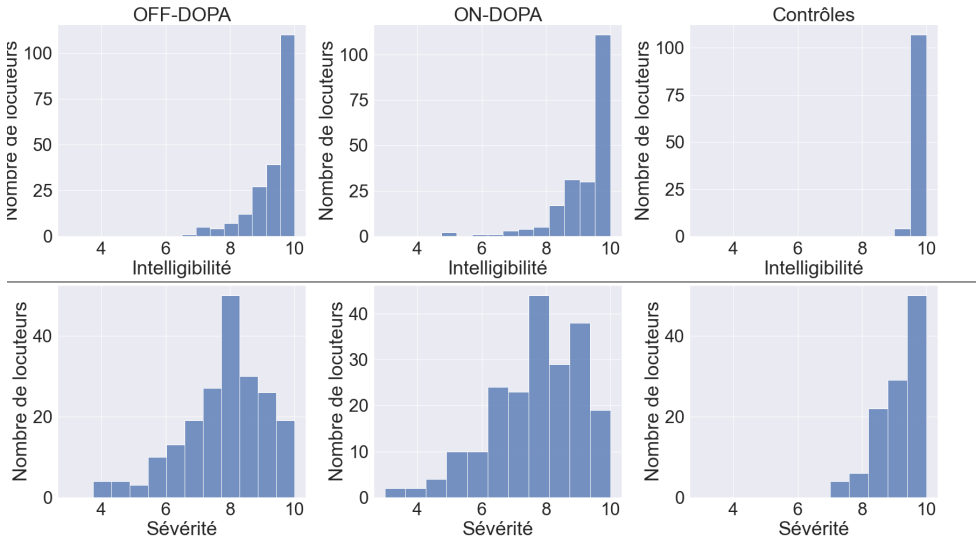


FIGURE 3.2 – Répartition des scores d'intelligibilité et de sévérité. En haut, la distribution du score d'intelligibilité et en bas, la distribution du score de sévérité. A gauche, en condition sans médication, au milieu avec médication (dopamine), et à droite le groupe contrôle.

2015 [Moisan et al., 2018]. En revanche, chez les patients contrôles, les hommes ne représentent que 41% des personnes. Ce phénomène pourrait être expliqué par le choix des participants à l'étude, car les conjoint(e)s des patients sont davantage susceptibles d'accepter de participer à ce genre d'études. La répartition du sexe des locuteurs est illustrée dans la partie droite de la figure 3.2.

L'évaluation perceptive de tous ces corpus a été réalisée par un groupe de six experts. Un score de sévérité (mesurant l'atteinte de la maladie) et un autre d'intelligibilité ont été produits sur la tâche de lecture (passage de la chèvre de M. Seguin). Quelques tests ont été également réalisés sur la tâche de description d'image, afin de pouvoir juger à partir d'un texte plus spontané, mais il n'y a pas eu de différence significative avec la même évaluation sur la parole lue. Pour évaluer l'intelligibilité, il a été demandé d'évaluer quatre critères qui sont des mesures d'altération de : la résonance, la prosodie, la voix, la prononciation phonémique. Ces critères sont échelonnés de 0 (pas d'altération) à 3 (grande altération). La somme de ces sous-scores constitue le score d'intelligibilité. Le score final est constitué de la moyenne des scores des six experts. Vous trouverez plus de précision sur la constitution du corpus et des évaluations perceptives dans [Woisard et al., 2022, Woisard et al., 2021].

Le sous-corpus correspondant à l'étude longitudinale a été livré en fin de projet, en 2023, ils n'ont pas encore été exploités par des études scientifiques.

3.3.4 GIS Parolothèque

Le Groupement d'Intérêt Scientifique PAROLOTHEQUE, est une structure dont la vocation est de permettre de développer la recherche scientifique sur des données cliniques d'enregistrements vocaux. Nous avons entrepris de mettre en place cette structure depuis 2012. Finalement, le document constitutif signé le 18 mai 2021 entre les établissements suivants : CNRS, l'Université Toulouse 3, l'Institut National Polytechnique de Toulouse, l'Université Toulouse 1, l'Université Toulouse 2, le Centre Hospitalier Universitaire de Toulouse, l'Institut Claudius Regaud, l'Université d'Avignon et l'Université d'Aix-Marseille. Les laboratoires constitutifs étant : l'IRIT, l'IFERISS, OCTOGONE-LORDAT, le CLLE, le LIA et le LPL. J'ai effectué plusieurs présentations afin de faire connaître cette structure [Farinas, 2017, Farinas, 2021, Ghio and Farinas, 2021, Farinas, 2021].

Par analogie aux tumorothèques, une parolothèque est une banque d'échantillons de parole enregistrés, obtenus à partir de bilans de trouble de la parole ou du langage ou à partir d'entretiens ou d'interviews de personnes concernées par les pathologies tumorales.

Bien que les enregistrements seront orientés par la recherche initiale (ou originelle), une automatisation de la transcription « sophistiquée » comprenant si besoin le repérage des différents locuteurs, la segmentation à partir d'une catégorisation sémantique rendra réalisable la réutilisation de la série d'interviews. Ces échantillons seront conservés dans un format numérique sur un serveur dédié au stockage d'enregistrements sonores et recensés dans une Base de données enrichies d'informations permettant leur l'archivage puis leur exploitation sur requête.

Ces données s'étendent du code de surface aux aspects contextuels, voire conversationnels, et peuvent faire l'objet d'une lecture psychologique, voir épidémiologique ou sociologique. La parolothèque sera ainsi alimentée par des recherches en SHS. Elle alimentera elle-même une recherche dans le domaine du traitement automatique des langues et fournira un service facilitant la recherche pour d'autres disciplines.

L'objectif principal de ce projet est d'établir un corpus de voix et de parole de patients atteints de cancer. Ce corpus pourra ensuite être analysé et étudié autant du point de vue de son contenant (parole) que de son contenu (discours). Il permettra de développer des axes de recherches variés et l'élaboration de publications scientifiques dans divers domaines (sciences humaines et sociales : SHS, épidémiologie, linguistique, informatique…).

Les données des projets C2SI, RUGBI, DAPADAF-E constituent dès à présent de corpus disponible dans le GIS PAROLOTHEQUE. Ces données vont donc pouvoir être exploitées au-delà des projets qui ont permis leur collecte et permettront à la collectivité scientifique de pouvoir prolonger les recherches qui ont déjà été menées.

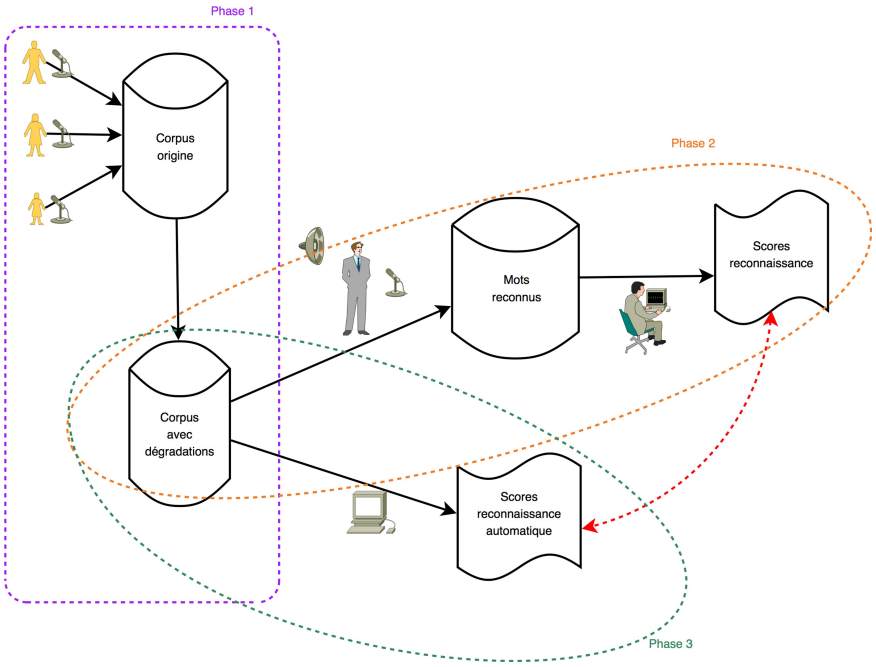


FIGURE 3.3 – Architecture de la mesure d'intelligibilité dans le projet Archean

3.4 Du point de vue de la perception de la parole

L'hypothèse principale de la production de mesure d'intelligibilité dans le cadre du projet Archean (cf. §5.4.9 page 79) est qu'un système de reconnaissance automatique de la parole grand vocabulaire va pouvoir produire des résultats dont on pourra trouver une bonne corrélation avec les résultats produits par les humains. Ce n'était *a priori* pas évident, car les facultés de transcription des humains sont meilleurs que les systèmes [Meyer et al., 2010].

La chaîne de traitement décrite sur la figure 3.3 présente :

— la chaîne de traitement automatique (en haut) : le corpus, avec ses différents niveaux de dégradation, est évalué par un système de reconnaissance de la parole généraliste ;

— l'évaluation perceptive (en bas) : ces mêmes fichiers sont évalués par plusieurs personnes. Des scores de distance entre la proposition de la personne et le mot attendu est calculé en utilisant une distance de Levenshtein [Levenshtein et al., 1966] qui prend en compte la proximité linguistique entre l'énoncé reconnu et attendu en comptabilisant les traits linguistiques en commun.

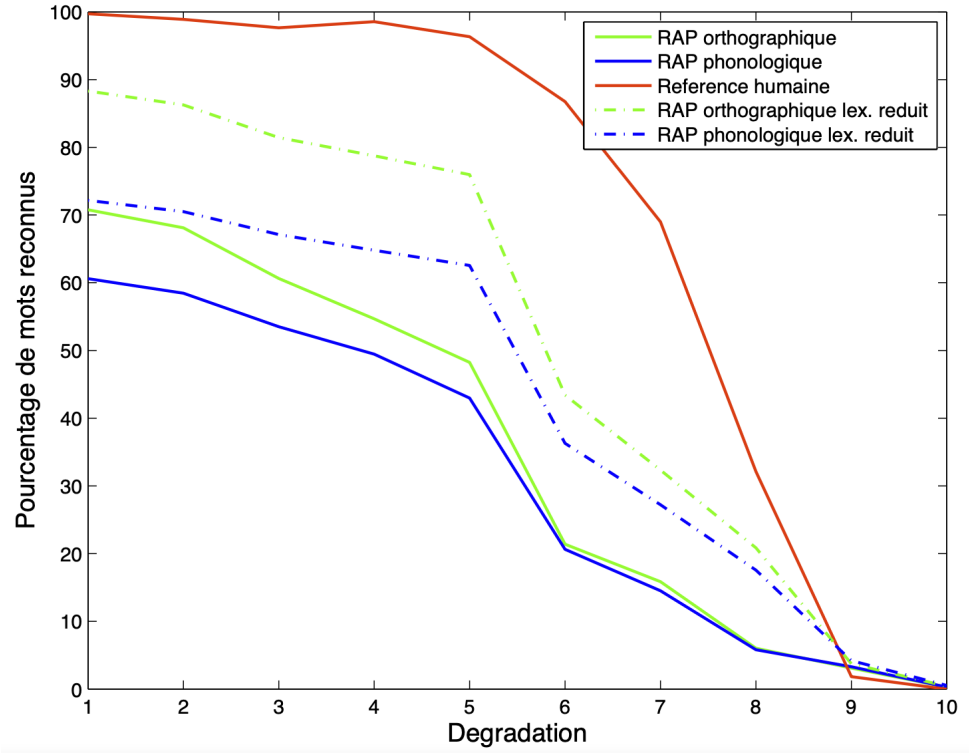


FIGURE 3.4 – Résultats des scores automatiques et humains dans le projet Archean

La comparaison entre les scores automatiques et des jugements humains se trouve sur la figure 3.4. Une régression polynomiale d'ordre 5 permet de lier les résultats, et obtenir un coefficient R^2 de 0,997. Vous trouverez plus de détails dans les papiers détaillant les expérimentations [Fontan et al., 2015b, Fontan et al., 2017].

3.5 Du point de vue de la production de la parole

La première vague de travaux en traitement automatique a consisté à tester d'extraire de nombreux paramètres du signal de parole afin d'essayer de faire des corrélations avec les scores de sévérité et

d'intelligibilité. Il a ensuite été essayé de tenter une approche par modèles non linéaires. Puis nous avons travaillé avec Mathieu sur son doctorat pour modéliser l'impact sur la communication. Et finalement on a essayé avec Vincent de proposer des solutions de modélisations automatiques utilisant des réseaux de neurones, mais en employant des techniques d'usage parcimonieux des données. Au cours du projet RUGBI, j'ai également traité la mesure d'intelligibilité sous l'angle de la prosodie. Je reviendrais sur cet aspect dans la prochaine partie (cf. chapitre 1).

3.5.1 Recherche de corrélation et calcul de régressions

Lors du stage de master de Brendan Gloinec (cf. section 10 page 93) nous avons voulu essayer de caractériser sévérité et intelligibilité sur un des premiers lots du corpus C2SI.

La modélisation de parole pathologique, en 2016, avait fait l'objet de la réalisation de challenges au niveau du congrès international Interspeech : « computational paralinguistics challenge : cognitive & physical load » organisé par Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi et Yue Zhang [Schuller et al., 2014]. Le challenge mettait à disposition un outil qui permettait d'extraire de très nombreux paramètres sur les fichiers analysés grâce à la boîte à outils OpenSmile (open-source Speech and Music Interpretation by Large-space Extraction) [Eyben et al., 2010]. Cet outil est capable de produire de très nombreux paramètres (plus de 1000 par fichier audio traité), et d'y associer différentes mesures statistiques pour les caractériser (ex : calcul de la moyenne sur le fichier traité, calcul du skewness, du kurtosis…).

Il suffit ensuite d'effectuer des calculs de corrélation avec ces paramètres, de manière à faire ressortir les plus intéressants avec les scores de sévérité et d'intelligibilité.

Nous souhaitions éprouver les paramètres tels que le skewness et le kurtosis de la fréquence fondamentale, ainsi que le calcul de HNR (Harmonic to Noise Ratio). Ces paramètres étaient connus pour fournir des informations sur la voix [Teixeira et al., 2013].

Il a obtenu des résultats de corrélation de 0,51 avec des paramètres acoustiques (basés sur la moyenne des MFCC) et de 0,70 a 0,71 avec des paramètres basés sur le flux spectral et les paramètres acoustiques. Notons que des paramètres dérivés du jitter ont également obtenu des corrélations assez proches. La différence au niveau des résultats sur la sévérité ou bien l'intelligibilité est plutôt logique : un système automatique basé sur des paramètres simples va obtenir de meilleures corrélations avec de score de sévérité, car ce score est plus proche des altérations que l'on retrouve dans le signal, alors que celui d'intelligibilité peut faire intervenir des processus de plus haut niveau. Notons au final que les paramètres qui représentent l'acoustique du signal semblent jouer un grand rôle pour ces scores.

Mathieu Balaguer (cf. section 6.2.3 page 99) a effectué un stage pour son Master 2 Recherche mention santé publique spécialité épidémiologie clinique en 2018. Son objectif était d'évaluer la validité des différents scores de mesure des troubles de la parole issus d'une analyse automatique du signal. Il a pu travailler avec une base de données contenant plus de fichiers enregistrés que ce dont disposait Brendan, et a travaillé sur plusieurs tâches de C2SI alors que Brendan n'avait utilisé que les A tenus. Il a calculé la corrélation entre le score de sévérité et celui d'intelligibilité et obtenu une corrélation de 0,91 entre les deux. Il a obtenu une corrélation avec la sévérité de 0,75 en amalgamant les paramètres suivants :

- Écart interquartile de la fréquence fondamentale,
- Instabilité de la hauteur sur la tâche de A tenu,
- Score de vraisemblance automatique sur la tâche de décodage automatique de la parole,
- Score de vraisemblance automatique sur la tâche de lecture,
- Cumul des rangs sur la tâche de décodage automatique de la parole,
- Taux d'anomalies sur la tâche de décodage automatique de la parole.

3.5.2 Approche par modèles non linéaires

Oriol Pont (cf. section 6.3.6 page 105) a été recruté sur C2SI pour tenter de voir si les méthodes non linéaires pouvaient avoir un intérêt pour notre travail sur la parole pathologique.

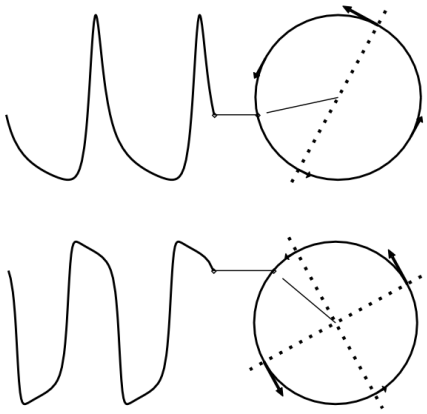


FIGURE 3.5 – Formes simples de non linéarité : en haut, signal résultant de la rupture de la symétrie de rotation. Il apparaît comme projection d'une vitesse de rotation non uniforme ; la vitesse de phase présente la symétrie de réflexion (par rapport à l'axe pointillé). En bas : forme de non linéarité plus simple impliquant deux axes de réflexion orthogonaux. Le long du cycle, deux régions de rotation rapide sont séparées par deux sections à rotation lente. Cela donne la forme carrée du signal (selon l'orientation de l'axe de symétrie)

Oriol a réalisé une implémentation des caractéristiques d'anharmonicité et de non linéarité pour les signaux de parole (quasi périodiques). Le signal de parole est représenté dans la base de Hanusse [Hanusse et al., 1994, Hanusse and Gomez-Gesteria, 1996, Hanusse, 2011] en le considérant comme une représentation multi-résolution issue de psin et pcos. Cela exploite la base trigonométrique de Hanusse [Hanusse et al., 1994]. Un signal périodique représente la projection d'un mouvement de rotation le long d'une trajectoire fermée sur une variété bidimensionnelle, qui est topologiquement équivalente à une rotation autour d'un cercle. Lorsque la vitesse de rotation est indépendante de la position ou de la phase, le signal observé est sinusoïdal, mais lorsque la vitesse devient dépendante de la phase, une anharmonicité apparaît. Voir la figure 3.5 pour une illustration de formes simple de non linéarité (extrait de [Hanusse, 2011]).

Il a calculé et implémenté sur des fenêtres glissantes différents paramètres :

— une « sparsité » synthétique (décrit les fluctuations d'amplitude, analogue à la représentation du « shimmer » dans la représentation de Fourier)
— le paramètre optimal de forme (représente le point de départ dans Fourier)
— une mesure de non linéarité (le gain de « sparsité » de Fourier)

La figure 3.6 montre les résultats de ces paramètres pour un témoin et un patient. Malheureusement nous n'avons pas obtenu de corrélations plus intéressantes que les autres travaux en cours.

3.5.3 Modélisation de l'altération de la communication

Mathieu Balaguer (cf. section 6.2.3 page 99) à soutenu sa thèse « Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé » en 2021. Il s'agit donc de trouver dans le signal de parole spontanée quels paramètres issus d'analyses automatiques permettent de prédire l'impact du trouble de parole sur les capacités

fonctionnelles de communication des patients (cf. figure 3.7 qui représente la modélisation jusqu'à la Qualité de Vie, inspiré de Wilson [Wilson and Cleary, 1995]). L'impact sur la communication n'est pas facile à déterminer, car aucun indicateur n'est validé en cancérologie ORL pour cette mesure. En ce qui concerne les mesures automatiques, les mesures actuelles concernent en général des paramètres bas niveau (cf. étude systématique faite en début de doctorat [Balaguer et al., 2020b]), mais l'on souhaitait intégrer des niveaux bien plus élevés.

Mathieu s'est inspiré du modèle psycholinguistique de Caron [Caron, 1989] pour produire des paramètres dans les trois grandes dimensions de ce modèle : articulatoire, catégorielle et conceptuelle.

Le niveau articulatoire concerne, dans sa composante de phonétique articulatoire, des mesures temporelles telles que la durée de parole ou le débit de parole et l'articulation, mais aussi l'étude de segments vocaliques et non vocaliques. Dans sa composante phonologique, les mesures phonémiques seront prises en compte, avec l'étude de la reconnaissance des phonèmes dans la parole des sujets (inventaire phonémique, classes phonétiques).

Le niveau catégoriel sera envisagé selon deux volets : le premier étudie le niveau lexical, via l'inventaire des mots reconnus par un système de reconnaissance automatique, et les caractéristiques associées (longueur et fréquence des mots, diversité et densité lexicale). Le second envisagera le niveau grammatical par l'étude des classes des mots reconnus.

Enfin le niveau conceptuel : étude des thématiques avec les sorties du système de reconnaissance au moyen d'une classification hiérarchique descendante, et étude du sentiment général.

Pour pouvoir travailler sur de la parole spontanée, Mathieu a été obligé de constituer un nouveau corpus auprès de 25 sujets traités pour un cancer de la cavité buccale ou de l'oropharynx. Il comprend une tâche de parole spontanée enregistrée au cours d'un entretien semi-dirigé, mais aussi des auto-questionnaires autorisant la mesure de la communication et des facteurs associés à la parole et à la communication. Concernant le premier aspect, un score de référence mesurant de façon holistique la communication a été construit. Il permet de combler le manque d'outils disponibles en cancérologie ORL pour cette mesure.

Cent quarante-neuf paramètres automatiques issus des différents niveaux du modèle psycholinguistique de communication de Caron ont été extraits. Puis, un processus de sélection a abouti à retenir 75 paramètres pertinents et non redondants. Enfin, pour le troisième aspect, une modélisation prédictive de l'altération de la communication a été menée au moyen des paramètres automatiques retenus (corrélation de 0,83 entre score prédit et score réel). La corrélation atteint même 0,89 en incluant à la modélisation des facteurs associés (constitution des cercles sociaux, état anxio-dépressif, déficits associés, auto-perception du handicap lié au trouble de parole). L'utilisation de l'analyse automatique de la parole permet donc une prédiction fiable de l'altération de communication ressentie par les patients [Balaguer, 2021].

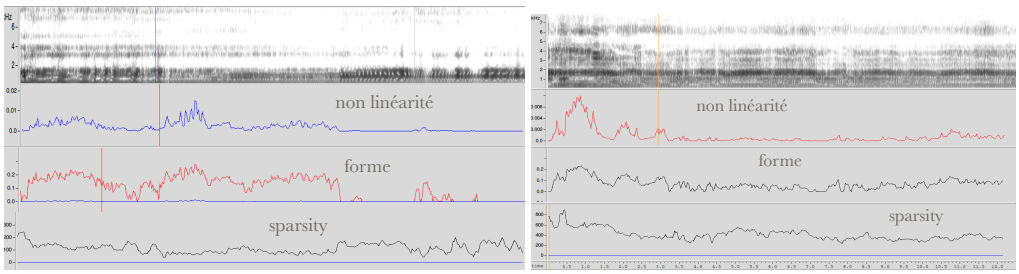


FIGURE 3.6 – Illustration du spectrogramme avec les représentations de la sparsité, du paramètre de forme et du paramètre de non linéarité sur un témoin (à gauche) et un patient (à droite)

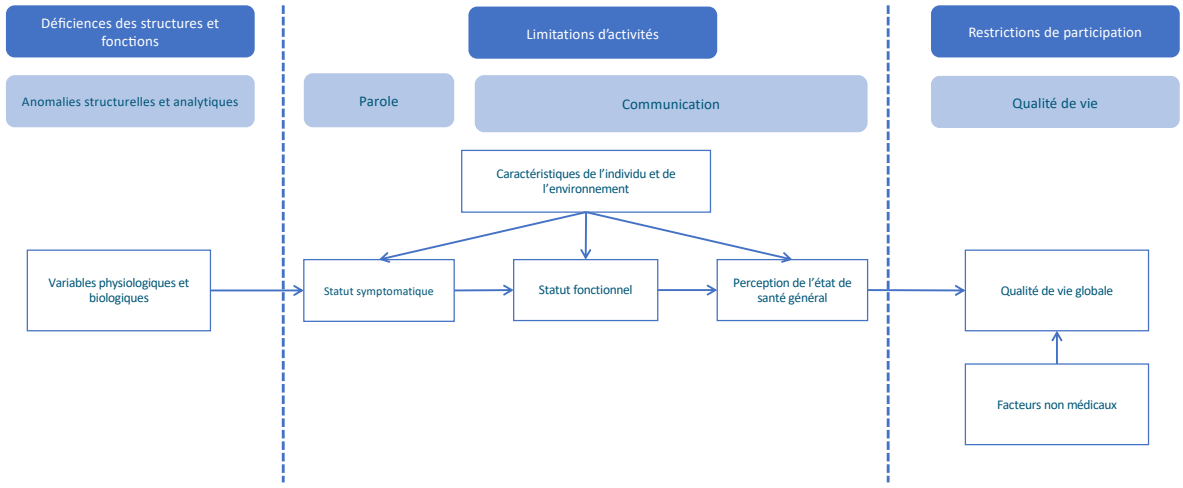


FIGURE 3.7 – Modèle de la Qualité de Vie de Mathieu Balaguer, inspiré de Wilson

3.5.4 Modélisation de la sévérité avec des méthodes parcimonieuses

Vincent Roger (cf. section 6.2.4 page 100 a effectué un doctorat sur « Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond : d'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique » qui a été soutenu en 2022 [Roger, 2022].

Il s'est attelé à proposer des apprentissages de modélisations avec des techniques modernes d'apprentissage par réseau de neurones, tout en étant confronté au problème de la parcimonie des données disponibles en parole pathologique. Sa problématique a donc été la suivante : est-ce que les techniques d'apprentissage peuvent, avec un corpus de données limité, modéliser le concept d'indice de sévérité du trouble de la parole tout en étant robustes ? Et, si oui, serait-il possible d'envisager une application médicale à ces travaux ?

La première tentative a été d'essayer de modéliser la sévérité à l'aide d'apprentissage par transfert et par apprentissage profond. Les résultats étant non utilisables, il a fallu se tourner sur les techniques dites « few shot » (apprentissage à partir de quelques exemples seulement). Ainsi, après de premiers essais prometteurs sur la reconnaissance de phonèmes, nous avons obtenu des résultats prometteurs pour catégoriser la sévérité des patients. Vous trouverez plus de détails dans cet article [Roger et al., 2022a].

Néanmoins, l'exploitation de ces résultats pour une application médicale demandait des améliorations. Nous avons donc réalisé des projections des données du corpus RUGBI (cf. §3.3.3). Comme certaines tranches de scores étaient séparables à l'aide de paramètres acoustiques, nous avons proposé une nouvelle méthode de mesure entropique. Celle-ci est fondée sur des représentations de la parole auto-apprise sur le corpus Librispeech [Panayotov et al., 2015] : le modèle PASE+ [Ravanelli et al., 2020], qui est inspiré de l'Inception Score (généralement utilisé en image pour évaluer la qualité des images générées par les modèles). Notre méthode nous permet de produire un score semblable à l'indice de sévérité avec une corrélation de Spearman de 0,87 sur la tâche de lecture du corpus cancer. L'avantage de notre approche est qu'elle ne nécessite pas des données du corpus cancer pour l'apprentissage. Ainsi, nous pouvons utiliser l'entièreté du corpus pour l'évaluation de notre système. Vous trouverez plus de détails dans l'article suivant [Roger et al., 2022c].

La qualité de nos résultats nous a permis d'envisager une utilisation en milieu clinique à travers une application sur tablette : des tests sont d'ailleurs en cours à l'hôpital Larrey de Toulouse (cf. section 5 page 76).

3.6 Conclusion

Cette thématique de recherche, autour de l'intelligibilité, a été alimentée par de nombreux projets : AGILE IT Archean, INCa C2SI, ANR RUGBI, H2020 TAPAS, TTT SAMI, UT3 BQR, PHRIP DAPADAF-E, AADI, ICC ADAPT. Cela a permis de créer une réelle dynamique dans l'équipe de recherche SAMOVA, mais également certains laboratoires (en particulier le LNPL/UT2J, le LIA/UA et le LPL/AMU) autour de la mesure d'intelligibilité et de la prise en compte de la qualité de vie des patients. Ce n'est pas terminé, nous avons encore des projets dans les cartons pour continuer ces collaborations, nous pouvons aller encore plus loin sur le sujet. En effet, nous pourrions envisager de pouvoir modifier le signal de parole afin de la rendre plus intelligible, en utilisant des aspects acoustiques, mais également prosodiques. Cela permettrait aux patients de disposer d'outils afin de mieux communiquer et ainsi gagner en qualité de vie.

FIGURE 3.8 – Le traitement automatique de la parole à la manière de Henri de Toulouse-Lautrec (généré par Midjourney, en février 2023, moteur de rendu n°4)

Deuxième partie

Projet de recherche

Introduction

Après avoir passé en revue dans la première partie les techniques que j'ai abordées dans mes travaux de recherche, mais également l'état de l'art actuel de ces techniques, je vais maintenant détailler mes prochains objectifs de recherche.

La technologie de reconnaissance automatique de la parole est maintenant arrivée dans une phase de production. Les défis restent encore nombreux. Je pense que l'on peut avec le renouveau des modélisations actuelles par réseaux neuronaux arriver à capturer des informations de plus en plus haut niveau. Et je reste persuadé que l'inclusion de la dimension prosodique aux systèmes de reconnaissance de la parole permettra de faire un pas vers un idéal de compréhension exhaustif de l'acte de communication. Mon incursion dans la modélisation de la prosodie appliquée à la parole pathologique m'a permis d'entrouvrir la porte sur un domaine qui est largement ouvert aux recherches scientifiques : l'automatisation des traitements et les modélisations de la prosodie sont, en effet, loin d'être résolues. La modélisation de l'intelligibilité devra combiner tous ces aspects pour arriver à un haut niveau de mesure.

Cette problématique de la mesure de l'intelligibilité m'a également poussé vers une thématique nouvelle : la prise en compte de la déglutition. En effet, dans le domaine médical et hospitalier, l'analyse de celle-ci permet de traiter les dysphagies, problèmes qui se retrouvent chez 50% des personnes âgées et 80% des personnes présentant des troubles neurologiques et présentant des cancers de la tête et du cou. Du point de vue médical les enjeux sont considérables. Les déglutitions, toux, hemmages, sont aussi des sons que l'on traitait jusqu'à présent dans les enregistrements audio, mais en général pour les écarter, car l'on s'intéressait, en général, uniquement aux informations verbales. Tout comme la prosodie accorde une grande importance aux silences entre les mots, la prise en compte de la déglutition permettra peut-être une approche bien plus complète du signal de parole !

Chapitre 1

Le traitement automatique de la prosodie

1.1 Voix pathologiques

Dans le cadre du projet RUGBI (cf. §5.4.15 page 83), l'objectif de recherche qui m'était dévoué, en collaboration avec Corine Astésano, était de prendre en compte l'aspect prosodique dans la recherche d'unités porteuses d'intelligibilité. Le corpus RUGBI (cf. §3.3.3 page 29) est constitué d'enregistrements :

— de personnes atteintes de cancers de la tête et du cou,
— de personnes atteintes de la Maladie de Parkinson.

Cette dernière pathologie, va avoir des conséquences directes sur la communication et la prosodie en particulier :

— une voix monotone, expression sans émotion,
— des variations dans le débit de la parole,
— et des palilalies, soit des répétitions de certaines parties des mots (semblables au bégaiement).

Le rapport avec la prosodie et l'intelligibilité dans le cas des personnes atteintes de cancers n'est pas aussi évident qu'avec les personnes atteintes de la Maladie de Parkinson. En effet, bien que l'on note de manière générale une baisse du débit, les structures prosodiques sont souvent bien plus marquées que chez les personnes témoins. En effet, pour pouvoir gagner en intelligibilité, lorsqu'il est impossible de prononcer bien distinctement tous les phonèmes, la structure prosodique est un moyen qui permet de gagner en intelligibilité.

De nombreuses théories ont été établies afin de décrire les règles relatives à l'accentuation et plus généralement à la structure prosodique des langues [Di Cristo, 2011]. Je reprends ici le cadre que a été suivi dans la thèse de Robin Vaysse [Vaysse, 2023] et que je vais également suivre. On se place donc dans une combinaison d'approche Auto Segmentale (AM) et dans le cadre de la phonologie métrique.

L'approche métrique Auto Segmentale (AM) a été proposée par [Pierrehumbert, 1980] et adaptée au français par [Jun and Fougeron, 2000]. Cette théorie basée sur la phonologie autosegmentale de [Goldsmith, 1976] étudie des segments tonals qui décrivent les variations locales de f_0 au niveau syllabique et les mets en relation avec les contours intonatifs plus globaux indépendamment de la syntaxe.

La phonologie métrique qui est une théorie proposée par [Liberman, 1975] et [Liberman and Prince, 1977] et adaptée au français par [Di Cristo, 2000]. L'approche métrique pose l'accentuation au cœur de la structuration prosodique en représentant les relations de hiérarchie accentuelle, indépendamment des corrélats acoustiques (rappelons que l'AM se fonde essentiellement sur les variations de f_0 pour déterminer les types d'accents et de frontières). Cette théorie formalise l'alternance rythmique des syllabes

accentuées ou non, en ajoutant également une notion de poids métrique au centre de l'organisation accentuelle. La prosodie permet donc une structuration hiérarchique du flux de parole. Ainsi, ce flux continu peut être segmenté en unités de différentes tailles : les syllabes qui se regroupent elles-mêmes en mots puis en groupes de mots, en phrases et enfin en énoncés.

Je vais donc considérer les niveaux suivants (du plus court au plus long) :

syllabe qui est le niveau où se matérialise l'accent de par des modulations physiques des sons.

pied qui contient une syllabe accentuée suivie ou précédée (en fonction de la langue) d'un nombre fini de syllabes inaccentuées.

mot prosodique (*pw*) qui est constitué d'un mot lexical (porteur de sens) et de son ou ses clitiques (ex : *un oiseau* ; niveau proposé pour le français par [Astésano, 2017])

syntagme accentuel (AP) qui contient un mot prosodique et, potentiellement, son ou ses adjectifs (ex : *un petit oiseau bleu*). En l'absence d'adjectifs, il est possible que le niveau du *pw* et celui de l'AP représentent la même chose.

syntagme intermédiaire (ip) niveau prosodique à part entière, qui présente une réalité physique de par un allongement syllabique et des mouvements de f_0 se situant à un niveau intermédiaire entre l'AP et l'IP [Michelas and D'Imperio, 2010]

syntagme intonatif (IP) est une unité contenant généralement plusieurs AP [Beckman, 1996] et qui est le domaine correspondant aux contours intonatifs [Delattre, 1966]

énoncé qui correspond à minima à une « phrase » complète et qui peut contenir une ou plusieurs IP.

Vous trouverez un exemple de structure prosodique dans la figure 1.1.

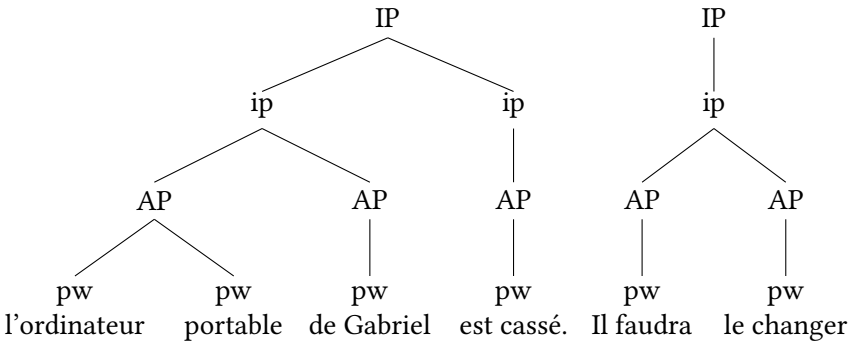


FIGURE 1.1 – Représentation de la structure prosodique de la phrase « L'ordinateur portable de Gabriel est cassé. Il faudra le changer. »

Robin Vaysse dans son travail de doctorat (cf. section 6.2.5 page 101) a produit une représentation automatique de la structure rythmique de séquences de quelques secondes (pouvant représenter une à deux phrases) : le spectre d'enveloppe d'amplitude (EMS) [Vaysse, 2023, Vaysse et al., 2023] (cf. figure 1.2). Il s'agit d'une représentation qui est obtenue en représentant les fréquences entre 0 et 10 Hz résultant d'une transformée de fourrier sur la courbe filtrée (pour se focaliser sur les voyelles) de l'amplitude de la parole.

Cette représentation ouvre de nouvelles perspectives dans l'étude automatique du rythme. Une analyse de longs enregistrements devrait permettre, grâce à une analyse par morceaux, de pouvoir en déduire des évolutions dans les répartitions d'énergie dans les différents niveaux prosodiques.

Nous avons également étudié la possibilité de faire cette représentation fréquentielle en utilisant un substrat composé à partir de la fréquence fondamentale f_0. Vu que la courbe de l'intonation est

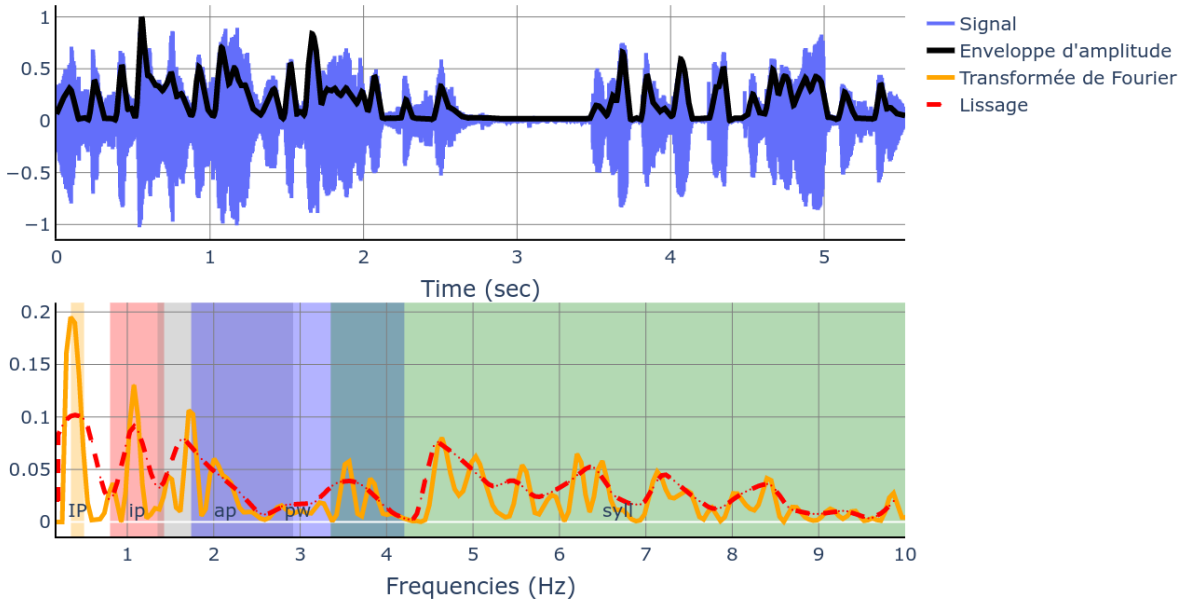


FIGURE 1.2 – Signal et enveloppe d'amplitude sur la figure du haut et sur la figure du bas : EMS d'un locuteur sain (C2SI n°33) sur l'extrait « Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres, il les perdait toutes de la même façon ». Les intervalles des niveaux prosodiques, manuellement annotés, ont été indiqués en couleur : orange pour l'IP (syntagme intonatif), rouge pour l'ip (syntagme intonatif), bleu pour le pw (mot prosodique), gris pour l'AP (syntagme accentuel) et vert pour la syllabe

discontinue, il est possible d'utiliser un modèle pour représenter une courbe continue de l'intonation. Il est possible d'utiliser MOMEL [Hirst, 2007], par exemple, pour obtenir ce résultat. La figure 1.3 donne un exemple de représentation à partir d'une courbe de l'intonation stylisée par MOMEL. Il y a peu d'énergie au-dessus de 4 Hz, ce qui est logique, nous ne disposons pas de représentation précise d'unités de type syllabique ou phonétique. Les informations présentes de 0 à 4 Hz sont, par contre, exploitables. Il serait intéressant de les combiner avec la représentation EMS. En effet, l'EMS n'est pas très fiable sur les très petites fréquences, car on est dépendant de la longueur du signal étudié. Une analyse comparative entre EMS et spectre de modulation, voire une combinaison des deux permettrait de rendre plus robuste l'affichage du spectre et mieux représenter les niveaux prosodiques.

Une autre amélioration permettrait de rendre la représentation plus universelle. Jusqu'à présent les différents niveaux prosodiques représentés sur les schémas proviennent d'annotation manuelle (par exemple sur la figure 1.2 les zones colorées représentant phonèmes, pw, AP, ip, IP) : ils sont segmentés manuellement sur le signal. La figure 1.4 montre un exemple de ces annotations sous le logiciel Praat [Boersma and Weenink, 1992].

Le lissage sur le spectre a été pour le moment dimensionné pour ne garder que trois points par seconde. On pourrait optimiser plus précisément afin de pouvoir obtenir exactement les pics d'énergie qui nous intéressent au niveau des niveaux IP, ip, AP et pw. Cela permettrait d'envisager de définir les zones les plus probables de ces niveaux, et donc de pouvoir annoter automatiquement ces zones. Pour y parvenir, il serait nécessaire de disposer d'annotations manuelles sur des exemples divers de parole. Il serait envisageable de lancer un sujet de recherche sur cette thématique, en continuant la collaboration avec Corine Astésano du Laboratoire de NeuroPsycholinguistique de l'Université Toulouse Jean Jaurès.

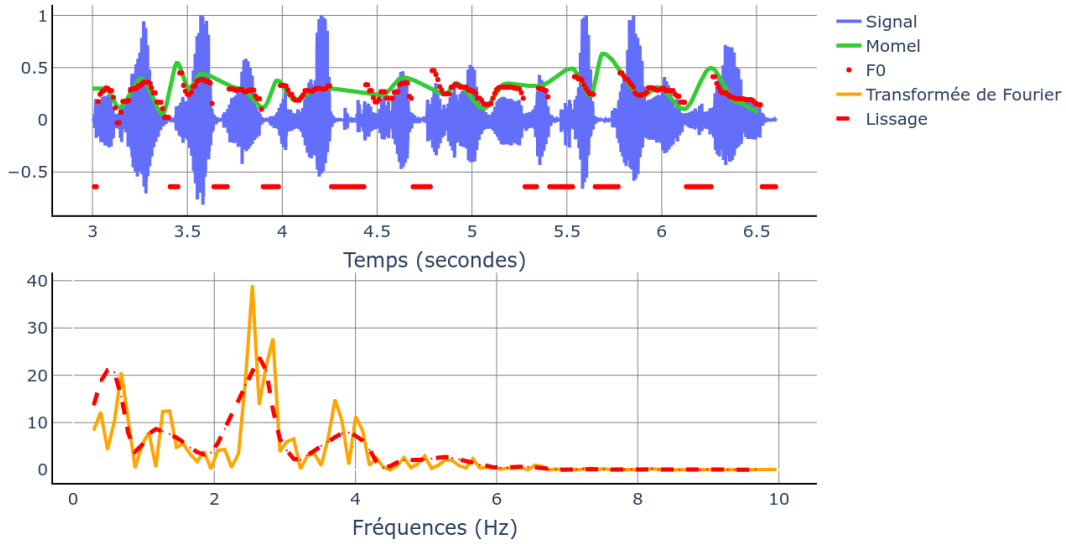


FIGURE 1.3 – Exemple de spectre de modulation de f_0. Le signal brut (en bleu) et sa courbe intonative (en vert) générée par MOMEL sont en haut. Le spectre de la courbe intonative est en bas. Le spectre brut est en orange et le spectre lissé est en rouge (en pointillés)

1.2 La modélisation des émotions

Depuis 2022, suite à la demande de collaboration de l'entreprise MyFamilyUp, j'ai abordé la prosodie sous un angle différent : celui de la modélisation des émotions. J'avais déjà abordé ce domaine lors de l'encadrement du stage de Abdelwahab Heba avec l'entreprise Intel en 2016 (cf. section 7 page 92). Dans le cadre de MyFamilyUp, leur objectif est d'arriver à détecter automatiquement des états émotionnels dans la parole pour le soutien psychologique. Arriver à détecter des émotions, à travers des entretiens, ou bien à travers l'interaction avec une application destinée à aider les aidants n'est pas chose facile. D'une part, car la détection d'émotions n'est pas triviale et peut s'avérer très différente d'un individu à l'autre. D'autre part, car cette détection ne se fait pas sur un simple échantillon de parole sur lequel on

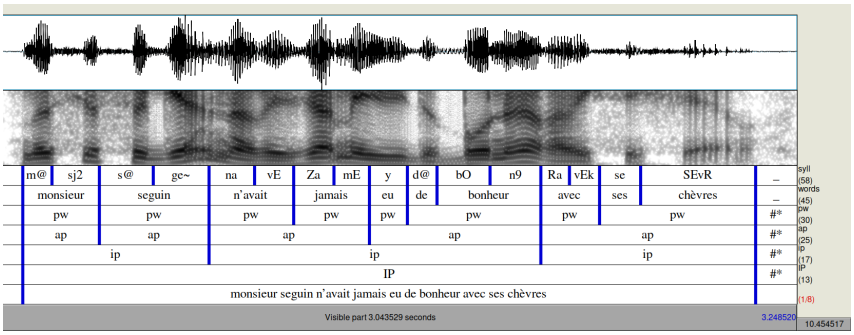


FIGURE 1.4 – Exemple d'annotation prosodique sur le logiciel Praat pour un locuteur témoin. On retrouve la forme d'onde du signal en haut, son spectrogramme en dessous et enfin les différentes zones concernent les annotations aux niveaux syllabiques (syll), pw, AP, ip et IP

doit donner une décision. Il est nécessaire de cumuler les indices et pouvoir prendre une décision sur un empan temporel qui peut s'avérer très large.

Une première méthode passera bien entendu par les méthodes classiques (cf. stage de master de Jérôme Susgin §23 page 95) que l'on retrouve dans l'état de l'art. Il s'agit de détecter des émotions en effectuant des projection autour des axes de « valence » et d'« arousal ». Pour entraîner ces systèmes, l'usage de corpus grand publics tels que RECOLA [Ringeval et al., 2013], IEMOCAP [Busso et al., 2008] ou MSP PODCAST [Lotfian and Busso, 2017] permettra la convergence de ces modèles. Mais cela sera-t-il applicable et transférable sur un autre cas d'usage ? Sur une autre langue (la plupart des corpus annotés ne le sont pas en français) ?

Il pourrait être intéressant de compléter cette approche, en exploitant la représentation prosodique détaillée en section 1.1. L'analyse des paramètres extraits du spectre de modulation de l'amplitude (EMS) sous forme de séquences pourrait permettre de détecter des changements ou des altérations dans l'état émotionnel du locuteur. Cette information (qui dans le cadre applicatif de MyFamilyUp sera combinée à d'autres sources, en particulier la détection d'émotions dans le texte) pourrait s'insérer dans le modèle de l'état psychologique des aidants.

FIGURE 1.5 – Enregistrement audio à la manière de Fernando Botero (généré par Midjourney en mars 2023, moteur de rendu n°4)

Chapitre 2

La modélisation de la déglutition

2.1 Les signaux issus du collier Swallis DSA®

La dysphagie oropharyngée est un problème majeur de santé publique. Ce trouble de la déglutition touche les personnes âgées (50%) et les patients atteints de troubles neurologiques et de cancers de la tête et du cou (80%), les principales complications étant les infections respiratoires et la dénutrition. Ces complications entraînent de longues hospitalisations, l'utilisation d'antibiotiques et d'aliments industriels modifiés, et une augmentation du taux de mortalité dans ces populations déjà fragiles. Le contrôle de ces complications et la réduction de leurs conséquences reposent sur la détection et le traitement précoces de la dysphagie. Or, à l'heure actuelle, les explorations cliniques de faible sensibilité, l'imagerie diagnostique (vidéofluoroscopie (VFS) et nasofibroscopie de déglutition (FEES)) jugée invasive, chronophage, peu accessible, et de surcroît pratiquée par un nombre limité de spécialistes, rendent le diagnostic et le suivi impossible pour la grande majorité de la population. Swallis Medical propose un collier de mesure appelé Swallis DSA®(cf. figure 2.1) qui permet de réaliser des analyses de manière non invasive. Le collier se place au niveau de la gorge, et permet d'acquisition de signaux provenant d'un microphone et d'accéléromètres.

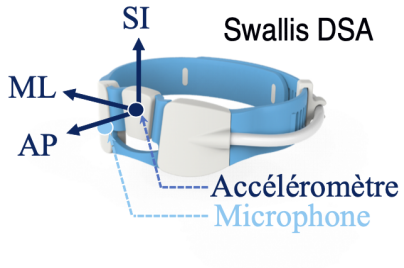


Figure 2.1 – Dispositif Swallis DSA®

La figure 2.2 présente une illustration des signaux obtenus après une toux et une déglutition. Y apparaissent de haute en bas :

- le signal de l'accéléromètre sur l'axe Supérieur-Inférieur (S-I)
- le signal de l'accéléromètre sur l'axe M-L Médial-Latéral (M-L)
- le signal de l'accéléromètre sur l'axe Antérieur-Postérieur (A-P)
- le signal du microphone
- le spectrogramme du signal du microphone

Les figures sont alignées au niveau temporel (axe des abscisses).

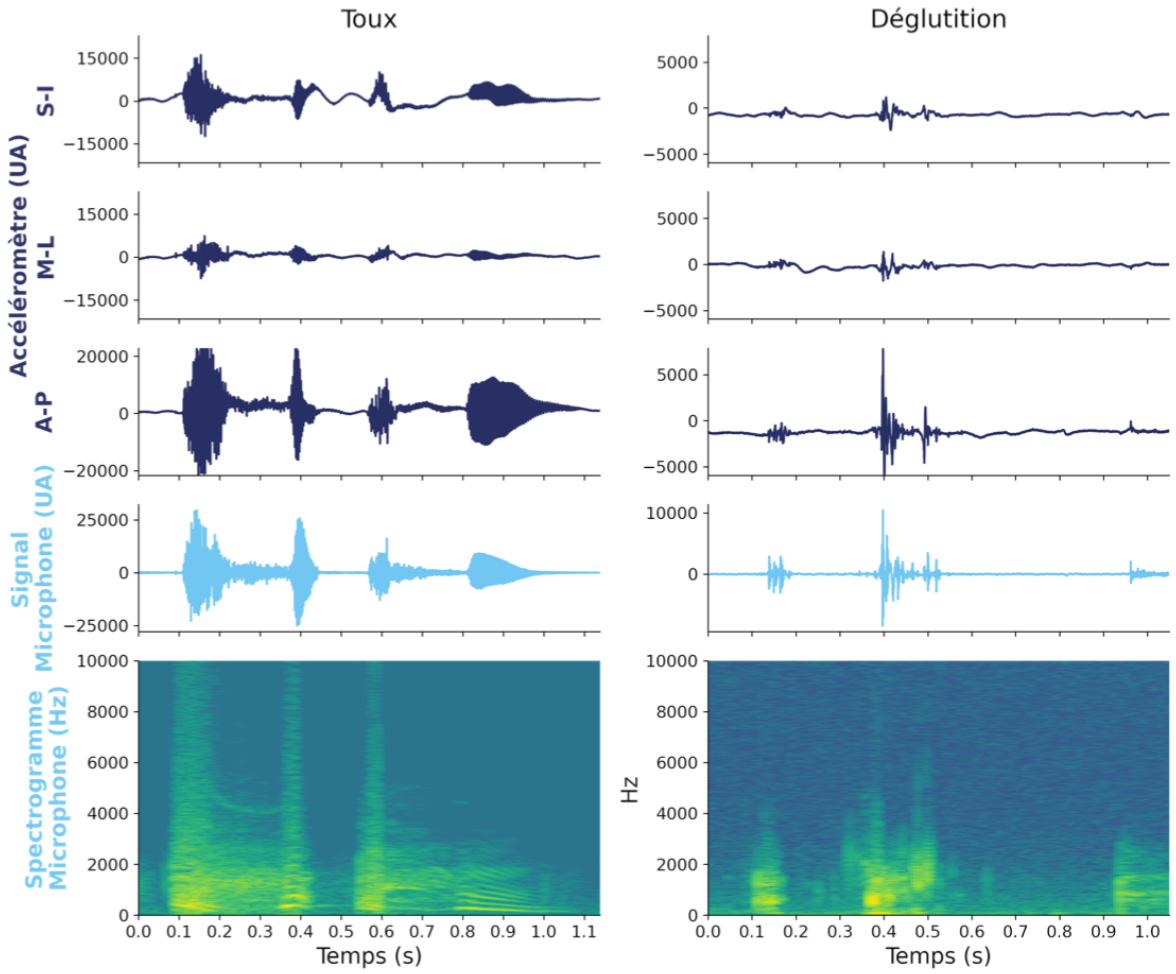


FIGURE 2.2 – Exemple de signaux de toux (à gauche) et de déglutition (à droite). Affichage des différents signaux et du spectrogramme du microphone. Figure extraite doctorat de Lila Gravellier.

La figure 2.3 représente une séquence plus longue, où apparaissent plusieurs déglutitions et une phase de phonation. Le signal de la phonation paraît saturé, car la représentation est amplifiée pour bien représenter les signaux peu énergétiques, mais le signal n'est pas écrêté en réalité, comme peut en témoigner le spectrogramme associé. Le signal de l'accéléromètre n'a pas été filtré pour les basses fréquences ici. En effet, les déplacements de la tête induisent des modifications lentes que l'on peut nettoyer en appliquant un filtre passe-haut. On notera que même si le mécanisme physiologique impliqué dans la déglutition consiste en une succession de trois phases (ascension du larynx, ouverture du sphincter haut et relâchement du larynx [Morinière et al., 2008]), leur réalisation acoustique est très variable dans le temps. Il arrive par moment que certaines phases disparaissent. La distribution des durées des déglutitions varie de 250 ms à 2 s.

Les signaux issus de l'accéléromètre sont parfois assez proches de ceux obtenus par le microphone. L'accéléromètre étant placé presque au même endroit que le microphone, au-dessus de la glotte, il est assez logique qu'il enregistre des déplacements très corrélés avec les sons produits lors de la déglutition.

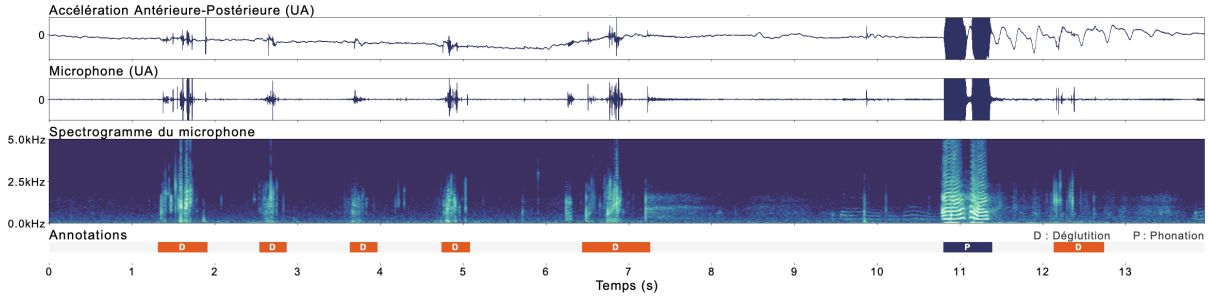


FIGURE 2.3 – Signaux de l'accélération Antérieure-Postérieure et du microphone du dispositif Swallis DSA®, avec le spectrogramme du microphone avec les annotations de déglutition et de zone de phonation. Figure extraite doctorat de Lila Gravellier.

2.2 Modélisation des signaux issus de la déglutition

La modélisation des signaux issus de la déglutition ne constitue pas un changement de paradigme très important pour une personne experte en traitement de la parole. Il s'agit de séquences temporelles, de formes qui disposent d'échelles de fluctuation sensiblement identiques à la parole. La nature multi-capteur de ce signal n'est également pas insurmontable.

Ma première approche avec ces signaux a été faite lors de la participation à l'encadrement d'une stagiaire en orthophonie Marie-Wenn Paugam « L'identification automatique des différents bruits de gorge chez le sujet sain : une étude pilote » encadrée par Virginie Woisard. Nous disposions alors en 2020 d'une version en développement du dispositif Swallis DSA®. N'ayant récolté que de 21 enregistrements, il était exclu d'envisager l'utilisation de réseaux de neurones profonds pour modéliser tous les événements de déglutition (nous souhaitions modéliser les déglutitions de différents liquides, mais également les toux, la phonation et les inspirations-expirations). J'ai donc réalisé des modèles de Markov en utilisant HTK pour représenter ces différents événements. J'ai modélisé des séquences de 7 à 9 états pour contraindre la taille minimale des événements. Les résultats au niveau du taux de reconnaissance étaient plutôt bon (97,2% pour la déglutition, 78,3% pour les mécanismes d'expulsion, 90,5% pour la phonation et 80,3% pour la respiration) mais il y avait de trop nombreuses insertions. Je n'avais modélisé que le signal du microphone, je n'ai pas eu le temps de rajouter aux modélisations la prise en compte des accéléromètres qui auraient sûrement diminué ces insertions.

Lila Gravellier (cf. section 6.2.6 page 102), dans le cadre de son doctorat en CIFRE chez Swallis Medical, a constitué une base de donnée de patients sains qui commence a être conséquente (49 personnes, 19 hommes et 23 femmes, de 22 à 57 ans, âge moyen de 34 ans) en suivant un protocole d'enregistrement très complet. Lila a pu entraîner des modèles de CNN pour représenter les déglutitions. Elle est en train d'essayer de rajouter des LSTM afin de mieux prendre en compte l'évolution temporelle.

Les déglutitions ayant une grande variabilité au niveau des trois différentes phases, je pense qu'il est indispensable de pouvoir modéliser les séquences temporelles de manière assez fine. Des modélisations présentes dans le chapitre 1 page 9 de la partie 1, il serait intéressant d'utiliser une topologie permettant l'omission de certains états comme cela était possible avec les HMM. Il faudra expérimenter différentes topologies de RNN. Ou bien mettre en place une structure assez souple au-dessus de représentations latentes apprises sur des enregistrements audio qui pourront être appris par transfert.

2.3 Vers une modélisation de l'efficacité du transport pharyngo-laryngé

Dans le cadre du projet PHLES-nid (cf. §5.4.21 page 86), il faudra réussi à caractériser finement les déglutitions dans le cadre de diverses pathologies. Car il sera demandé de modéliser les indicateurs d'efficacité pharyngolaryngée en référence aux grandes fonctions impliquant le pharyngo-larynx : respiration, déglutition, mécanisme de protection des voies aériennes et phonation.

Les indicateurs recherchés sont les suivants :

— l'efficacité du transport pharyngé
— efficacité contre l'aspiration
— réflexes de défense des poumons
— force de la toux
— contraste vocal
— clarté de la voix

Nous disposerons, à la fin du projet en 2026, d'environ 400 enregistrements d'un protocole comprenant de nombreux événements. Nous disposerons, à des fins de recherche, de capteurs complémentaires : nasofibroscopie, débit d'air nasal, électromyogramme de surface et oxymètre de pouls. Les événements seront annotés et des indicateurs cliniques seront disponibles. Nous serons amenés à produire des corrélations et régressions afin de pouvoir spécialiser nos modèles.

FIGURE 2.4 – Auscultation de la gorge à l'hôpital à la manière de Amedeo Modigliani (généré par Midjourney en mars 2023, moteur de rendu n°4)

Synthèse

Le premier domaine où je souhaite poursuivre concerne le traitement automatique autour de la prosodie. Le doctorat de Robin Vaysse a ouvert un grand champ de perspectives au niveau de la représentation automatique du rythme de la parole. À court terme, je vais m'investir sur le traitement automatique de la modélisation des émotions avec le doctorat CIFRE d'Adrien Lafore (cf. §6.2.7 page 103), où je souhaite exploiter le spectre de la modulation de l'amplitude. Mais il serait intéressant de poursuivre la collaboration avec Corine Astesano pour pouvoir continuer le travail de Robin. Cela permettrait de traiter des enregistrements de parole classique, mais également intéresser les aspects plus atypiques de parole : le domaine de la parole clinique, où la difficulté va être induite par les atteintes neurologiques, mais également les altérations de la sphère concernant l'Oto-Rhino-Laryngologie, dues à des opérations chirurgicales, mais également à des traitements (chimiothérapie ou radiothérapie) ; mais également de nombreux autres domaines où la parole diffère d'une représentation canonique : apprentissage des langues secondes, parole de personnes âgées, parole des aidants de personnes handicapées, parole de locuteurs régionaux… Ces apports automatiques permettraient aux linguistes, médecins, orthophonistes, de disposer de complément d'information facilement accessible. Mais la matérialisation de données prosodiques permettrait également d'améliorer les systèmes de reconnaissance ou de synthèse automatique de la parole.

Le second domaine concerne la recherche d'une meilleure qualité de vie et le fait de vouloir avoir un impact sociétal avec les recherches menées. Nous avons posé de nombreuses pierres : constitution de corpus (cf. §3.3 page 27), facilitation d'accès aux données (cf. Parolothèque §5.4.20 page 86), production scientifique (cf. §3 page 25), essais cliniques (cf. Protopitch SAMI §5 page 76, projet ADAPT §5.4.22 page 87). Mais il reste de quoi faire pour aller plus loin sur la modélisation de l'intelligibilité et la compréhensibilité, et ses conséquences sur la qualité de vie. Le traitement des données issues de la déglutition va permettre d'influencer les pratiques cliniques en apportant plus de facilité dans les examens de suivi. La poursuite du co-encadrement du doctorat de Lila Gravellier (cf. §6.2.6 page 102), mais également le projet PHLES-nid (cf. §5.4.21 page 86) et le co-encadrement du doctorat de Philippe Allet (cf. §6.2.8 page 104) va permettre d'avancer sur la modélisation et la caractérisation de la déglutition. La facilitation apportée par l'apport d'outils automatiques pour caractériser cette déglutition permettrait d'envisager des usages hors de l'hôpital, ou bien sur des dispositifs légers (téléphone, tablette). Cela pourrait améliorer le suivi des patients, les médecins pourraient avoir plus d'informations sur les patients. Mais à plus long terme, il serait possible d'avoir une action sur la qualité de vie en elle-même. Pas seulement à travers des mesures automatiques pour mieux pouvoir quantifier cette qualité de vie. Il est imaginable que le retour sur la qualité de vie des patients puisse influencer les traitements pratiqués par les médecins : il serait en effet possible de privilégier les interventions qui garantissent un meilleur niveau de vie.

Troisième partie

Curriculum Vitæ

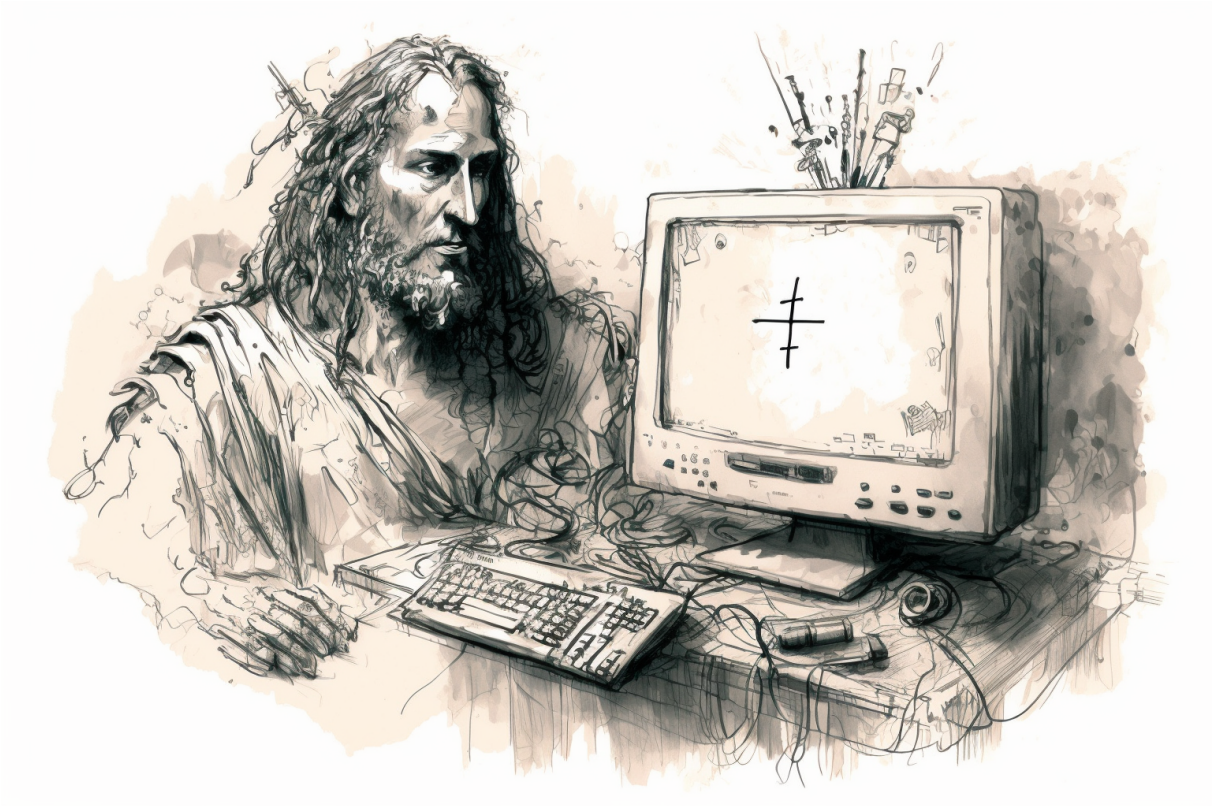


FIGURE 3 – Informatique à la manière de Léonard de Vinci (généré par Midjourney en mars 2023, moteur de rendu n°4)

Curriculum Vitæ

Sommaire

1 Identité

Nom Farinas

Prénom Jérôme

Né le 16 août 1974

Nationalité française

Fonction Maître de conférence en informatique

Corps Maître de conférence

Date d'affectation 1er septembre 2003

Grade Hors Classe

Échelon 5 (septembre 2021)

Doctorat obtenu le 15 novembre 2002 à l'Université Paul Sabatier (Toulouse III)

École doctorale Mathématiques, Informatique, Télécommunications de Toulouse (ED 475 MITT)

PEDR depuis le 1er octobre 2019

2 Parcours universitaire

1992-1993 Mathématiques Supérieures – Lycée Michel Montaigne (Bordeaux)

1993-1995 DEUG sciences : Mathématique, Informatique et Applications aux sciences – Université Paul Sabatier

1994-1995 DULS espagnol – Université Paul Sabatier

1995-1996 Licence informatique – Université Paul Sabatier

1996-1997 Maîtrise informatique – Université Paul Sabatier

5-14 oct 1997 International school on Neural Nets – Vietry (Italie), 3ème école d'été "E.R. Caianiello", « Speech Processing, Recognition and Articial Neural Networks »

1997-1998 DULS anglais – Université Paul Sabatier

1997-1998 DEA Informatique de l'Image et du Langage – Université Paul Sabatier « La prosodie pour l'identification automatique des langues »

1998-2002 Doctorat en Informatique – Université Paul Sabatier, Mention très honorable : « Une modélisation automatique du rythme pour l'identification des langues »
Jury :
- Daniel Dours, Université Toulouse 3 (président)
- Jean-François Bonastre, Université d'Avignon (rapporteur)
- Daniel Hirst, Laboratoire Parole et Langage, Aix-en-Provence (rapporteur)
- Édouard Geoffrois, DGA-DCE centre technique d'Arcueil (examinateur)
- François Pellegrino, Université Lyon 2 (examinateur)
- Régine André-Obrecht, Université Toulouse 3 (directrice de thèse)

3 Parcours professionnel

1998-2002 Vacataire à l'Université Paul Sabatier, l'ENSEEIHT et l'INSA Toulouse

2001-2003 Attaché Temporaire à l'Enseignement et à la Recherche Université Paul Sabatier

Depuis 2003 Maître de conférences Université Paul Sabatier – Laboratoire de recherche : Institut de Recherche en Informatique de Toulouse

4 Activités d'enseignement

Le tableau 1 détaille mes principales contributions et le lien vers les supports crées tout au long de mon activité d'enseignement.

Table 1 – Principaux enseignements créés (avec lien vers supports)

Unité d'Enseignement	Niveau	Implication	Audience
Introduction au matériel informatique, système Unix et réseaux	M2 BI	création CM,TP (CM : 68 slides)	18 étudiants
Représentation et manipulation de contenus 3D, Image et Son	M1 info	création TP et éval. QCM (6 TP : moodle)	72 étudiants
Langage pour Recherche Information	L3 SID	création CM/TD/TP (CM : 209 slides)	36 étudiants
Programmation en bioinformatique	M1 MABS	création CM/TD/TP (livret : 119 pages)	18 étudiants
Bases de l'informatique	L1 SFA	création 2CM/2TP/QCM (moodle)	254 étudiants
BIOMIP4 Informatique	L2 Bio	création CM/TP (moodle)	24 étudiants
Algorithmique et programmation	L1 SFA	création 1CM/2TP	509 étudiants
Indexation A/V Media Parole	M2 SID	création CM/TP (CM : 115 slides)	36 étudiants
Calcul numérique pour les neurosciences	M2 NCC	création CM/TD/TP (CM : 52 slides)	18 étudiants
Environnements Sonores	M2 IGAI	création CM/TD/TP/QCM (CM : 133 slides)	18 étudiants
Modélisation Calcul Scientifique	M1 info	création 1TD, TP (moodle)	130 étudiants
Technologies Vocales	M2 IARF	création CM/TD/TP (CM/TD : 382 slides)	32 étudiants

4.1 Enseignements 2021-2022

La figure 4 représente le volume (ramené à des heures équivalent TD) réalisé lors de l'année universitaire 2021-2022 trié par le niveau universitaire des étudiants. Le volume total est de 239 HETD. Au niveau licence, des cours sont réalisés en 1ère et 2ème année et concernent de l'initiation à la programmation et à la théorie de l'information. Les enseignements en Master concernent principalement des enseignements sur le traitement automatique de la parole et l'apprentissage automatique. Les enseignements en niveau doctorat sont constitués d'un cours sur l'initiation de la programmation à destination d'un public ayant été très peu sensibilisé aux traitements automatiques dans leur cursus : les étudiants de l'École Doctorale Comportement, Langage, Éducation, Socialisation, Cognition (CLESCO).

La table 2 détaille les matières enseignées, ainsi que leur volume horaire associé. Une grande partie des enseignements concerne des étudiants de la filière Biologie et Santé et vise à leur apprendre les rudiments d'algorithmique, de complexité et de programmation. L'enseignement spécialisé sur la reconnaissance automatique de la parole est réalisé à des parcours différents en Master 2 : le parcours

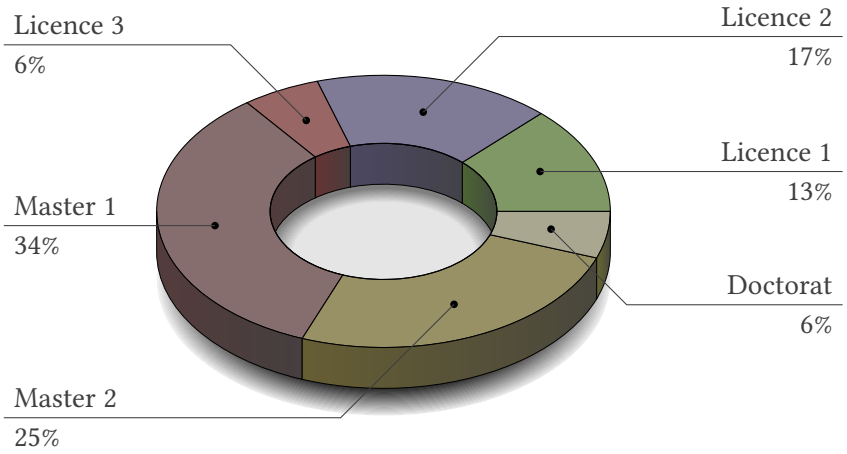


FIGURE 4 – Répartition des enseignements en HETD réalisés en 2021/2022 par niveau universitaire (sur 239 HETD)

Statistique et Informatique Décisionnelle et le parcours informatique, le parcours Science du Numérique (3SN) de la 3ème année d'ingénieur de l'école d'ingénieur ENSEEIHT.

TABLE 2 – Matières enseignées en 2021-2022

Matière	**Formation**	**Type**
Algorithmique	L1 info	TP
Initiation à l'algorithmique et à la programmation	L2 BioMIP	CTD
Programmation Python	L3 BioMIP	CTD
Intelligence Artificielle	L3 info	TP/Stage
Projet tuteuré	M1 bioinformatique	CM/Projet
Algorithmique et complexité	M1 bioinformatique	CTD
Introduction au traitement du signal, aux signaux sonores et aux images	M1 info	TP
Calcul Scientifique et apprentissage automatique	M1 info	TD
Analyse et exploitation de données	M2 SID	CM/TD/TP
Reconnaissance des formes et technologies vocales	M2 info	CM/TD/TP
Audionumérique Parole et Musique	3EN ENSEEIHT	CM/TD/TP
Initiation programmation MATLAB	ED CLESCO	CM/TD/TP

La figure 5 détaille la répartition des enseignements réalisées par type lors de l'année universitaire 2021-2022. Afin de gagner en lisibilité, les enseignements de type « Cours-Travaux Dirigés » ont été répartis de manière égale entre « Cours » et « Travaux Dirigés ». Le terme « Projet » désigne : les projets qui sont réalisés dans le tronc commun du Master 1 qui permettent de compléter les enseignements pratiques et les projets tuteurés en fin d'année du Master 1 de bioinformatique qui permettent à des étudiants en groupe de 3 ou 4 de mener à bien une réalisation pratique mettent en œuvre les connaissances théoriques vues lors des mois précédents.

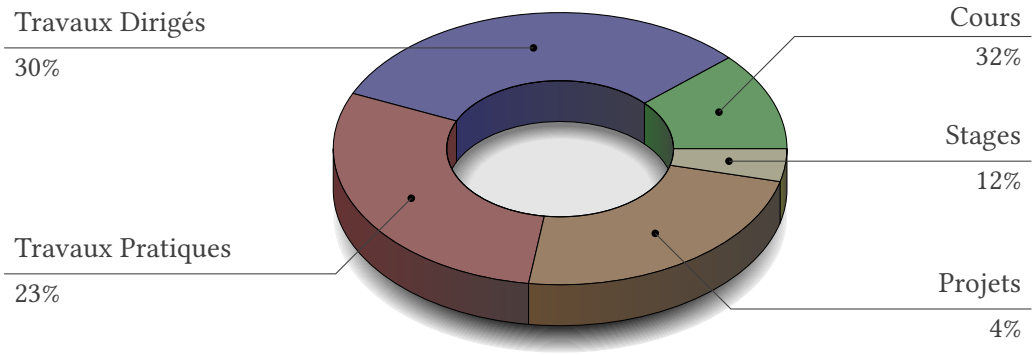


FIGURE 5 – Répartition du volume des enseignements par type en 2021/2022

4.2 Formations suivies

- Tutoriel de George Saon (IBM T. J. Watson Research Center, Yorktown Heights, NY) et Jen-Tzung Chien (National Chiao Tung University, Hsinchu, Taiwan) sur « **Advances in Large Vocabulary Continuous Speech Recognition** », InterSpeech'2013, Lyon, France, 2013
- Tutoriel de Jacqueline Vaissière (Laboratoire de Phonétique et de Phonologie, Paris, France) sur « **Spectrogram reading course** », InterSpeech'2013, Lyon, France, 2013
- **Comprendre la génération Y**, formation F1315 assurée par François Debly par le Service Universitaire de Pédagogie, le mardi 26 novembre 2013
- **Enseigner avec les mindmaps**, formation F1333 assurée par Isabelle Chênerie par le Service Universitaire de Pédagogie, le vendredi 10 janvier 2014
- **Amphi interactif**, formation F1421 assurée par Franck Amadieu par le Service Universitaire de Pédagogie, le lundi 13 octobre 2014
- **L'évaluation formative : comment évaluer pour remédier aux difficultés d'apprentissage**, formation F1422, assurée par Laurent Talbot par le Service Universitaire de Pédagogie, le mardi 4 novembre 2014
- Tutoriel de Li Deng (Deep Learning Technology Center Microsoft Research, Redmond, USA) « **Deep Learning for Speech/Language Processing, machine learning and signal processing perspectives** », Interspeech'2015, Dresde, Allemagne, 6 septembre 2015
- Tutoriel de Bernd T. Meyer (Medizinische Physik and Cluster of Excellence Hearing4all, University of Oldenburg, Allemagne), Hynek Hermansky (Center for Language and Speech Processing, The Johns Hopkins University) et Nelson Morgan (International Computer Science Institute and University of California at Berkeley, Berkeley, CA, USA) « **Bio-inspired speech recognition : From human perception to deep networks** », Interspeech'2015, Dresde, Allemagne, 6 septembre 2015
- **Recruter un chercheur post-doctorant** Formation organisée par Danièle Véret à la DR 14 CNRS, Délégation Midi-Pyrénées, Toulouse, 23 au 24 septembre 2015
- Tutoriel de Oldooz Hazrati, Hussnain Ali, John H. L. Hansen (The University of Texas, Dallas, USA) and James M. Kates (University of Colorado, Boulder, USA) « **Hearing Assistive Technologies : Challenges and Opportunities** », Interspeech'2016, San Francisco, USA, 8 septembre 2016
- Journée de formation AFCP « Statistiques et données phonétiques atypiques », Nicolas Audibert, Frédérique Lethué, Olivier Crouzet, Université Paris Diderot, Paris, 28 juin 2017
- Tutoriel de Hung-yi Lee (Department of Electrical Engineering, National Taiwan University, Taiwan) et Yu Tsao (Research Center for Information Technology Innovation, Academia Sinica,

Taipei, Taiwan) « **Generative adversarial network and its applications to speech signal and natural language processing** », Interspeech'2019, Graz, Autriche, 25 septembre 2019
- Tutoriel de Takaaki Hori (Mitsubishi Electric Research Laboratories), Tomoki Hayashi (Department of Information Science, Nagoya University), Shigeki Karita (NTT Communication Science Laboratories) et Shinji Watanabe (Center for Language and Speech Processing, Johns Hopkins University), « **Advanced methods for neural end-to-end speech processing – unification, integration, and implementation** », Interspeech'2019, Graz, Autriche, 25 septembre 2019
- Tutoriel de Piotr Żelasko (Johns Hopkins University, États Unis d'Amérique), Sanjeev Khudanpur (Johns Hopkins University, États Unis d'Amérique) et Daniel Povey (Xiaomi, Chine) « **Speech Recognition with Next-Generation Kaldi (K2, Lhotse, Icefall)** », Interspeech'2021, Brno, République Tchèque, 30 août 2021
- Tutoriel de Mirco Ravanelli (Mila - Université de Montréal, Canada), Titouan Parcollet (LIA - Avignon University), Aku Rouhe (Aalto University, Finlande) « **SpeechBrain : Unifying Speech Technologies and Deep Learning With an Open Source Toolkit** », Interspeech'2021, Brno, République Tchèque, 30 août 2021
- Tutoriel de Amalia Arvaniti (Radboud University, Pays Bas), Cong Zhang (Radboud University, Pays Bas), Kathleen Jepson (Radboud University, Pays Bas) et Katherine Marcoux (Radboud University, Pays Bas) « **Intonation Transcription and Modelling in Research and Speech Technology Applications** », Interspeech'2021, Brno, République Tchèque, 30 août 2021
- Tutoriel de Fei Chen (Department of Electrical and Electronic Engineering, Southern University of Science and Technology, China) and Yu Tsao (The Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan) sur « **Speech Assessment Metrics : from psychoacoustics to Machine Learning** », Interspeech'2023, Dublin, Ireland, 20 août 2023
- Tutoriel de Fangjun Kuang (Xiaomi, China) Matthew Wiesner (John Hopkins University, USA), Piotr Zelasko (Meaning, USA), Desh Raj (John Hopkins University, USA), Dan Povey (Xiaomi, China) Sanjeev Khudanpur (John Hopkins University, USA), Leibny Paola Garcia Perera (Johns Hopkins University, USA) and Jan "Yenda" Trmal (Johns Hopkins University, USA) sur « **Open Source Tools for ASR with LHoste and Icefall** », Interspeech'2023, Dublin, Ireland, 20 août 2023

4.3 Responsabilités

- Responsable de l'année Licence 3 du Cursus Master en Ingénierie Statistique et Informatique Décisionnelle de 2013 à 2016 : mise en place de l'organisation des enseignements, coordination des équipes pédagogiques, suivi des stages en entreprises, promotion de la formation en France (75% de recrutement en L3 se fait en dehors de Toulouse), planification des examens, relation avec les intervenants industriels extérieurs, participation à la mise en place de la convention de partenariat avec l'école Toulouse Business School, suivi des étudiants.
- Membre du conseil de perfectionnement du CMI SID de 2013 à 2016 (une réunion plénière par an) : mise en place des nouvelles orientations de la formation en fonction des retours du milieu de la recherche et du milieu industriel
- Co-porteur depuis 2015 de la mention bio-informatique pour l'accréditation 2016-2021 et 2021-2026 : préparation du dossier d'accréditation, animation de réunions pédagogiques, création du nouveau parcours de master.
- Responsable M1 (avec Roland Barriot) et M2 (avec Gwennaele Fichant) du parcours Bioinformatique et biologie des systèmes depuis septembre 2016

— Participation au dossier IDEX formation « Innovation en licence » BIOMIP avec Christel Lutz et Noémie Davezac. Ce projet vise à proposer un cursus pédagogique « BIOlogie avec renforcement en Mathématiques, Informatique, Physique ». Je suis responsable avec Julien Pinquier de la proposition informatique.
— Responsable UEs :
  — L2 Biomip "Algorithmique, Unix et programmation R",
  — L3 Biomip "Programmation Python",
  — M1 bioinfo "Algorithmique et complexité",
  — M1 bioinfo "Projet Tuteuré",
  — M2 IARF "Reconnaissance des formes et Technologies Vocales",
  — M2 SID "Analyse et exploitation des données",
  — ENSEEIHT 3SN "Audionumérique Parole et Musique".

5 Activités de recherche

5.1 Liste des publications

5.1.1 Analyses synthétiques

Le nombre de publications est détaillé dans le tableau 3.

Table 3 – Nombre de publications

Type	Nombre
Articles de revues internationales	13
Articles de revues nationales	3
Conférences et ateliers internationaux	28
Conférences et ateliers nationaux	28
Conférences sans actes publiés	23
Conférencier invité	14
Déclaration d'invention et dépôt logiciel	5

Publications les plus significatives :

1. Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2005). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, 47(4):436–456
2. Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., and Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language and Hearing Research*, 60:2394–2405
3. Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., and Roger, V. (2021). C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55:173–190
4. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023b). Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation. *International Journal of Language and Communication Disorders*, 58(1):39–51

5. Vaysse, R., Astesano, C., and Farinas, J. (2022a). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *Journal of the Acoustical Society of America (JASA)*, 152(5):3091–3101

Ces articles dans des revues internationales sont représentatives de différents travaux que j'ai menés. La première détaille la modélisation du rythme pour l'identification des langues et la proposition de pseudo-syllabe issu de mon doctorat. La deuxième détaille les résultats qui ont été obtenus dans la prédiction automatique de la presbyacousie en utilisant un système de reconnaissance automatique de la parole. La troisième détaille la constitution du corpus de parole pathologique qui a donné lieu aux recherches sur la mesure automatique de l'intelligibilité chez les personnes atteintes de cancer ORL. La quatrième présente l'un des résultats du travail avec Mathieu Balaguer : l'impact des troubles de la parole sur la communication dans les cercles sociaux. La cinquième présente un des résultats du doctorats de Robin Vaysse sur la modélisation de la prosodie.

La liste exhaustive des publications classées par type d'ouvrage est détaillée ci-dessous.

5.1.2 Articles de revues internationales

1. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023b). Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation. *International Journal of Language and Communication Disorders*, 58(1):39–51
2. Vaysse, R., Astesano, C., and Farinas, J. (2022a). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *Journal of the Acoustical Society of America (JASA)*, 152(5):3091–3101
3. Roger, V., Farinas, J., and Pinquier, J. (2022a). Deep neural networks for automatic speech processing: a survey from large corpora to limited data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(19):15. Soumis : 19 novembre 2021, accepté : 15 juillet 2022, publié le 17 août 2022
4. Balaguer, M., Pommée, T., Pinquier, J., Farinas, J., Woisard, V., and Sordes, F. (2022). Development and preliminary validation of the questionnaire 'Evaluation of the constitution of social circles (ECSC)' in patients treated for cancer of the upper aerodigestive tract. *Folia Phoniatrica et Logopaedica*
5. Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astesano, C., Pinquier, J., and Lepage, B. (2022). Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The carcinologic speech severity index. *Head & Neck*, 44(1):71–88. publié en ligne 2/11/2021
6. Balaguer, M., Champenois, M., Farinas, J., Pinquier, J., and Woisard, V. (2021a). The (head and neck) carcinologic handicap index: validation of a modular type questionnaire and its ability to prioritise patients' needs. *European Archives of Oto-Rhino-Laryngology*, 278:1159–1169
7. Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., and Roger, V. (2021). C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55:173–190
8. Balaguer, M., Pommée, T., Farinas, J., Fichaux-Bourin, P., Puech, M., Pinquier, J., and Woisard, V. (2020a). Validation of the French versions of the Speech Handicap Index and the Phonation Handicap Index in patients treated for cancer of the oral cavity or oropharynx. *International Journal of Phoniatrics, Speech Therapy and Communication Pathology: Folia Phoniatrica et Logopaedica*, 72(6):464–477

9. Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V., and Speyer, R. (2020b). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review. *Journal of the sciences and specialities of the Head and Neck*, 42(1):111–130

10. Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., and Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language and Hearing Research*, 60:2394–2405

11. Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Alazard-Guiu, C., Robert, M., and Gatignol, P. (2015). Automatic Assessment of Speech Capability Loss in Disordered Speech. *ACM Transactions on Accessible Computing (TACCESS)*, 6(3):8:1–8:14

12. Reby, D., André-Obrecht, R., Galinier, A., Farinas, J., and Cargnelutti, B. (2006). Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (Cervus elaphus) stags. *Journal of the Acoustical Society of America (JASA)*, 120(6):4080–4089

13. Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2005). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, 47(4):436–456

5.1.3 Articles de revues nationales

1. Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., and Woisard, V. (2021f). Paramètres perceptifs expliquant la sévérité du trouble de parole mesurée automatiquement en cancérologie ORL. *Rééducation orthophonique*, 286:1–13. "Rééducation Orthophonique" est une revue scientifique trimestrielle, réalisée par la Fédération Nationale des Orthophonistes. Chaque numéro est thématique

2. Fontan, L., Magnen, C., Tardieu, J., Ferrané, I., Pinquier, J., Farinas, J., Gaillard, P., and Aumont, X. (2015b). Comparaison de mesures perceptives et automatiques de l'intelligibilité : application à de la parole simulant la presbyacousie. *Traitement Automatique des Langues*, 55(2):151–174

3. Farinas, J., Rouas, J.-L., Pellegrino, F., and André-Obrecht, R. (2005). Extraction automatique de paramètres prosodiques pour l'identification automatique des langues. *Traitement du Signal*, 22(2):81–97

5.1.4 Conférences et ateliers internationaux avec actes édités et comité de lecture

1. Balaguer, M., Gelin, L., Woisard, V., Farinas, J., and Pinquier, J. (2021b). Measurement of speech intelligibility after oral or oropharyngeal cancer by an automatic speech recognition system. In *12th International Workshop MAVEBA (Models and analysis of vocal emissions for biomedical applications)*, Firenze, Italy. Università degli Studi Firenze

2. Vaysse, R., Farinas, J., Astésano, C., and André-Obrecht, R. (2021b). Automatic extraction of speech rhythm descriptors for speech intelligibility assessment in the context of Head and Neck Cancers. In *INTERSPEECH*, Proceeding of Interspeech 2021, Brno, Czech Republic. ISCA

3. Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., and Woisard, V. (2020c). Functional impact of speech disorders in patients treated for oral or oropharyngeal cancer, assessed by perceptual and automatic measurements (education paper). In *Motor Speech Conference, Hyatt Centric Santa Barbara, California, USA, 20/02/2020-23/02/2020*

4. Pellegrini, T., Farinas, J., Delpech, E., and Lancelot, F. (2019). The Airbus Air Traffic Control speech recognition 2018 challenge: towards ATC automatic transcription and call sign detection. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2019), Graz, Autriche, 15-19/09/2019*, pages 2993–2997, Graz, Autriche. International Speech Communication Association (ISCA)

5. Balaguer, M., Farinas, J., Pinquier, J., and Woisard, V. (2019b). Construction of an automatic Carcinologic Speech Severity Index (C2SI) score (regular paper). In *World Congress of the International Association of Logopedics and Phoniatrics (IALP 2019), Taipei, Taiwan, 18–22/08/2019*, page to appear. IALP : International Association of Logopedics and Phoniatrics

6. Farinas, J. (2019). Interests of using Automatic Speech recognition for Speech-Language Therapists (regular paper). In *World Congress of the International Association of Logopedics and Phoniatrics (IALP 2019), Taipei, Taiwan, 18–22/08/2019*. IALP : International Association of Logopedics and Phoniatrics

7. Gaume, B., Tanguy, L., Fabre, C., Ho-Dac, L.-M., Pierrejean, B., Hathout, N., Farinas, J., Pinquier, J., Danet, L., Peran, P., De Boissezon, X., and Jucla, M. (2018). Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures (regular paper). In *Natural Language Processing and Cognitive Science, Krakow, Poland, 11–12/09/18*, pages 19–26. Jagiellonian Library

8. Laaridh, I., Tardieu, J., Magnen, C., Gaillard, P., Farinas, J., and Pinquier, J. (2018a). Perceptual and Automatic Evaluations of the Intelligibility of Speech Degraded by Noise Induced Hearing Loss Simulation (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India, 2–6/09/18*, pages 2953–2957. International Speech Communication Association (ISCA)

9. Astesano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pont, O., Pouchoulin, G., Puech, M., Robert, D., Sicard, E., and Woisard, V. (2018). Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer (regular paper). In Calzolari, N., editor, *Language Resources and Evaluation Conference (LREC 2018), Miyazak, Japon, 07/05/2018-12/05/2018*, pages 4265–4271. European Language Resources Association (ELRA)

10. Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H., and Farinas, J. (2016). Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA, 8–12/09/16*, pages 2686–2690. International Speech Communication Association (ISCA)

11. Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., and Aumont, X. (2016). Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA, 8–12/09/16*, pages 650–654. International Speech Communication Association (ISCA)

12. Fontan, L., Farinas, J., Ferrané, I., Pinquier, J., and Aumont, X. (2015a). Automatic intelligibility measures applied to speech signals simulating age-related hearing loss (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany, 6–10/09/15*, pages 663–667. International Speech Communication Association (ISCA)

13. Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., and Robert, M. (2014). The Goodness of Pronunciation algorithm applied to disordered speech (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2014), Singapour, 14–18/09/14*, pages 1463–1467. International Speech Communication Association (ISCA)

14. Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., and Farinas, J. (2011). The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news (regular paper). In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., and Odijk, J., editors, *Language Resources and Evaluation Conference (LREC 2010), Valletta, Malte, 19/05/10-21/05/10*, pages 1686–1689. European Language Resources Association (ELRA)

15. Arias, J. A., André-Obrecht, R., and Farinas, J. (2008c). Unsupervised signal segmentation based on temporal spectral clustering. In *European Signal and Image Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, 25–29/08/08*, page (online). EURASIP

16. Arias, J. A., André-Obrecht, R., and Farinas, J. (2008a). Automatic low-dimensional analysis of audio databasis. In *International Workshop on Content-Based Multimedia Indexing (CBMI 2008), London, UK, 18-20/06/08*, pages 556–559. IEEE

17. Pellegrino, F., Farinas, J., and Rouas, J.-L. (2004). Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech. In Bel, B. and Marlien, I., editors, *International Conference on Speech Prosody 2004, Nara, Japon, 23/03/04-26/03/04*, pages 517–520, ISBN 2-9518233-1-2. ISCA Special Interest Group on Speech Prosody (SproSIG)

18. Parlangeau, N., Farinas, J., Fohr, D., Illina, I., Magrin-Chagnolleau, I., Mella, O., Pellegrino, F., Pinquier, J., Senac, C., and Smaili, K. (2003). Audio Indexing On The Web: A Preliminary Study Of Some Audio Descriptors. In *7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, Florida, USA, 27–30/07/03*. International Institute of Informatics and Systemics

19. Rouas, J.-L., Farinas, J., and Pellegrino, F. (2003a). Automatic Modelling of Rhythm and Intonation for Language Identification. In *15th International Congress of Phonetic Sciences (15th ICPhS), Barcelona, Spain, 3-9/08/2003*, pages 567–570

20. Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2003b). Modeling Prosody for Language Identification on Read and Spontaneous Speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003), Hong Kong, China, 06–10/04/2003*, volume I, pages 40–43. IEEE

21. Rouas, J.-L., Farinas, J., and Pellegrino, F. (2002a). Merging segmental, rhythmic and fundamental frequency features for automatic language identification. In *Eusipco 2002, Toulouse, 03–06/09/02*, volume III, pages 591–594. EURASIP

22. Farinas, J., Pellegrino, F., Rouas, J.-L., and André-Obrecht, R. (2002). Merging segmental and rhythmic features for Automatic Language Identification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002), Orlando, USA, 13–17/05/2002*, volume I, pages 753–756. IEEE

23. Pellegrino, F., Chauchat, J.-H., Rakotomalala, R., and Farinas, J. (2002). Can automatically extracted rhythmic units discriminate among languages? In *International Conference of Speech Prosody 2002, Aix-en-provence, France, 11–13/04/02*, pages 563–566. International Speech Communication Association (ISCA)

24. Farinas, J. and André-Obrecht, R. (2001b). Phonotactic modeling of broad acoustic classes for a differentiated approach of Automatic Language Identification. In *17th International Congress on Acoustics, ICA'2001, Rome, Italie, 02/09/01-07/09/01*, Italie. ASA

25. Farinas, J. and Pellegrino, F. (2001). Comparison of two approaches to Language Identification. In *7th International Conference on Speech Communication and Technology, Eurospeech 2001, Aalborg, Danemark, 3–7/09/2001*, volume I, pages 399–402. International Speech Communication Association (ISCA)

26. Farinas, J., Pellegrino, F., and André-Obrecht, R. (2000). Automatic Language identification: from a phonetic differentiated model to a complete system. In *Workshop on Friendly Exchanging through the Net, COST 254'2000, Bordeaux, 23–24/03/2000*, pages 97–102, ENITA/ENSERB, Bordeaux. C. Germain, E. Grivel and O. Lavialle

27. Pellegrino, F., Farinas, J., and André-Obrecht, R. (1999b). Vowel System Modeling: A Complement to Phonetic Modeling in Langage Identification. In *Multi-Lingual Interoperability in Speech Technology11, (RTO MP-28), Leusden, The Netherlands, 13–14/09/1999*, pages 119–124. RTO/NATO 2000

28. Pellegrino, F., Farinas, J., and André-Obrecht, R. (1999a). Comparison of Two Phonetic Aproaches to Language Identification. In Olaszy, G., Németh, G., and Erdohegyi, K., editors, *6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hongrie, 5–9/09/1999*, volume I, pages 399–402. International Speech Communication Association (ISCA)

5.1.5 Conférences et ateliers nationaux avec actes édités et comité de lecture

1. Gravellier, L., Coz, M. L., Farinas, J., and Pinquier, J. (2023a). Détection automatique de la déglutition dans les signaux d'auscultation cervicale à haute résolution. In *29ième Colloque sur le traitement du signal et des images*, pages 789–792, Grenoble. GRETSI - Groupe de Recherche en Traitement du Signal et des Images. `https://gretsi.fr/data/colloque/pdf/2023_gravellier1269.pdf` du 6 août au 9 Sept 2023

2. Vaysse, R., Astésano, C., and Farinas, J. (2023). Représentation automatique du rythme de la parole pathologique via le spectre de modulations d'amplitude. In de Recherche en Informatique de Toulouse, I., editor, *9èmes Journées de Phonétique Clinique (JPC 2023)*, pages 59–61, Toulouse, France. Université de Toulouse. ISBN: 978-2-917490-35-8

3. Gravellier, L., Le Coz, M., Farinas, J., Neveu, F., and Pinquier, J. (2023b). Étude des signaux vibroacoustiques de déglutition chez les sujets sains. 9èmes Journées de Phonétique Clinique (JPC 2023). Poster - ISBN: 978-2-917490-35-8

4. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023a). Analyse des performances de systèmes de reconnaissance automatique de la parole spontanée après cancer oral ou oropharyngé. 9èmes Journées de Phonétique Clinique (JPC 2023). Poster - ISBN: 978-2-917490-35-8

5. Vaysse, R., Ghio, A., Astésano, C., Farinas, J., and Viallet, F. (2022b). Analyse macroscopique des variations et modulations de F0 en lecture dans la maladie de Parkinson : données sur 320 locuteurs. In *Actes XXXIVe Journées d'Études sur la Parole (JEP2022)*, pages 307–315, Noirmoutier, France. Association Française de la Communication Parlée

6. Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2022c). Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures. In *Actes XXXIVe Journées d'Études sur la Parole (JEP2022)*, pages 117–125, Noirmoutier, France. Association Française de la Communication Parlée

7. Balaguer, M., Gelin, L., Woisard, V., Farinas, J., and Pinquier, J. (2021c). Mesure de l'intelligibilité après cancer oral ou oropharyngé par un système de reconnaissance automatique de la parole. In *1ère Journée Scientifique d'Orthophonie*, Congrès en ligne, France. SURO Société Universitaire de Recherche en Orthophonie

8. Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020c). Une nouvelle mesure de la réverbération pour prédire les performances a priori de la transcription de la parole. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).*, pages 226–234, Nancy, France. ATALA

9. Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020b). Analyse de l'effet de la réverbération sur la reconnaissance automatique de la parole. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, pages 235–243, Nancy, France. ATALA

10. Ferreira, S., Farinas, J., Pinquier, J., and Rabant, S. (2019a). Analyse du bruit pour la prédiction de la qualité de la transcription automatique de la parole (regular paper). In *Colloque sur le Traitement du Signal et des Images (GRETSI 2019), Lille, France, 26/08/19-29/08/19*

11. Farinas, J., Pellegrini, T., and Pinquier, J. (2019). Comparaison de systèmes automatiques de reconnaissance grand vocabulaire appliqué à de la parole pathologique (regular paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14/05/2019-16/05/2019*, pages 53–54. Centre international de Phonétique Appliquée (CIPA)

12. Woisard, V., Farinas, J., and Astesano, C. (2019b). Intelligibilité de la parole et qualité de vie. Réflexions à partir des résultats de l'étude « carcinologic speech severity index » (short paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14-16/05/2019*, pages 15–16. Centre international de Phonétique Appliquée (CIPA). (Conférencier invité)

13. Pinquier, J., Farinas, J., De Boissezon, X., Peran, P., Danet, L., and Jucla, M. (2019). EVOLEX : apport de la reconnaissance vocale pour le diagnostic des dysfonctionnements cognitifs légers (poster). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14–16/05/2019*, pages 105–106. Centre international de Phonétique Appliquée (CIPA)

14. Balaguer, M., Woisard, V., Farinas, J., and Pinquier, J. (2019c). Impact du trouble de la production de la parole sur les actes communicationnels de la vie quotidienne dans les cancers de la cavité buccale et de l'oropharynx

15. Pommée, T., Mauclair, J., Woisard, V., Farinas, J., and Pinquier, J. (2019). Génération de la « banane de la parole » en vue d'une évaluation objective de l'intelligibilité (regular paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Université de Mons, 14–16/05/2019*, pages 107–108. Centre international de Phonétique Appliquée (CIPA)

16. Balaguer, M., Boisguerin, A., Galtier, A., Puech, M., Farinas, J., Pinquier, J., and Woisard, V. (2019a). Facteurs influençant l'intelligibilité et la sévérité du trouble chronique de la parole des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. In *Journées de Phonétique Clinique (JPC 2019), Université de Mons, 14–16/05/2019*, pages 23–24. Centre international de Phonétique Appliquée (CIPA)

17. Ferreira, S., Farinas, J., Pinquier, J., and Rabant, S. (2018). Prédiction a priori de la qualité de la transcription automatique de la parole bruitée (regular paper). In *XXXIIe Journées d'Etudes sur la Parole (JEP 2018), Aix-En-Provence, France, 4–8/06/2018*, pages 249–257. Association Francophone de la Communication Parlée (AFCP)

18. Laaridh, I., Tardieu, J., Magnen, C., Gaillard, P., Farinas, J., and Pinquier, J. (2018b). Évaluations perceptive et automatique de l'intelligibilité de la parole dégradée par simulation de la surdité professionnelle. In *Journées d'Etudes sur la Parole (JEP 2018), Aix-en-Provence, 4-8/06/2018*, pages 392–400. Association Francophone de la Communication Parlée (AFCP)

19. Meignier, S., Merlin, T., Lévy, C., Larcher, A., Charton, E., Bonastre, J.-F., Besacier, L., Farinas, J., and Ravera, B. (2008). Mistral : plate-forme open source d'authentification biométrique. In *Journées d'Etudes sur la Parole (JEP 2008), Avignon, France, 9–13/06/2008*, pages 81–84. Association Francophone de la Communication Parlée (AFCP)

20. Arias, J. A., André-Obrecht, R., and Farinas, J. (2008b). Représentations de séquences de parole en espaces de faible dimensionalité. In *XXVIIèmes Journées d'Etudes sur la Parole (JEP 2008)*, pages 373–376, Avignon, France. Association Francophone de la Communication Parlée (AFCP)

21. Arias, J. A., André-Obrecht, R., Farinas, J., and Pinquier, J. (2006). Etude de la réduction non linéaire de la dimension du signal de parole en vue de modélisations discriminatives par SVM. In *Journées d'Etudes sur la Parole*, pages 77–80, Dinard, France. Association Francophone de la Communication Parlée (AFCP)

22. Gutierrez, J., Rouas, J.-L., Farinas, J., and André-Obrecht, R. (2004). Stratégies de fusion des décisions multiexpert en identification automatique des langues. In *Identification des langues et des variétés dialectales par les humains et par les machines - Modélisation pour l'identification des langues (MIDL)*, pages 71–76, Paris, France. ENST - Télécom Paris

23. Rouas, J.-L. and Farinas, J. (2004). Comparaison de méthodes de caractérisation du rythme des langues. In *Identification des langues et des variétés dialectales par les humains et par les machines - Modélisation pour l'identification des langues*, pages 45–50, Paris, France. ENST - Télécom Paris

24. Rouas, J.-L., Farinas, J., and Pellegrino, F. (2004). Evaluation automatique du débit de la parole sur des données multilingues spontanées. In *XXVe Journées d'Etude sur la Parole (JEP'2004), Fès, Maroc, 19–21/04/2004*, pages 437–440. Association Francophone de la Communication Parlée (AFCP)

25. Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2002b). Fusion de paramètres rythmiques et segmentaux pour l'identification automatique des langues. In *XXVIéme Journées d'Etude sur la Parole (JEP'2002), Nancy, 24–27/06/2002*, pages 105–108. Groupe Francophone de la Communication Parlée (GFCP)

26. Farinas, J. and André-Obrecht, R. (2001a). Modélisation phonotactique de grandes classes phonétiques en vue d'une approche différenciée en identification automatique des langues. In *XVIIIème colloque GRETSI sur le traitement du signal et des images*, Toulouse, France. Télécommunications Spatiales et aéronautiques (TéSA)

27. Pellegrino, F., Farinas, J., and André-Obrecht, R. (2000). Identification Automatique des Langues par une modélisation différenciée des systèmes vocaliques et consonantiques. In *XXIIème congrès francophone AFRIF-AFIA de la Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 131–139, Paris, France. AFRIF-AFIA

28. Farinas, J. and André-Obrecht, R. (2000). Identification Automatique des Langues : variations sur les multigrammes. In *XXIIIème Journées d'Etude sur la Parole (JEP)*, pages 373–376, Aussois, France. Groupe Francophone de la Communication Parlée (GFCP)

5.1.6 Conférences sans actes édités, communications orales ou par affiche

1. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023c). Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: preliminary validation. In *32nd World Congress of the IALP*, Auckland (Nouvelle Zelande), New Zealand. IALP International Association of Communication Sciences and Disorders
2. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023d). Measurement of communication impairment after treatment for oral and oropharyngeal cancer by automatic analyses of spontaneous speech associated with biopsychosocial factors. In *32nd World Congress of the IALP*, Auckland (Nouvelle Zelande), New Zealand. IALP International Association of Communication Sciences and Disorders
3. Farinas, J. (2023). Évaluation objective de la compréhensibilité d'un patient ayant des difficultés à s'exprimer. Institut Carnot Cognition `https://www.youtube.com/watch?v=_XOJDHBiWEM`
4. Farinas, J. (2022). Jérôme FARINAS, Maître de Conférence UT3 au département SI – Équipe SAMoVA, explique son travail de recherche sur le traitement automatique de la parole. Université Toulouse 3 `https://www.youtube.com/watch?v=7Aj8HUVAzgA`
5. Farinas, J., Pinquier, J., Ghio, A., and Petiot, J. (2022). Plateforme traitement de Parole Atypique. Evènement Kick-off de l'Institut Carnot Cognition 2022. Poster
6. Balaguer, M., Pinquier, J., Farinas, J., Lepage, B., and Woisard, V. (2021d). Construction d'un index holistique d'impact sur la communication des troubles de la parole chez des patients traités pour un cancer oral ou oropharyngé. 15ème conférence francophone d'EPIdémiologie CLINique (EPICLIN 2021) et les 28èmes Journées des Statisticiens des Centres de Lutte Contre le Cancer
7. Petiot, J., Gravellier, L., Jucla, M., Monnier, N., Quillion-Dupre, L., Péran, P., Danet, L., De Boissezon, X., Farinas, J., and Pinquier, J. (2021). EVOLEX : la reconnaissance vocale au service du diagnostic des dysfonctionnements langagiers. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021
8. Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2021e). Construction d'un index holistique d'impact sur la communication des troubles de la parole chez des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021
9. Vaysse, R., Astesano, C., and Farinas, J. (2021a). Analyse des performances des algorithmes d'estimation de la fréquence fondamentale dans le cadre de la voix pathologique. Séminaire AFCP – Phonétique Clinique. Poster
10. Gallois, Y., Farinas, J., Paugam, M.-W., Nicolini, L., and Woisard, V. (2021). L'identification automatique des différents bruits de gorge chez le sujet sain : une étude pilote. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021
11. Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2021). Une méthode automatique non supervisée pour évaluer le score de sévérité de la parole chez les patients traités pour un cancer ORL. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021
12. Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020a). A new set of superimposed speech features to predict a priori the performance of automatic speech recognition systems. Speech In Noise Workshop, Toulouse, France, 10–11/01/2020
13. Fontan, L., Farinas, J., Segura, B., Stone, M., and Füllgrabe, C. (2020). Using automatic speech recognition to predict aided speech-in-noise intelligibility. Speech In Noise Workshop, Toulouse, France, 10-11/01/2020

14. Fontan, L., Laaridh, I., Farinas, J., Pinquier, J., Le Coz, M., and Füllgrabe, C. (2019). Using automatic speech recognition for the prediction of impaired speech identification. Speech In Noise Workshop
15. Ferreira, S., Pinquier, J., Farinas, J., Mauclair, J., and Rabant, S. (2019b). A new measure to predict the a priori performance of automatic transcription systems on reverberated speech. Speech In Noise Workshop, Ghent, Belgique, 10-11/01/2019
16. Farinas, J. and Pellegrini, T. (2018). Analysis of Data Provided for the 2018 AIRBUS Air Traffic Control Challenge (2018 AIRBUS Air Traffic Control Challenge Workshop, Toulouse, France, 04/10/2018). https://www.irit.fr/recherches/SAMOVA/pagechallenge-airbus-atc-workshop.html
17. Fontan, L., Pellegrini, T., Farinas, J., Mauclair, J., Laborde, V., Sahraoui, H., Aumont, X., Olcoz, J., and Abad, A. (2015c). Vers des outils automatiques pour l'évaluation de locuteurs atypiques. Colloque Interphonologie du Français Contemporain Evaluation de la parole non native et corpus oraux, Paris, 08/12/2015
18. Bredin, H., Koenig, L., and Farinas, J. (2010). IRIT TRECVid 2010: Hidden Markov Models for Context-aware Late Fusion of Multiple Audio Classifiers. TREC Video Retrieval Evaluation, Gaithersburg, MD, USA, 15–17/11/2010
19. Sanchez-Soto, E. and Farinas, J. (2007). Irit system description for Nist 2007 Language Recognition Evaluation. NIST evaluation workshop, Orlando, USA, 11–12/12/2007
20. de Calmès, M., Farinas, J., Ferrané, I., and Pinquier, J. (2005). Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire. Atelier ESTER, Avignon, 30–31/03/2005
21. Farinas, J., Pellegrino, F., and André-Obrecht, R. (2001). Automatic Language Identification: From a Phonetic Differentiated Model to a Complete System. 5th workshop on Electronics, Control, Modelling, Measurement and Signals, Toulouse, 30/05–1/06/2001
22. Farinas, J. (2001). A differentiated approach and prosody improvements in automatic language identification. Student Forum, IEEE Signal Processing Society, International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2001), Salt Lake City, Etats-Unis, 7–11/05/2001
23. Farinas, J. (1999). Comment modéliser la prosodie dans un but d'identification des langues ? Rencontres Jeunes Chercheurs en Parole (RJC Parole'99), Avignon, 18–19/11/1999

5.1.7 Conférencier invité

1. Farinas, J., Astesano, C., and Vaysse, R. (2023). Caractérisation automatique du rythme de la parole. In *Journée scientifique de Toulouse Mind and Brain Institute (TMBI 2023)*, Toulouse, France. Toulouse Mind and Brain Institute
2. Farinas, J. (2021). Evaluation automatique de l'intelligibilité pour des patients présentant une atteinte de la voix. In *Premier webinaire de Start in Lab Santé 2021*, Toulouse, France. Digital 113
3. Ghio, A. and Farinas, J. (2021). La laryngophoniatrie du futur. In *Congrès National de la Société Française d'ORL*, Paris, France
4. Woisard, V., Farinas, J., Ahmed, B., and Wren, Y. (2019a). Place of Automatic Speech Recognition for Assessing Speech Disorders. Présentation orale. Educational Committee for Phoniatrics Committee at International Association of Logopedics and Phoniatrics (IALP)
5. Farinas, J. (2018). Voice and Speech Perception: Human vs Automatic: How can Computer Speech Recognition be used? Tutoriel. 29th Congress of Union of The European Phoniatricians, Helsinki, Finland

6. De Boissezon, X., Danet, L., Fabre, C., Farinas, J., Gaume, B., Hathout, N., Ho-Dac, L.-M., Jucla, M., Peran, P., Pierrejean, B., Pinquier, J., and Tanguy, L. (2018). Le projet EvoLex : Aller plus loin dans l'étude de la fluence et de l'accès au lexique. Qui-Quoi-Où de la recherche sur langage, culture et société à Toulouse, CHU Purpan, Pavillon Baudot, 14/05/2018
7. Woisard, V., Astesano, C., and Farinas, J. (2018). Carcinologic Speech Severity Index Project: A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer. Qui-Quoi-Où de la recherche sur langage, culture et société à Toulouse, CHU Purpan, pavillon Baudot, 14/05/2018
8. Farinas, J. and Fredouille, C. (2017). La place du traitement automatique de la parole (Journée d'Etudes "Intelligibilité de la parole", Maison de la Recherche, D29, 24/03/2017). (Conférencier invité)
9. Farinas, J. (2017). GIS Parolothèque : une plateforme de recherche sur des données médicales de parole (Infrastructure pour l'expérimentation en ligne, la science participative/collaborative ou la recherche inter-disciplinaire par les données (InfraMed), Toulouse, France, 04/12/2017). (Conférencier invité)
10. Farinas, J. (2016). IRIT, un acteur de la recherche en Santé et Autonomie, Axe stratégique "Systèmes Informatiques pour la Santé et l'Autonomie", composante gestion de données (Université d'été de la e-santé, Castres, France, 05/07/2016). (Conférencier invité)
11. Farinas, J. (2008b). Structuration de documents Audio Vidéo. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 24-26/09/2008
12. Farinas, J. (2008a). Reconnaissance Automatique de Langues. Tutoriel. Ecole Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 24-26/09/2008
13. Farinas, J. (2007). Reconnaissance Automatique de Langues. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 4-6 septembre 2007
14. Farinas, J. (2006). Identification et Classification Automatique de Langues. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 4-6 sept. 2006

5.1.8 Thèses et habilitations

1. Farinas, J. (2002). *Une modélisation automatique du rythme pour l'identification des langues* . Thèse de doctorat, Université Paul Sabatier, Toulouse, France. (Soutenance le 15/11/2002)

5.2 Dépots logiciels et déclarations d'invention

1. Décodeurs Acoustico-Phonétiques multilingue (DAPm)[5]. Il s'agit du dépot de six décodeurs acoustico-phonétique bande étroite (anglais, allemand, hindi, japonais, mandarin et espagnol) et un bande large (français). Les décodeurs en bandes étroites ont été appris sur le corpus Oregon Graduate Institute-Multi Language Telephone Speech [Muthusamy et al., 1992]. La bande large est issue du système appris sur la campagne d'évaluation ESTER [de Calmès et al., 2005]. Les modèles sont sous licence libre GNU GPL 3.
2. Analyse prosodique et lexicale : déclaration d'invention datant du 26 août 2014 et concernant les techniques à la base du contrat de maturation Toulouse Tech Transfert-Authôt (cf. §5.6.1 page 89). Les inventeurs sont Julien Pinquier (33%), Thomas Pellegrini (33%) et moi-même (33%). Mes contributions concernent les modélisations automatiques de la prosodie que j'ai produit depuis le doctorat.

5. https://www.irit.fr/recherches/SAMOVA/pagedap.html

3. PROSODIE [6] : dépôt logiciel à l'Agence de Protection des Programmes par Toulouse Tech Transfert le 30 décembre 2014 concernant Thomas Pellegrini, Julien Pinquier et Jérôme Farinas. Les logiciels correspondent à mes programmes de modélisation prosodique, aux programmes de segmentation acoustique de Julien et aux programmes de modélisation lexicale de Thomas. Logiciel de ponctuation automatique licencié à la société Authôt le 3 juillet 2015 (version française) et le 10 février 2016 (avenant concernant une version anglaise des modèles).

4. EVOLEX : dépôt logiciel datant de 2018 (dépôt à l'Agence pour la protection des Programmes IDDN.FR.001.300011.000.S.P.2018.000.10000) et concrétisant la collaboration entre Xavier de Boissezon (spécialiste en médecine physique et de réadaptation à l'hôpital de Rangueil CHU de Toulouse) et Julien Pinquier, Maxime Le Coz et moi-même de l'équipe SAMOVA de l'IRIT. Nous avons réalisé une platefome expérimentale permettant d'automatiser et d'analyser les passations d'évocation lexicale, de fluence verbale et de dénomination réalisés par les orthophonistes et les médecins.

5. SAMI : dépôt datant du 15 novembre 2021, concernant le logiciel Système Automatique de Mesure d'Intelligibilité (SAMI), numéros TTT : I-21-2485 et CNRS : 15505-01. Les inventeurs sont : Julien Pinquier, Vincent Roger, Jérôme Farinas, Gauthier Arcin et Virginie Woisard. Il s'agit d'un logiciel d'analyse audio à base d'IA mesurant l'intelligibilité pour le suivi et la rééducation de patients atteints de maladie(s) affectant la voix, basé sur une mesure d'intelligibilité. Le logiciel a été développé dans le cadre d'un projet de pré-amturation ProtoPITCH de Toulouse Tech Transfert. Il s'agit d'un logiciel sur tablette mobile permettant de faire de l'acquisition et de la lecture de résultats, et d'un serveur réalisant le traitement des requêtes et le calcul des mesures d'intelligibilité.

5.3 Prix et distinctions

1. Concours « La fabrique » AVIVA 2016

   Jérôme Farinas, Isabelle Ferrané et Julien Pinquier avec société Archean Technologie ont été les grands gagnants du concours La Fabrique lancé par AVIVA en 2016 [7] pour la technologie CarlW, système d'aide au réglage de prothèses auditives développé suite aux projet AGILE IT et PHONICS (cf. §5.4.9).

   CarlW a été grand gagnant de la catégorie « Agir pour une santé durable » et a obtenu une aide financière de 50 000 € [8].

   Le concours AVIVA a rassemblé plus de 200 participants. En avril 2017 il a été annoncé que 64 projets ont été sélectionnés dans ce concours et ont reçu des prix allant de 5 000 à 50 000 €(4 grands gagnants).

5.4 Projets de recherche

La table 4 résume mes participations aux différents projets. Le détail des projets ainsi que mon implication est détaillé en suivant.

6. http://www.cnrs.fr/lettre-innovation/actus.php?numero=314
7. https://www.irit.fr/CarlIW-un-systeme-de-mesure-de-la
8. https://agence-api.ouest-france.fr/uploads/article/63f4a80f687e632e836d9d13ca35c6e615311e0
pdf

Table 4 – Responsabilités dans des projets de recherche

Fonction	Européen	National	Régional
Porteur du projet		2	
Responsabilité scientifique		6	2
Participation	3	13	4

5.4.1 Projet RNRT Global Architecture for Search and Indexing (AGIR)

Projet pré-compétitif RNRT de 1999 à 2001 qui a réuni l'Institut National de l'Audiovisuel, la Compagnie des Signaux, la société Arts Video, la société MEMODATA, l'IRISA, l'INRIA Rhône-Alpes, l'Université Joseph Fourier Grenoble-Alpes, le LiP6 et l'INT Evry.

5.4.2 Projet CNRS APN SHS Language Identification Afro-asiatic languages and dialects

Projet APN SHS du CNRS réunissant le laboratoire Dynamique du Langage à Lyon, l'Institut des Sciences Cognitives et l'University of California Berkeley et l'IRIT de 2000 à 2002 sur la reconnaissance automatique des langues pour les langages et dialectes afro-asiatiques.

5.4.3 Projet région Rhône-Alpes Emergence 2000 Language Identification on European languages

Projet Rhône-Alpes Region Emergence 2000 avec le laboratoire Dynamique du Langage à Lyon, l'Institut des Sciences Cognitives, l'University of California Berkeley et l'IRIT entre 2000 et 2003 sur l'identification des langues européennes.

Cette collaboration a permis de travailler la modélisation du rythme [Pellegrino et al., 2002].

5.4.4 Projet CNRS Automatic Retrieval of Audio and Speech Informations (RAIVES)

Le projet RAIVES est un projet CNRS qui a réuni le laboratoire Dynamique du Langage, le laboratoire LORIA de l'INRIA et l'équipe SAMOVA de l'IRIT de 2002 à 2003. Faisant le constat que les documents sonores font en 2002 partie de ce que l'on appelle le « web invisible », le projet a pour objectif une structuration de ces documents sonores, en particulier radiophoniques, à partir de l'indexation par leur contenu, de manière à leur donner un sens du point de vue d'un utilisateur du web, et de produire à partir de ces documents des connaissances exploitables. Ce contenu pourrait alors être accessible aux moteurs de recherche et devenir disponible aux internautes au même titre que le contenu textuel de pages HTML.

Le travail a porté sur les descripteurs du contenu d'un document radiophonique suivant :

- distinction des segments de Parole/Musique,
- détection de « sons clés »,
- identification de la langue,
- découpage en locuteurs associé à une éventuelle identification de ces locuteurs,
- détection de mots clés,
- identification de thèmes.

Mon implication a plus particulièrement porté sur l'identification des langues sur les média radiophoniques [Parlangeau et al., 2003].

5.4.5 Projet ITEA KLIMT

Le projet ITEA KLIMT de 2002 à 2004 a eu pour objectif de répondre au besoin d'intelligence dans un environnement multimédia distribué. Il en a résulté la livraison d'une plate-forme de traitement de contenu multimédia avec une architecture orientée services.

Ce projet a réuni de nombreux partenaires :

— Français : THALES, LiP6, 4ème Millénaire, Isoft, Sinequa, UMPC, Eurécom, IRIT ;
— Italiens : Zanussi, Softeco Sismat, TTS, Politecnico di Milano ;
— Espagnols : Meta4, I&IMS, B-kin, ESI.

Les livrables me concernant se basaient sur la reconnaissance automatique des langues.

5.4.6 Projet ANR Technologies Logicielles 2006 MISTRAL

Projet ANR appel Technologies Logicielles de 2006 réunissant le Laboratoire d'Informatique d'Avignon (porteur du projet), Thales, le Laboratoire d'informatique du Mans (LIUM), la société Calistel, le CLIPS/IMAG, EURECOM et l'IRIT de 2007 à 2009.

Le projet MISTRAL propose un cadre « open source » pour l'authentification biométrique. Le premier objectif de MISTRAL est de faciliter l'accès aux technologies biométriques à la fois pour les chercheurs universitaires, les enseignants et pour les ingénieurs des entreprises, en leur fournissant une boîte à outils logicielle complète, puissante, flexible, facile à comprendre et capable d'utiliser la majorité des applications biométriques sur un large éventail d'environnements matériels. Le deuxième objectif de MISTRAL est de promouvoir les recherches et les applications biométriques en organisant les communications et les échanges au sein d'un large éventail d'utilisateurs de logiciels MISTRAL, composé d'acteurs issus du monde universitaire et industriel.

J'ai eu en charge l'adaptation de la plateforme ALIZE/MISTRAL, spécialisée à l'origine sur la reconnaissance du locuteur, à l'identification automatique des langues. Eduardo Soto Sanchez (cf. §6.3.1 page 105) a été recruté sur ce projet pour mener à bien cette mission. Ce travail a débouché par une participation à la campagne d'évaluation NIST LRE 2007 [Sanchez-Soto and Farinas, 2007] (cf. §5.5.3).

5.4.7 Projet ANR Masses de Données Connaissances Ambiantes EPAC

Le projet ANR 2006 Masses de Données Connaissances Ambiantes EPAC a rassemblé l'équipe parole du laboratoire d'informatique du Mans (LIUM), l'équipe TAP du laboratoire d'informatique d'Avignon (LIA), l'équipe BDTLN du laboratoire d'informatique de Tours (LI) et l'équipe SAMOVA de l'IRIT de 2007 à 2010.

Le but du projet EPAC est de proposer des méthodes d'extraction d'information et de structuration et de signature de documents qui seraient spécifiées aux données audio et appliquées à un vaste ensemble de documents audio (1930 heures de données enregistrées). Ces méthodes ont pris en compte tous les canaux d'information présents dans ces données et les ont analysés à partir de différents niveaux de granularité : segmentation du signal (parole/musique/jingle), caractérisation de l'environnement (foule, interview, débat…), identification et suivi du locuteur, transcription vocale, détection et suivi des sujets, analyse vocale, interactions conversationnelles, opinion des orateurs, etc.

J'ai été responsable scientifique de l'IRIT pour le projet. Nous avons travaillé sur les segmentations acoustiques parole/musique/bruit, la segmentation en locuteurs et la détection de thème et participé à la production du corpus de parole spontannée dérivé d'ESTER [Estève et al., 2011].

5.4.8 Projet ANR Défis 2008 ARTIS

Le projet ANR Articulatory inversion from audio-visual speech for augmented speech presentation (ARTIS), appel Défis 2008 a été réalisé en collaboration avec LTCI (Paris), Gipsa-Lab (Grenoble) et l'IRIT de 2009 à 2012.

J'ai travaillé avec Hélène Lachambre et Régine André Obrecht pour proposer une inversion articulatoire de la parole à partir de modèles de Makov cachés sous HTK Toolkit [Young et al., 2002].

5.4.9 Projet Région AGILE IT Équipement électronique intelligent de mesure de compréhension de la parole

Le projet régional AGILE IT 2012 a rassemble la Société Archean Technologies, la Maison des Sciences de l'Homme et de la Société de Toulouse et l'équipe SAMOVA de l'IRIT de 2013 à 2015.

Ce projet a visé la conception et la réalisation d'un équipement électronique intelligent pour la mesure de la compréhension de la parole à destination des audioprothésistes en France et à l'exportation. En se calquant sur le fonctionnement perceptif et cognitif humain, cet équipement a permis d'évaluer automatiquement les performances de prothèses auditives, afin de procéder à un réglage optimal en fonction du profil pathologique de la personne à appareiller.

L'objectif est de réduire la dépendance et d'accroître l'autonomie des personnes appareillables attentes de surdité partielle, par :

- un accroissement de l'acceptabilité des prothèses auditives par les patients ;
- un enrichissement de l'ergonomie d'adaptation ;
- un guidage de l'audioprothésiste pour l'augmentation du confort de la personne, en particulier dans la phase initiale de réglage et de contrôle de la prothèse auditive ;
- une réduction des freins actuels à l'équipement.

J'ai été responsable scientifique pour l'IRIT sur ce projet. Nous avons participé à la conception du protocole expérimental et du corpus. Nous avons mis en place une mesure de l'intelligibilité automatique basée sur la perception humaine [Fontan et al., 2015b, Fontan et al., 2015a, Fontan et al., 2015c, Fontan et al., 2016, Fontan et al., 2017].

5.4.10 Projet CLE PHONICS

Le projet PHONICS a prolongé la collaboration entre la Maison des Sciences de l'Homme, la société de Toulouse, Archean Technologies et de la Société de Toulouse et l'équipe SAMOVA de l'IRIT de 2015 à 2018.

PHONICS a visé à étendre le champ d'application d'un outil développé dans le cadre d'un projet AGILE-IT, impliquant les mêmes partenaires. Ce projet a permis de développer une solution de préréglage automatique des aides auditives afin d'optimiser et de garantir leur utilisation par les patients souffrant de presbyacousie. Ce système réduira la durée de la phase d'adaptation aux aides auditives en évaluant les confusions dans l'intelligibilité et la compréhension de la parole dues à la presbyacousie.

Le premier objectif était d'étendre l'utilisation de l'outil AGILE-IT à l'anglais américain. Pour ce faire, les tests d'intelligibilité et de compréhension développés initialement pour le français avec des auditeurs anglophones ont été dupliqués en créant du matériel linguistique adapté aux caractéristiques de l'anglais [Laaridh et al., 2018a].

Le deuxième objectif était d'appliquer l'outil de mesure des compétences à la surdité professionnelle. Les caractéristiques physiologiques de la surdité professionnelle sont différentes de celles de la presbyacousie. En particulier, il existe des zones mortes cochléaires, c'est-à-dire que la détérioration de certaines parties des cellules ciliées empêche le codage tonotopique. Ces zones mortes sont généralement accompagnées d'acouphènes ainsi que d'un phénomène d'hyperacousie créant des problèmes pour

comprendre la parole en silence et dans le bruit. Concernant ces symptômes, ils se traduisent par un audiogramme montrant une perte neurosensorielle dans les hautes fréquences et une encoche localisée dans la région des 4 kHz. Le profil d'audiogramme dépend du traumatisme et évolue dans le temps, en particulier les fréquences voisines de la région la plus touchée peuvent être affectées progressivement. Pour faire face à ce nouveau problème, le protocole défini pour la mesure de la compréhension dans le cas de la presbyacousie sera reproduit en intégrant les caractéristiques des surdités par traumatisme en milieu professionnel. Par conséquent, la simulation de ces types de surdité a été adaptée en fonction des nouveaux paramètres acoustiques impliqués. Les tests d'intelligibilité et de compréhension ont été proposés avec cette nouvelle simulation afin d'obtenir des scores correspondant à cette surdité qui seront ensuite soumis à un système de reconnaissance vocale [Laaridh et al., 2018b].

Le dernier objectif consistait à envisager les transpositions de l'application de l'outil de mesure de la compréhension à l'évaluation de la réadaptation des voix pathologiques.

5.4.11 Projet Action on Hearing Loss

Le projet « Using automatic speech recognition to predict speech-in-noise perception for simulated age-related hearing loss » est financé par « Action on hearing Loss » et établit une collaboration entre Christian Füllgrabe (Université de Nottingham, Royaume Unis, coordinateur scientifique), Lionel Fontan (Archean Labs, Montauban) et Jérôme Farinas entre 2017 et 2019.

Action on Hearing Loss est une action du Royal National Institute for Deaf People : un organisme de bienfaisance enregistré en Angleterre et au Pays de Galles (207720) et en Écosse (SC038926). Enregistrée en tant que société caritative à responsabilité limitée par garantie en Angleterre et au Pays de Galles sous le numéro 454169.

Le but du projet proposé était d'établir une base de données de référence comportementale pour les auditeurs de langue maternelle anglaise, à laquelle comparer les prédictions des machines. Contrairement aux travaux précédents, toutes les expériences mesurent la performance non seulement dans le silence, mais aussi en présence de différents bruits de fond (bruit de parole par rapport au bruit de babillage) présentés selon une gamme de rapports signal/bruit. L'enregistrement des stimuli de test et les passations des participants ont été effectués à l'Institut de Recherche sur l'Audition (IHR) à Nottingham tandis que le développement, l'apprentissage et les tests des systèmes de reconnaissance automatique de la parole seront effectués à l'IRIT.

J'ai été en charge de la conception des modèles acoustiques bruités à partir du corpus ESTER. Ces modèles seront utilisés pour les analyses des stimuli en conditions bruitées.

5.4.12 Projet INCA Carcinologic Speech Severity Index

Ce projet financé par l'Institut National du Cancer a réuni l'unité Voix et Déglutition du CHU de Toulouse, le laboratoire LPL d'Aix-en-provence, la Maison des Sciences de l'Homme, la société de Toulouse, le laboratoire Octogone-Lordat de l'Université Jean Jaures, le laboratorie d'informatique d'Avignon, l'unité de soutien méthodologique à la recherche Service d'Epidémiologie de l'université Paul Sabatier et l'équipe SAMOVA de Toulouse de 2014 à 2018. J'ai été responsable scientifique pour l'IRIT.

La diminution de la mortalité par cancer de la tête et du cou (HNC) souligne l'importance de réduire l'impact sur la qualité de vie (QoL). Néanmoins, les outils habituels d'évaluation de la qualité de vie ne sont pas pertinents pour mesurer l'impact du traitement sur les principales fonctions impliquées par les séquelles. Il manque des outils validés pour mesurer les résultats fonctionnels du traitement carcinologique, en particulier pour les troubles de la parole. Certaines évaluations sont disponibles pour les troubles de la voix dans le cancer du larynx, mais elles sont basées sur de très mauvais outils pour

les cancers buccaux et pharyngés impliquant plus de problèmes d'articulation vocale que de problèmes de voix.

Dans ce contexte, il a été proposé de développer un indice de gravité des troubles de la parole pour décrire l'impact des protocoles thérapeutiques dans le but de compléter les taux de survie.

L'intelligibilité de la parole est la façon habituelle de quantifier la gravité des troubles neurologiques de la parole. Mais cette mesure n'est pas valable dans la pratique clinique en raison de plusieurs difficultés comme l'« effet de familiarité » et la faible reproductibilité entre juges. De plus, les scores d'intelligibilité de la transcription ne reflètent pas exactement la compréhension de l'auditeur.

Il est reconnu qu'une évaluation objective et impartiale du déficit de communication causé par un trouble de la parole nécessite des outils de traitement automatique de la parole. L'idée est de collecter des enregistrements audio de la parole du patient et de calculer un score d'intelligibilité des énoncés. En 2012, Middag et collègues ont présenté une nouvelle méthode de prédiction de l'intelligibilité de la parole d'une manière robuste contre les changements dans le texte et contre les différences d'accent des néerlandophones applicables aux patients traités pour un cancer ORL [Middag et al., 2014].

Par conséquent, l'hypothèse est qu'une technique d'évaluation automatique peut mesurer l'impact des troubles de la parole sur les capacités de communication en donnant un indice de gravité dans la production de la parole des patients traités pour un cancer de la tête et du cou et en particulier pour un cancer buccal et pharyngien. Cet indice a été appelé : l'indice de gravité de la parole carcinologique (C2SI).

J'ai participé dans ce projet à la définition du protocole expérimental. J'ai co-encadré Brendan Gloinec (cf. §10 page 93) sur une analyse systématique des paramétrisation acoustiques à utiliser pour discriminer au mieux les atteintes des patients. J'ai également encadré Oriol Pont (cf. §6.3.6 page 105) sur l'essai de paramétrisation non linéaire afin d'affiner la discrimination des patients. J'ai également encadré Mathieu Balaguer (cf. §13 page 93) pour la finalisation du score en utilisant toutes les productions automatiques.

5.4.13 Projet ANR Voice4PD-MSA

Le projet ANR « Diagnostic différentiel entre la maladie de Parkinson et l'atrophie multi-systématisé par analyse numérique de la parole » (Voice4PD-MSA) rassemble l'INRIA Bordeaux, le CHU de Bordeaux, le CHU de Toulouse, l'Institut de Mathématique de Toulouse et l'IRIT de 2016 à 2020.

La maladie de Parkinson (PD) et l'atrophie multi-systématisée (AMS) sont des maladies neurodégénératives. Cette dernière fait partie du groupe des maladies parkinsoniennes atypiques et a un pronostic défavorable. Aux premiers stades de la maladie, les symptômes de la PD et de l'AMS sont très semblables, particulièrement chez les patients atteints d'ASM-P où le parkinsonisme prédomine. Le diagnostic différentiel entre l'AMS-P et la PD peut être très difficile aux premiers stades de la maladie, alors que la certitude du diagnostic précoce est importante pour le patient en raison des pronostics divergents. En effet, malgré des efforts récents, aucun marqueur objectif validé n'est actuellement disponible pour guider le clinicien dans ce diagnostic différentiel. Le besoin de tels marqueurs est donc très élevé dans la communauté neurologique, compte tenu notamment de la gravité du pronostic de l'AMS-P.

L'objectif novateur du projet est de développer un marqueur numérique objectif non invasif pour faciliter le diagnostic différentiel précoce entre la PD et l'AMS-P, et le diagnostic de la PD par rapport aux témoins sains (HC). L'ambition est de développer un outil numérique portable et à très bas coût. En cas de succès, cela permettrait une utilisation à grande échelle et dans le monde entier ; chaque neurologue aurait la possibilité d'utiliser le marqueur numérique dans sa clinique ou son cabinet privé. L'ambition et l'originalité de ce projet est donc de développer un outil vocal numérique pour la discrimination objective entre les différentes conditions.

J'ai contribué à la définition du protocole expérimental et à l'encadrement, conjoint avec les collègues de l'Institut de Mathématiques de Toulouse, de Robin Vaysse (cf. §14 page 93) qui a réalisé une étude préliminaire sur ce projet.

5.4.14 Projet H2020-MSCA-ITN-ETN TAPAS

Projet européen de formation doctorale initiale programme Horizon 2020 Marie Sklodowska Curie (MSCA-ITN-ETN) de 2017 à 2021 qui regroupe plusieurs partenaires bénéficiaires :

- Institut de recherche IDIAP (Suisse)
- Université Friedrich-Alexander Erlangen Nuernberg (Allemagne)
- Université de Ghent IMEC (Belgique)
- INESC Lisbonne (Portugal)
- Ludwig-Maximilians-Universität Muenchen (Allemagne)
- The Netherlands Cancer Institute - Antoni van Leeuwenhoek (Pays Bas)
- Philips Research Eindhoven (Pays Bas)
- Radboud Universiteit, Nijmegen (Pays Bas)
- Universität Augsburg (Allemagne)
- Université Toulouse III-Paul Sabatier (France)
- University of Sheffield (Angleterre)
- Hopital Universitaire d'Antwerp (Belgique)

De plus en plus de personnes en Europe souffrent de troubles débilitants de la parole (par exemple, à cause d'un accident vasculaire cérébral, de la maladie de Parkinson, etc.). Ces groupes sont confrontés à des problèmes de communication qui peuvent conduire à l'exclusion sociale. Ils sont de plus en plus marginalisés par une nouvelle vague de technologie vocale qui est de plus en plus intégrée dans la vie quotidienne mais qui n'est pas résistante aux discours atypiques. TAPAS est un projet qui vise à transformer le bien-être de ces personnes.

Le programme de travail TAPAS cible trois problèmes de recherche clés :

- Détection : développer des techniques de traitement de la parole pour la détection précoce des conditions qui ont un impact sur la production vocale. Les résultats seront des outils diagnostiques peu coûteux et non invasifs qui permettront de détecter rapidement l'apparition de maladies évolutives comme la maladie d'Alzheimer et la maladie de Parkinson.
- Thérapie : utiliser des techniques de traitement de la parole nouvellement apparues pour produire des outils d'orthophonie automatisés. Ces outils rendront la thérapie plus accessible et plus ciblée. Une meilleure thérapie peut augmenter les chances de recouvrer la parole intelligible après des événements traumatisants comme un accident vasculaire cérébral ou une chirurgie buccale.
- Assistance à la vie autonome : redessiner la technologie vocale actuelle afin qu'elle fonctionne bien pour les personnes ayant des troubles de la parole et qu'elle aide également à faire des choix cliniques éclairés. Les personnes atteintes de troubles de la parole ont souvent d'autres affections concomitantes qui les rendent dépendantes des soignants. Les outils d'aide à la vie assistée par la parole sont un moyen de permettre à ces personnes de vivre de façon plus autonome.

TAPAS adopte une approche interdisciplinaire et multisectorielle. Le consortium comprend des cliniciens praticiens, des chercheurs universitaires et des partenaires industriels possédant une expertise dans les domaines de l'ingénierie de la parole, de la linguistique et des sciences cliniques. Tous les membres ont une expertise dans certains éléments de la parole pathologique. Ce riche réseau formera une nouvelle génération de 15 chercheurs, en les dotant des compétences et des ressources nécessaires à un succès durable.

J'ai participé au début de la formation des deux doctorants Eugenia Rikova et Timothy Pommée [Pommée et al., 2019] qui ont été accueilli par le CHU de Toulouse et l'IRIT.

5.4.15 Projet ANR RUGBI

Le projet multidisciplinaire ANR « Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders » (ANR-18-CE45-008 RUGBI) rassemble l'unité Voix et Déglutition du CHU de Toulouse, le laboratoire Octogone-Lordat de l'Unievrsité Jean Jaures, le laboratoire d'Informatique d'Avignon (LIA) de l'Université d'Avignon et des Pays de Vaucluse, le laboratoire Parole et Langage d'Aix-en-Provence (LPL) et l'équipe SAMOVA de l'IRIT de 2019 à 2022. Il s'agit d'un projet répondant à l'appel générique 2018 de la section CE45 (Mathématique, informatique, automatique, traitement du signal pour répondre aux défis de la biologie et de la santé) qui a été financé grâce au plan sur l'Intelligence Articifielle.

Dans le cadre des troubles de la production de la parole observés dans les cancers ORL, les pathologies neurologiques, sensorielles ou structurelles, l'objectif du projet RUGBI est d'améliorer la mesure du déficit en intelligibilité. En effet, les troubles de la production de la parole peuvent entraîner une grave perte d'intelligibilité, rendant difficile pour les patients la communication avec leur entourage et limitant leur vie professionnelle et/ou sociale. Classiquement, l'évaluation clinique de l'intelligibilité repose sur une évaluation perceptuelle globale jugée insatisfaisante par sa subjectivité, son manque de précision et sa durée qui conduisent à des mesures erronées de l'intelligibilité du patient. Par ailleurs, les tâches de production vocale dédiées à ce type d'évaluation (répétition de mots, phrases, lecture) sont loin d'être adaptées à une mesure précise de l'intelligibilité et ne permettent qu'une évaluation globale de la déficience fonctionnelle vocale. Le projet RUGBI propose de surmonter ces limites en développant un nouvel outil d'évaluation objectif basé sur :

- l'identification des unités linguistiques pertinentes d'un point de vue acoustique et prosodique,
- l'identification des tâches linguistiques sensibles.

L'objectif du projet RUGBI est donc de compléter les outils du thérapeute par une mesure précise, robuste et rapide permettant de développer un projet thérapeutique optimisé en vue d'une amélioration tangible de l'intelligibilité. Pour cela, le projet RUGBI s'appuie sur des corpus importants et des productions vocales déjà disponibles de sujets sains (190) et de patients (365) présentant des pathologies d'origine structurelle (cancers ORL, données issues du projet C2SI, cf. §5.4.12) et neurologique (maladie de Parkinson, données issues de la base de donnée Park360 hébergé sur Speedi-DB [9] de patients parkinsoniens [Ghio et al., 2006, Ghio et al., 2012]), dans l'exécution de différentes tâches linguistiques, et pour une partie d'entre eux, à différents stades de la maladie. Ces corpus sont un atout considérable pour la conduite des deux domaines d'étude du projet, respectivement basés sur :

- la perception de l'intelligibilité de la parole,
- la modélisation automatique du traitement de la parole, et plus particulièrement, sur l'apprentissage profond et ses propriétés de représentation des données qui devront être exploitées ici.

Dans ce contexte, l'objectif central du projet rassemble l'expertise de ses membres du domaine médical, du domaine des sciences du langage et de l'ingénierie de la parole et du langage pour relever les défis de la biologie et de la santé.

Je suis porteur de ce projet. Les contributions de l'IRIT se feront à travers le doctorat co-encadrée entre Octogone-Lordat et l'IRIT de Robin Vaysse sur les modélisatins prosodiques (cf. §6.2.5 page 101) ; le doctorat de Mathieu Balaguer co-encadrée par le CHU de Toulouse et l'IRIT (cf. §6.2.5 page 101) et le doctorat de Vincent Roger (cf. §6.2.4 page 100) sur la modélisation de l'intelligibilité.

9. https://speedi-db.lpl-aix.fr/physio/info.php

5.4.16 Projet FEDER EVOLEX

Le projet régional sous fonds FEDER EVOLEX s'inscrit dans la Fédération Hospitalo-Universitaire (FHU) des Handicaps Cognitifs, Psychiques et Sensoriels (HoPeS) qui réunit des acteurs du CHU de Toulouse (Institut des Handicaps Neurologiques, Psychiatriques et Sensoriels), de l'université de Toulouse et des instituts de recherche (Inserm, CNRS). Il représente l'étape nécessaire et indispensable pour optimiser et valider l'utilisation d'un logiciel adapté à une pratique clinique et de recherche et permettant l'administration de tâches neuropsycholinguistiques à des sujets présentant un dysfonctionnement cognitif parfois léger tel qu'un « chemobrain ». Afin de s'adapter aux besoins spécifiques de ces patients et de leurs soignants, ce projet implique à la fois des chercheurs en informatique (Julien Pinquier, Jérôme Farinas), en neuro et psycholinguistique (Patrice Péran, Mélanie Jucla), et des cliniciens spécialistes de la prise en charge de ces troubles (Pr Xavier de Boissezon, Lola Danet, Pr Anne Laprie, actuellement responsable scientifique à Toulouse d'un projet d'envergure nationale qui s'intéresse à cette problématique des dysfonctionnements cognitifs chez l'enfant et chez l'adulte à la suite d'un traitement anti-cancer). A terme, le potentiel applicatif de ce projet et son transfert rapide en recherche clinique en oncologie, mais également dans de nombreuses autres disciplines (neurologie, gériatrie, médecine physique et de réadaptation, neuropédiatrie, etc.) contribue ainsi à la stratégie S3 [10] de l'académie de Toulouse. Le projet implique également la société Covirtua [11] pour le développement de la gestion utilisateur et le transfert potentiel vers une solution commerciale.

L'IRIT a en production un serveur [12] pour héberger les travaux de recherche du projet et mettre en place des solutions en reconnaissance automatique de la parole qui puissent alimenter les travaux cliniques et phycho-linguistiques. Jim Petiot (cf. 20) a été recruté pour deux ans en tant qu'ingénieur d'étude pour s'occuper des services et de la mise en place matérielle du serveur. Lila Gravellier (cf. 21) à travers son stage de recherche de master 2 a participé à la mise en place de la solution de transcription.

La solution web est basée sur la solution utilisée par Covirtua pour ses propres services web. Elle permet des passations avec seulement une connexion au serveur EVOLEX. La solution de reconnaissance automatique de la parole utilise maintenant des modèles de neurones profonds et les modèles ont été adaptés aux différentes tâches disponibles (fluence verbale, génération verbale et de dénomination d'images).

5.4.17 Projet PHRIP DAPADAF-E

Le projet PHRIP-19-0004 est un Programme Hospitalier de Recherche Infirmière et Paramédicale, décerné par la Direction générale de l'offre de soins (DGOS) pour un montant de 242 038 €. Il a été attribué à Mathieu Balaguer (cf. §6.2.3) et se déroulera du 1er janvier 2020 au 30 juin 2024. Il s'agit d'un projet visant à tester la « Validité d'une tâche de Décodage Acoustico-Phonétique Automatisée sur les Déficits Anatomo-Fonctionnels dans l'Évaluation paramédicale des troubles de la parole des patients traités pour un cancer de la cavité buccale ou de l'oropharynx » (DAPADAF-E).

L'objectif principal consiste à évaluer la validité d'un score automatique issu d'une tâche de DAP sur les déficits analytiques et dynamiques oro-pharyngés. Cela va se réaliser à travers deux études :

- l'étude de la corrélation entre le score automatique par classe de phonème issu du DAP et le score issu du testing musculaire pour chaque segment anatomique oro-pharyngé.
- Étude de la corrélation entre les deux listes pour l'analyse de la fiabilité.

Les objectifs secondaires concernent :

10. Dans le cadre de la programmation 2014-2020 des fonds européens, l'Union européenne a demandé à toutes les régions d'Europe d'élaborer une « stratégie de spécialisation intelligente » (« smart spécialization strategy ») pour la recherche et l'innovation sur leur territoire.
11. https://www.covirtua.com
12. https://evolex.irit.fr/

— la comparaison des performances d'une évaluation automatique et d'une évaluation perceptive sur les déficits anatomiques et dynamiques oro-pharyngés,
— la production d'un score perceptif humain issu de la transcription des productions par un jury d'orthophonistes,
— la mesure de l'impact du trouble de parole évalué automatiquement par le DAP sur le handicap de parole ressenti par le patient,
— et la production de sous-scores : signes physiques, retentissement fonctionnel, et retentissement psychosocial, jusqu'à la production d'un score de Parole Handicap Index (PHI).

L'étude réalisé rentre dans le cadre du doctorat de Mathieu, et ma participation concerne plus particulièrement l'analyse du handicap de parole ressenti à travers les sous-scores du PHI.

5.4.18 Projet PATY

Le projet plateforme de Parole ATYpique est un des projets sélectionnés en 2020 par l'Institut Carnot Cognition, dans le cadre de son appel à ressourcement auprès des laboratoires de l'Institut. L'objectif général du projet PATY est de mettre en place en 10 mois une plateforme expérimentale, qui permettrait d'améliorer l'accessibilité d'outils automatiques sur de la parole atypique. Cela permettra de faciliter l'accessibilité de résultats de recherche et de développer des collaborations avec des partenaires académiques ou industriels. La plateforme sera développée par un ingénieur d'étude. Un stage au laboratoire LPL permettra de collaborer pour la mise en place du serveur, et réalisera une expérimentation sur celui-ci. Les services web de traitement automatique de la parole atypique développés pourront être diffusé en open source aux laboratoires intéressés du Carnot Cognition. Sur les 20 000 € de financement obtenu, l'IRIT dispose de 16 400 € pour réaliser les 6 mois de financement de Jim Petiot (cf. 20) en tant qu'ingénieur d'étude, et le LPL dispose de 3 600 € pour financer une étude pour tester le dispositif. Le projet a débuté le 15 janvier 2021. L'équipe SAMOVA fournit le serveur web qui hébergera le service [13].

5.4.19 Projet AADI

Le projet AADI « Aphasie, Analyse du Discours en Interactions - constitution de bases de données et nouvelles méthodes d'exploitation » est issu d'un financement de la Région Occitanie, issu du Fond Européen de Développement Régional n° 2019-A03105-52.

Ce projet a collecté des enregistrements issus d'entretiens par internet impliquant de la conversation, du discours spontanée ainsi que des tâches courtes plus formelles comme la dénomination d'images.

Les objectifs du projet AADI sont les suivants :
— mieux comprendre l'aphasie pour aider à la prise en charge
— constituer une base de données informatisée sur le langage et l'aphasie
— concevoir des outils informatiques d'évaluation pour la prise en charge des patients.

Ce projet est ainsi au carrefour d'une recherche pluridisciplinaire en sciences humaines, cliniques et neurosciences. Trois laboratoires de recherche sont impliqués : UMR 5267 PRAXILING Montpellier 3 (porteur du projet), EA4156 Octogone-Lordat de l'Université de Toulouse Jean Jaurès, l'UMR5554 Institut des Sciences de l'Evolution de Montpellier et une entreprise partenaire : Archean Technologies. L'équipe SAMOVA intervient en tant que prestataires afin de fournir des formations sur l'automatisation et l'utilisation d'outils informatiques. Mathieu Balaguer, Julien Pinquier et Jérôme Farinas ont participé à ces formations.

13. https://paty.irit.fr/demo/

5.4.20 Groupement d'Intérêt Scientifique PAROLOTHEQUE

Le Groupement d'Intérêt Scientifique PAROLOTHEQUE, est une structure dont la vocation est de permettre de développer la recherche scientifique sur des données cliniques d'enregistrements vocaux.

Le document constitutif signé le 18 mai 2021 entre les établissements suivants : CNRS, l'Université Toulouse 3, l'Institut National Polytechnique de Toulouse, l'Université Toulouse 1, l'Université Toulouse 2, le Centre Hospitalier Universitaire de Toulouse, l'Institut Claudius Regaud, l'Université d'Avignon et l'Université d'Aix-Marseille. Les laboratoires constitutifs étant : l'IRIT, l'IFERISS, OCTOGONE-LORDAT, le CLLE, le LIA et le LPL.

Par analogie aux Tumorothèques, une PAROLOTHEQUE est une banque d'échantillons de paroles enregistrées, obtenue à partir de bilans de trouble de la parole ou du langage, ou à partir d'entretiens ou d'interviews de personnes concernées par les pathologies tumorales.

Bien que les enregistrements seront orientés par la recherche initiale o u originelle), une automatisation de la transcription « sophistiquée » comprenant si besoin, le repérage des différents locuteurs, la segmentation à partird'une catégorisation sémantique rendra réalisable la réutilisation de la série d'interviews. Ces échantillons seront conservés dans un format numérique sur un serveur dédié au stockage d'enregistrements sonores et recensés dans une Base de données enrichies d'informations permettant leur l'archivage puis leur exploitation sur requête. Ces données feront l'objet des démarches notamment éthiques, requises afin d'en assurer la conformité avec la législation et la réglementation en vigueur. Pour cela un avis d'un comité d'éthique du CHU de Toulouse a été obtenu le 17 mai 2016 et une déclaration CNIL a été faite le 24 juillet 2015 sous la référence 1876994vO.

Ces données s'étendent du code de surface aux aspects contextuels voire conversationnels et peuvent faire l'objet d'une lecture psychologique voire épidémiologique ou sociologique.

La PAROLOTHEQUE sera ainsi alimentée par des recherches en SHS réalisées par les laboratoires et fournira un service facilitant la recherche pour d'autres disciplines. L'objectif principal de ce projet est d'établir un corpus de voix et de paroles de patients atteints de cancer. Ce corpus pourra ensuite être analysé et étudié autant du point de vue de son contenant (parole) que de son contenu (discours). Il permettra de développer des axes de recherches variés et l'élaboration de publications scientifiques dans divers domaines (sciences humaines et sociales : SHS, épidémiologie, linguistique, informatique...).

5.4.21 Projet PHLES-NID

Le projet PhLEs-NID est un projet porté par Virginie Woisard du Centre Hospitalier Universitaire de Toulouse financé par l'Agence Nationale de la Recherche dans le CES Technologie pour la santé, sous le n° ANR-21-CE19-0057 débutant le 1er février 2022, pour une durée de 48 mois et d'un montant de 649252 €.

Les partenaires sont les suivants :

- Centre Hospitalier Universitaire de Toulouse (Coordinateur) Swallis Medical
- Université Toulouse III (Laboratoire IRIT)
- Université d'Aix Marseille (Laboratoire LPL)
- Centre Hospitalier Universitaire de Bordeaux
- Centre Hospitalier Universitaire de Tours
- Centre Hospitalier Universitaire de Rouen
- GIP MiPih
- Oslo University

Le projet PhLEs-NID propose de développer un dispositif non invasif en réponse à la problématique des troubles de la déglutition. Ce dispositif associera plusieurs capteurs capables de détecter différents types de signaux en lien avec le fonctionnement pharyngolaryngé. Il permettra grâce à leur

traitement par de l'intelligence artificielle la prise en charge précoce des troubles de la déglutition. L'interdisciplinarité de l'équipe composée de chercheurs, d'ingénieurs, d'informaticiens, de cliniciens et de méthodologistes travaillant dans des laboratoires, des hôpitaux et des entreprises produira une solution innovante,sensible et spécifique, accessible à la grande majorité de la population.

L'objectif principal de l'étude clinique est d'obtenir, par modélisation, des indicateurs d'efficacité pharyngolaryngée (PhLE). Le critère d'évaluation principal sera l'indicateur d'efficacité du transport pharyngé mesuré par un test de déglutition de référence (fluoroscopie de déglutition ou test de déglutition par nasofibroscopie). Pour les autres indicateurs PhLES, les critères d'évaluation seront mesurés par les résultats des tests de référence. Les objectifs secondaires sont :

- l'identification des troubles de la déglutition et la modélisation de la sévérité du trouble sur la base de critères de travail issus d'un consensus supervisé par un groupe d'experts.
- la modélisation prédictive des complications liées aux troubles de la déglutition et de l'oralité,
- la facilité d'utilisation de l'appareil dans la pratique clinique, en accordant une attention particulière à la tolérance du patient et à la facilité d'utilisation évaluée au moyen de questionnaires et/ou d'entretiens.

Deux équipes de l'IRIT sont concernées par ce projet : l'équipe APO avec Sandrine Mouysset et l'équipe SAMOVA avec Jérôme Farinas. La collaboration permettra de travailler particulièrement sur l'établissement des indicateurs d'efficacité pharyngolaryngée et la modélisation prédictive des complications liées aux troubles de la déglutition et de l'oralité.

5.4.22 Projet ADAPT

Le projet Aide à l'Analyse et au DiAgnostic de la Parole pathologique pour les Thérapeutes (ADAPT) est un des projets portés par Mathieu Bagaguer et sélectionné en 2022 par l'Institut Carnot Cognition, dans le cadre de son appel à ressourcement auprès des laboratoires de l'Institut.

Ce projet propose une collaboration entre l'IRIT avec Mathieu Balagier, Julien Pinquier et Jérôme Farinas et le Laboratoire d'Informatique d'Avignon avec Corinne Fredouille.

L'analyse automatique de la parole pathologique est un axe de recherche en plein développement, en raison de son apport dans la compréhension des processus linguistiques spécifiques dans la parole altérée, et dans le diagnostic des patients. Or, actuellement, peu d'outils simples et fiables sont disponibles pour réaliser ces analyses. ADAPT s'appuie sur la plateforme PATY, développée dans le cadre d'un financement de l'Institut Carnot Cognition. L'objectif est d'enrichir et développer cette plateforme en y ajoutant de nouvelles fonctionnalités comme des nouveaux modèles de reconnaissance adaptés à la parole après cancer ou des systèmes de reconnaissance du locuteur (segmentation et regroupement en locuteurs, suivi de locuteurs dans le continuum de parole spontanée entre locuteur et interlocuteur), répondant aux besoins de transcription automatique de la parole cancérologique par locuteur, rapportés par les chercheurs et cliniciens. Il s'agit également d'adapter l'interface pour permettre à l'utilisateur d'accéder directement aux résultats dans des fichiers personnalisables.

5.5 Campagnes d'évaluation

5.5.1 NIST Language Recognition Evaluation 2003

L'évaluation 2003 sur l'évaluation de la reconnaissance de la langue est relativement proche de la première campagne réalisée en 1996 par l'institut de normalisation américain (NIST) [Martin and Przybocki, 2003]. L'objectif était d'établir une nouvelle base de référence de la capacité de performance en matière de reconnaissance linguistique de la parole téléphonique conversationnelle et de jeter les bases d'autres efforts de recherche dans ce domaine. Les principales données d'évaluation étaient des extraits de conversations en douze langues du corpus CallFriend. Ces segments d'essai avaient une

durée d'environ trois, dix ou trente secondes. Six sites de trois continents ont participé à l'évaluation. Les meilleurs résultats en matière de rendement ont été considérablement améliorés par rapport à ceux de l'évaluation précédente. La proposition de l'IRIT est le fruit d'une collaboration avec le laboratoire Dynamique du Langage de Lyon et a consisté à une modélisation acoustique spécialisé dans chaque langue suivi d'une modélisation de langage en parallèle (PPRLM).

5.5.2 Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques (ESTER) 2005

La première campagne a été organisée dans le cadre du project EVALDA, financé par le Ministère de la Recherche dans le cadre de l'appel à projet Technolangue, sous l'impulsion de l'Association Francophone de la Communication Parlée, du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et de ELDA (Evaluations and Language resources Distribution Agency). Organisée en deux phases, cette première campagne s'est terminée par une campagne de tests menée en janvier 2005.

La campagne d'évaluation ESTER vise à la mesure des performances des systèmes de transcription d'émissions radiophoniques. Les transcriptions sont enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des entités nommées, etc. La transcription enrichie a donc pour but d'obtenir une transcription lisible d'une part et, d'autre part, une représentation structurée du document à des fins d'extraction d'informations. L'évaluation de la qualité des informations annexes en complément de l'évaluation de la transcription orthographique permet d'établir une référence des niveaux de performances de chacune des composantes d'un système d'indexation, tout en donnant une idée des performances du système complet.

L'IRIT a participé aux deux phases de cette première campagne dans toutes les catégories : segmentation des événements sonores (détection parole et musique, réalisé par Julien Pinquier), transcription (j'ai réalisé les modèles acoustiques et l'architecture globale du système, Isabelle Ferrané s'est chargée des modèles de langage et Martine de Calmès des dictionnaires) [de Calmès et al., 2005]. Il s'agit d'une première réalisation qui nous a permis de pouvoir utiliser un système complet de bout en bout.

5.5.3 NIST Language Recognition Evaluation 2007

Depuis 2003 la campagne d'évaluation sur la reconnaissance des langues s'est enrichie au niveau du nombre de langues traitées (20 langues), dispose maintenant d'évaluation de dialectes (anglais, hindi, mandarin, espagnol) dans des conditions fermées ou ouvertes (présence ou non de langue inconnue par le système) [Martin and Garofolo, 2007]. Le système présenté par l'IRIT a été construit sur la plateforme ALIZE/MISTRAL et était constituée d'un modèle universel du monde (UBM) adapté à chaque langue/dialecte [Sanchez-Soto and Farinas, 2007].

5.5.4 Projet Européen QUÆRO 2008-2013

Quæro (« je cherche » en latin) est un programme de recherche et d'innovation lancé le 26 octobre 2004, destiné à développer des « outils intégrés de gestion des contenus multimédias », dont des extensions multimédias pour des moteurs de recherche de nouvelle génération qui permettront de rechercher par le contenu non seulement du texte, mais aussi des images, du son et de la vidéo. Initié par la France et soutenu par Oséo, le programme inclut une participation allemande. Son lancement a bénéficié de la volonté politique de relancer de grands programmes d'innovation industrielle et de réagir face à la montée en puissance des moteurs de recherche américains.

L'IRIT était le partenaire du projet Quæro chargé des évaluations. J'ai été chargé de l'évaluation des tâches 5.3 (identification des langues) et 11.4 (empreintes de morceaux musicaux).

5.5.5 NIST TrecVid 2010

L'objectif principal de TREC Video Retrieval Evaluation (TRECVID) est de promouvoir le progrès dans l'analyse basée sur le contenu et la récupération de la vidéo numérique par le biais d'une évaluation ouverte, basée sur des mesures. TRECVID est une évaluation qui tente de modéliser des situations du monde réel ou des tâches importantes de composants impliquées dans de telles situations.

La participation de l'IRIT est le fruit d'une collaboration entre Hervé Bredin (fusion tardive), Lionel Koenig (paramétrisation audio et modèle de Markov cachés) et moi-même (modélisation acoustique en utilisant ALIZE/MISTRAL) [Bredin et al., 2010].

5.5.6 AIRBUS Challenge ATC 2018

Campagne d'évaluation de la transcription de la parole de contrôle du trafic aérien, organisée par AIRBUS (Estelle DELPECH, François LANCELOT), Safety Data Analysis Services (Céline Raynal), ENAC (Etienne Ceretto) et IRIT (Thomas Pellegrini et Jérôme Farinas). L'évaluation a réuni 22 participants internationaux, et permis d'établir un état de l'art au niveau de la reconnaissance de la parole et de la reconnaissance des identifiants des avions [Farinas and Pellegrini, 2018, Pellegrini et al., 2019].

J'ai participé comme consultant à la définition de la campagne d'évaluation. J'ai ensuite analysé les résultats avec Thomas et nous les avons présentés lors de la journée de restitution. Nous avons ensuite organisé la rencontre à l'IRIT [14].

5.6 Projets de valorisation

5.6.1 Valorisation TTT-Authôt Analyse prosodique et lexicale

Maturation scientifique sous l'égide de Toulouse Tech Transfert à destination de l'entreprise Authôt SAS (projet n°UPS-14-T5-118/TTT-I-14-0946). Deux licences ont été signées le 3 juillet 2015 et le 10 février 2016. Cela portait sur un système de ponctuation automatique, pouvant utiliser des informations acoustiques, prosodiques ou bien lexicales (lié au dépôt logiciel §3 page 76) et à la déclaration d'invention 2 page 75).

5.6.2 Experimentations TTT-AudioToolBox

Contrat de prestation de service avec la Société d'Accélération du Transfert de Technologies Toulouse Tech Transfert [15] n° UPS-2017-662/TTT-P0082 portant sur la réalisation d'une comparaison scientifique des systèmes de reconnaissance automatique de la parole en français, composée des phases suivantes :

- une étude sur le choix et la définition de métriques d'évaluation pertinentes,
- une mise en place d'un corpus varié correspondant à des situations intéressantes scientifiquement et commercialement,
- et une évaluation des principaux systèmes commerciaux et universitaires et production d'un rapport de synthèse.

L'équipe de TTT comprenait : Jerôme Lelasseux (responsable pôle AN), Antonin Crosson (responsable commercial), Sebastien Lebbe (responsable scientifique) et Sylvie Da Pare (responsable technique). L'encadrement à été réalisé à 33%, avec Julien Pinquier (33%) et Thomas Pellegrini (33%) entre décembre 2017 et février 2018. François Xavier Decroix a été recruté à la suite de sa soutenance de thèse, pour réaliser

14. https://www.irit.fr/recherches/SAMOVA/pagechallenge-airbus-atc-workshop.html
15. https://www.toulouse-tech-transfer.com/

l'étude. Les résultats de l'étude ont été valorisés lors d'une communication aux Journées de Phonétique Clinique à Mons en mai 2019 [Farinas et al., 2019].

5.6.3 Prestation WISHINOV

Contrat de prestation de service (n°UPS-2017-653) avec la société Wishinov [16] destiné à développer un système de reconnaissance du locuteur (par technique UBM-GMM) et un système de reconnaissance de mot clef (par programmation dynamique cf. §1.2 page 11). La recherche par mot clé est destinée à être embarquée sur le dispositif Wizzili, un assistant personnel pour gérer les tâches courantes dans la maison [17].

5.6.4 Prestation ARCHEAN

Contrat de prestation de service (n°UPS-2017-103) avec la société Archean Technologies [18] portant sur la réalisation de modèles phonétiques. Les trois livraisons concernent :

— des modèles acoustiques appris sur de la parole radiophonique,
— des modèles adaptés à des données fournies par Archean,
— des modèles spécialisés appris sur des données japonophones.

Ces modèles sont destinés aux applications de mesure de l'intelligibilité, mais également les modèles de reconnaissance de la parole pour les apprenant de langue seconde pour les japonais.

5.6.5 Prestation COVIRTUA

Contrat de prestation de service (n°UPS-0001912) entre la startup COVIRTUA [19] et Jérôme Farinas de l'équipe SAMOVA de l'IRIT.

L'IRIT s'engage à :

1. mettre à disposition de la société Covirtua un système de reconnaissance automatique de la parole grand vocabulaire français
2. rédiger de la documentation sur son utilisation et sur les modifications qui peuvent y être apportées
3. adapter ce système vers une reconnaissance de mots clés spécifique aux besoin logiciel de Covirtua
4. mettre en place ces systèmes à travers un accès à distance sur internet (à travers une interface de communication de type API).

Ce travail est principalement basé sur des modèles acoustiques et linguistiques qui ont été réalisés dans l'équipe SAMOVA de l'IRIT. Les outils et la chaîne de traitement utilisée sont basés sur la plateforme logicielle KALDI, logiciel libre, librement disponible sur internet. Le travail demandé à SAMOVA consiste à respecter le cahier des charges de Covirtua pour adapter ce système afin de réduire les modèles pour les rendre plus performant sur les tâches de reconnaissance de Covirtua. Il s'agit du développement d'une « preuve de concept » à ce stade, et non pas d'une intégration dans un produit fini. Le travail demandé consiste également à apporter à Covirtua l'assistance nécessaire pour l'utilisation de cette transcription automatique.

16. http://www.wishinnov.com
17. https://www.wizzili.com
18. http://www.archean.tech
19. https://www.covirtua.com

5.7 Administration de la recherche

- Élu et membre de la cellule opérationnelle du Conseil Scientifique de l'Université Paul Sabatier (de mars 2000 à mars 2002)
- Nommé personne qualifiée à la commission scientifique du CNESER pour représenter la Confédération des Étudiants Chercheurs en mars 2002
- élu au collège scientifique de la 27e section de l'Université Paul Sabatier (de 2013 à 2015). Constitution de comités de sélection de la section 27 sur l'université Paul Sabatier, définition de la politique des examens des dossiers d'ATER, sélection des dossiers d'ATER.
- Membre du comité scientifique du congrès national bisannuel des Journées d'Etudes sur la Parole (environ 6 revues par édition) depuis 2006.
- relecteur des congrès internationaux INTERSPEECH et ICASSP (environ 6 à 10 revues par an depuis 2010)
- Relecteur du journal Multimedia Tools And Application (MTAP) de Springer depuis 2007 (environ deux revues par an)
- Membre du comité d'organisation du « thème Apprentissage » du LABEX du Centre International de Mathématiques et d'Informatique de Toulouse (CIMI [20]). Participation à la rédaction du projet et participation à la sélection des membres invités.
- Rapporteur dossiers CIFRE (ANRT) depuis 2016
- Rapporteur de projets ANR depuis 2016

6 Activités d'encadrement

Le nombre d'encadrements (en Master, Doctorat et Post-doctorat) est détaillé dans le tableau 5

Table 5 – Nombre d'encadrements

Niveau	Nombre
Master 1	6
Master 2	20
Doctorants (en cours)	5 (+3)
Post-Doctorants (en cours)	7 (+1)

6.1 Encadrements en Master

1. **José Anibal Arias Aguilar (01/09/2003 – 31/08/2004)**
   José Anibal Arias Aguilar a réalisé un stage dans le cadre du DEA Informatique de l'Image et du Langage sur le sujet des « Méthodes à vecteurs de support et de l'indexation sonore » [Arias, 2004]. L'encadrement à été réalisé à 33%, avec Julien Pinquier (33%) et Régine André Obrecht (33%) entre le début septembre 2003 et la fin août 2004. José a ensuite poursuivi sa formation par un doctorat.
2. **Loïc Lefloch (01/09/2004 – 31/08/2005)**
   Loïc Lefloch a réalisé un stage dans le cadre du Master Recherche parcours Image, Information et Hypermédia (2IH) sur une « Étude contrastive de décodeurs acoustico-phonétique » [Lefloch, 2005]. Je l'ai encadré à 100% entre début septembre 2004 et fin août 2008. Loïc a permis de tester un très grand nombre de paramétrisations différentes pour la modélisation acoustique des sons.

20. http://www.cimi.univ-toulouse.fr/

3. **Archange Giscard Destiné (01/09/2010 – 31/08/2011)**
   Archange Giscard Destiné a réalisé un stage dans le cadre du Master 2 Informatique et Télécommunications, parcours AVI, et a été encadré à 50%, avec Isabelle Ferrané (50%) entre début septembre 2010 et fin août 2011. Le stage intitulé « Recherche des expressions clés caractéristiques de l'interaction entre locuteurs dans les documents audiovisuels » [Destiné, 2011] était issu des problématiques gérées par l'équipe SAMOVA dans le cadre du projet EPAC.

4. **Paul Chambon (01/03/2014 – 31/07/2014)**
   Paul Chambon a réalisé son stage dans le cadre du Master 2 parcours Intelligence Artificielle et Reconnaissance des formes, encadré à 50% avec Julie Mauclair (50%) de début mars à fin août 2014. Le stage intitulé « Mesures et paramètres dans le cadre de comparaison de signaux de voix pathologique » a permis de débuter la recherche sur le projet INCA C2SI.

5. **Katia Vidal (01/06/2014 – 31/07/2014)**
   Katia Vidal a réalisé un stage dans le cadre du Master 1 bioinformatique, parcours Biologie des Systèmes intitulé « Comparaison de scores acoustiques d'alignements phonétiques par programmation dynamique en python ». Elle a été encadré à 33% par Thomas Pellegrini, 33% par Maxime Lecoz et 33% par moi-même. Son travail a permis à l'équipe SAMOVA de disposer d'un code fonctionnel pour utiliser l'algorithme de « Goodness of Pronunciation » introduit par Witt [Witt et al., 1999, Witt and Young, 2000] dans un objectif de mesure de la prononciation d'une langue seconde.

6. **Vincent Laborde (01/03/2015 – 31/08/2015)**
   Vincent Laborde a réalisé son stage dans le cadre du Master 2 informatique, parcours Image et Mutimedia, encadré à 50% avec Lionel Fontan de début mars à fin août 2015. Le stage intitulé « Mesures automatiques d'intelligibilité sur des signaux de parole » a permis de travailler sur l'algorithme de « Goodness of Prononciation » a permis de transposer cette approche servant à évaluer le niveau de langue d'apprenant de langue étrangère, à la problématique de la mesure de l'intelligibilité. Ce travail a donné lieu a plusieurs publications [Laborde et al., 2016, Fontan et al., 2015c].

7. **Abdelwahab Heba (29/02/2016 – 26/08/2016)**
   Abdelwahab Heba a réalisé son stage dans le Master 2 informatique, parcours Intelligence Artificielle et Reconnaissance des Formes, dans le cadre d'une coopération entre l'entreprise Intel Toulouse et l'équipe SAMOVA. L'encadrement a été assuré par Benoît Guilhaumon (entreprise Intel, 20%), José Mendes Carvalho (entreprise Intel, 20%), Julien Pinquier (20%), Isabelle Ferrané (20%) et moi-même (20%). Le stage intitulé « Development of a smart algorithm for content analysis of non-verbal information in speech » a permis d'aboutir à la réalisation d'un démonstrateur de détection d'émotions obtenu par une analyse de voix en temps réel. L'objectif était de fournir des extractions d'informations non verbales afin de pouvoir enrichir l'interface de développement proposée par Intel aux constructeurs informatiques.

8. **Sébastien Ferreira (29/02/2016 – 31/08/2016)**
   Sébastien Ferreira a réalisé son stage dans le Master 2 informatique, parcours Intelligence Artificielle et Reconnaissance des Formes, dans le cadre d'une coopération entre la société TELEQUID et l'équipe SAMOVA. Le stage intitulé « Synchronisation de flux TV sur smartphone » a permis de développer une solution inspirée des solutions d'empreintes musicales pour détecter en temps réel le flux de la télévision en cours de diffusion, pour alimenter une application visant à enrichir les contenus diffusés. L'encadrement a été réalisé par Benjamin Ahsan (entreprise TELEQUID, 33%), Julien Pinuqier (33%) et moi-même (33%). Sébastien a ensuite obtenu un financement CIFRE avec l'entreprise Authôt.

9. **Lucien Mahot (29/02/2016 – 31/08/2016)**
   Lucien Mahot a réalisé son stage dans le cadre du Master 2 informatique, parcours Image et Mutimedia et de la coopération entre la société TELEQUID et l'équipe SAMOVA. L'encadrement a été réalisé par Benjamin Ahsan (entreprise TELEQUID, 33%), Julien Pinquier (33%) et moi-même (33%). Le stage est intitulé « Analyse audio par empreinte pour la synchronisation de flux TV » est complémentaire à celui de Sébastien Ferreira, et a permis d'implémenter une méthode de comparaison de morceaux par empreinte audio.

10. **Brendan Gloinec (01/04/2016 – 15/10/2016)**
    Brendan Gloinec a réalisé son stage dans le cadre de la 3ème année du parcours Electronique de l'école d'ingénieur INP-ENSEEIHT. L'encadrement a été réalisé par Julien Mauclair (25%) Etienne Sicard (25%), Oriol Pont (25%) et moi-même (25%). Le stage est intitulé « Développement de Mesures Objectives sur l'Intelligibilité de Voix Pathologiques ». Brendan a essayé une approche systématique en générant de très nombreux paramètres acoustiques des enregistrements des patients du projet INCA C2SI. Il a ensuite effectué une sélection de paramètres afin de mettre en valeur les plus pertinents.

11. **Maëlys Salingre (29/05/2017 – 25/08/2017)**
    Maëlys Salingre a réalisé son stage de son Master 1 à l'Université Paris Diderot (parcours de Linguistique Informatique) dans le cadre d'une collaboration entre la société Authôt et l'équipe SAMOVA. L'encadrement a été réalisé par Stéphane Rabant (entreprise Authôt, 20%) et moi-même (80%). Le stage est intitulé « Étude et développement d'un système d'identification automatique des langues et des dialectes en s'appuyant sur l'analyse des signaux de parole » [Salingre et al., 2017]. L'objectif était de modéliser de manière automatique la prosodie afin de permettre l'identification de parlers régionaux.

12. **Mohamed El Mostadi (05/02/2018 – 04/08/2018)**
    Mohamed El Mostadi a réalisé son stage de Master 2 de l'école Polytech Montpellier (parcours Microélectronique et automatique) dans le cadre d'une collaboration entre la société Renault SW Labs (ex. Intel Toulouse) et l'équipe SAMOVA. L'encadrement a été réalisé par José Mendes-Carvalho (entreprise Renault, 25%), Julien Pinquier (25%), Isabelle Ferrané (25%) et moi-même (25%). Le stage concerne la « Recherche d'information non verbale dans la voix ».

13. **Mathieu Balaguer Navarro (29/01/2018 – 31/08/2018)**
    Mathieu Balaguer Navarro a réalisé son stage de Master 2 en santé Publique, parcours Epidémiologie Clinique. L'encadrement a été assuré par Virginie Woisard (50%) et moi-même (50%). Le stage est intitulé « Construction d'un score Carcinologic Speech Severity Index (C2SI) automatique » et concerne le projet INCA C2SI. Il a permis de proposer un score de mesure de l'intelligibilté et de déterminer les meilleures corrélations avec les analyses automatiques. Cela a donné lieu à plusieurs publications [Balaguer et al., 2019b, Balaguer et al., 2019c, Balaguer et al., 2019a].

14. **Robin Vaysse (16/04/2018 – 31/07/2018)**
    Robin Vaysse a réalisé son stage de Master 1 informatique parcours Statistiques et Informatique Décisionnelle en colaboration entre l'équipe IMT et SAMOVA. L'encadrement a été assuré par Laurent Risser (IMT, 25%), Sébastien Dejean (IMT, 25%), Julie Mauclair (25%) et moi-même (25%). Le stage est intitulé « Analyse statistique de données issues du traitement de la parole pour aider au diagnostic de la maladie de Parkinson et de l'atrophie multisystématisée » et repose sur les données du projet ANR Voice4PD-MSA. Robin a pu essayer différentes paramétrisations pour essayer de discriminer au mieux les pathologies, et il a également testé des idées pour produire des indicateurs sur les tâches de diadococinésie à partir de tempogrammes [Le Coz, 2014].

15. **Yassir Bouiry (01/04/2018 – 31/08/2018)**
Yassir Bouiry a effectué son stage de Master 2 du département informatique de l'INSA de Lyon dans le cadre d'une collaboration entre l'entreprise Wishinnov [21] et SAMOVA. L'encadrement a été assuré par David Mills (entreprise Wishinnov, 20%), Grégoire Tyrou (entreprise Wishinnov, 20%), Julien Pinquier (20%), Isabelle Ferrané (20%) et moi-même (20%). Le stage est intitulé « Enrichissement de la reconnaissance de parole par des marqueurs prosodiques ». Yassir a réalisé un état de l'art sur le domaine et a procédé à l'enregistrement d'un corpus pour évaluer les différentes méthodes qu'il a implémentées.

16. **Laïta Favier (01/06/2018 – 30/09/2018)**
Laïta Favier a effectué son stage dans le cadre de la 2ème année du parcours électronique et traitement du signal de l'école d'ingénieur INP-ENSEEIHT. Le stage a pour sujet la « Visualisation des caractéristiques de la parole dans le cadre de l'étude des interactions vocales avec un assistant virtuel ». L'encadrement a été assuré par Julien Pinquier (33%), Isabelle Ferrané (33%) et moi-même (33%). Laïta a pu approfondir les représentations de la prosodie dans le cadre de la collaboration avec Wishinnov et SAMOVA.

17. **Antoine Viette (18/03/2019 – 20/09/2019)**
Antoine Viette a effectué un stage dans le cadre de la 3ème année du parcours électronique et traitement du signal de l'école d'ingénieur INP-ENSEEIHT. Il s'agit d'une collaboration entre l'ENAC et SAMOVA. L'encadrement a été assuré par François-Régis Colin (ENAC, 25%), Jean Paul-Imbert (ENAC, 25%), Isabelle Ferrané (25%) et moi-même (25%). Le stage est intitulé « Détection des CallSigns et de la phraséologie d'urgence dans des enregistrements audio de la fréquence radio utilisée par les contrôleurs aériens ». Cette problématique est née suite à l'organisation de la campagne d'évaluation des systèmes de la reconnaissance de la parole et de la conférence qui a eu lieu à Toulouse en octobre 2018 [Farinas and Pellegrini, 2018]. L'ENAC a des besoins en traitement automatique de la parole et leur recherche concerne en particulier l'automatisation des traitements des phases critiques dans la communication entre tour de contrôle et pilotes d'avion.

18. **Baptiste Moret (18/03/2019 – 31/08/2019)**
Baptiste Moret a effectué un stage dans le cadre de la 3ème année du parcours électronique diplôme d'ingénieur en électronique et traitement du signal de l'école d'ingénieur INP-ENSEEIHT. Le financement provient de l'Institut des Sciences du Cerveau, de la Cognition et du Comportement de Toulouse (Toulouse Mind and Brain Institute [22]). L'encadrement a été réalisé par Corine Astésano (50%) et moi-même (50%). Le stage est intitulé « État de l'art de l'existant et choix des meilleurs modèles pour l'extraction des paramètres prosodiques caractérisant la dysfluence ». Après un état de l'art des méthodes de modélisation de la prosodie, des outils permettant la visualisation de ces modélisations sur des enregistrements audio dans le but de mettre en évidence des unités linguistiques pertinentes pour l'évaluation des dysfluences dans le cadre du projet ANR RUGBI.

19. **Zakaria Boubkary (27/06/2019 – 15/09/2019)**
Zakaria Boubkary a effectué un stage dans le cadre de la 2ème année du parcours Informatique de l'école d'ingénieur INP-ENSEEIHT et d'une collaboration entre la société Swallis Medical [23] et les équipes APO et SAMOVA. L'encadrement est réalisé par Linda Nicolini (33%), Sandrine Mouysset (33%) et moi-même (33%). Le stage est intitulé « Analyse de données audio de déglutition ». L'objectif est d'étudier les signaux audio et d'accéléromètres posés sur la glotte de personnes

21. http://www.wishinnov.com
22. http://tmbi.fr
23. http://www.startup-semia.com/?startup=swallis-medical

réalisant des déglutitions et de faire un état de l'art des différentes techniques qui permettent de traiter ces signaux.

20. **Jim Petiot (01/01/2019 – 15/06/2021)**
Jim Petiot a été recruté après son Master 2 informatique parcours Intelligence Artificielle et Reconnaissance des Formes à l'Université Paul Sabatier en tant qu'ingénieur d'étude sur le projet EVOLEX (cf. 5.4.16) pour participer à la mise en place d'un serveur destiné à la recherche et le développement de services afférents. Il est encadré à 50% par Julien Pinquier et 50% par moi-même.

    Le contrat de Jim a été étendu pour réaliser la mise en place de la plateforme PATY (cf. 5.4.18), encadré par Julien Pinquier, Jérôme Farinas et Hérvé Bredin.

21. **Lila Gravellier (20/01/2020 – 20/07/2020)**
Lila Gravellier a effectué un stage dans le cadre de la 3ème année du parcours Signal, Image, Communication et Multimedia de l'école d'ingénieur INP PHELMA à Grenoble et de son master 2 "Electronic Systems Design" de Norwegian university of science and technology (NTNU) à Trondheim en Norvège. Son stage porte sur la « Reconnaissance automatique de la parole pour des tests d'évaluation automatisés à destination d'orthophonistes et médecins ». Elle est encadrée par Julien Pinquier (50%) et moi-même (50%). Elle participe à la mise en place d'une solution à l'état de l'art en reconnaissance de la parole (HMM-DNN sous KALDI) et à l'optimisation des performances des différents tests orthophoniques qui sont mis en place dans le cadre du projet EVOLEX (cf. 5.4.16).

22. **Gildas Cherrier (28/06/2021 – 12/09/2021)**
Gildas Cherrier a effectué un stage dans le cade de la 2ème année du parcours Informatique et Télécommunications de l'ENSEEIHT. Son stage est intitulé « Spécialisations de systèmes de reconnaissance de la parole et mise en place sous forme de service web ». Il est encadré à 100% par moi-même. Gildas a produit plusieurs déclinaisons de modèles de reconnaissance automatique de la parole destinés à alimenter le service web pour l'entreprise Covirtua (cf. §5.6.5 page 90).

23. **Jérôme Susgin (01/03/2022 – 28/08/2022)**
Jérôme Susgin a effectué un stage de master 2 dans le cadre de la 3ème année de l'école d'ingénieur de la formation « Systèmes Robotiques et Interactifs » (SRI) de l'UPSSITECH financé par l'entreprise MyFamilyUp. Jérôme Susgin a été encadré à 33% par Véronique Moriceau de l'équipe MELODI de l'IRIT, Jérôme Bertrand de MyFamilyUp et 33% par Jérôme Farinas. My Family Up est une Start Up toulousaine qui développe des bases de données exclusives et un premier niveau d'IA sur la dématérialisation du diagnostic psychologique (psychologie de l'enfant, psychologie des seniors, etc) afin de proposer des services innovants de détection, de prévention et d'accompagnement. Dans ce cadre, elle s'est rapprochée de l'Institut de Recherche en Informatique de Toulouse et notamment de l'équipe de recherche SAMoVA (Structuration, Analyse et Modélisation de documents Vidéo et Audio) et de l'équipe de recherche MELODI (MEthodes et ingénierie des Langues, des Ontologies et du DIscours) pour élaborer un programme de Recherche et Développement sur l'identification des sentiments et états émotionnels dans la communication orale. L'objectif du stage sera de réaliser un l'état de l'art en matière de modélisation de la prosodie et de l'identification des émotions, de l'identification des paramètres de critérisation, de la sélection, de la catégorisation en base de données, de la critérisation numérique d'enregistrements audio de citoyens lambda afin de pouvoir bénéficier d'un corpus exploitable. Le stagiaire pourra également mettre en place une modélisation automatique des sentiments à partir de signaux issus de corpus audio disponibles pour analyser les sentiments. Jérôme Susgin a réalisé l'état de l'art et mis en place un système de reconnaissance d'émotions, sur la base d'un corpus public MSP-PODCAST sur le supercalculateur CNRS Jean Zay.

24. **Ibrahim Abdullah (01/03/2022 – 31/08/2022)**
Ibrahim Abdullah a effectué un stage dans le parcours Signal Imagerie et Applications Audio-Vidéo Médicales et Spatiales dans le master 2 Électronique, Énergie électrique et Automatique. Il s'agit d'un co-encadrement avec Sandrine Mouysset (équipe Algorithmes Parallèles et Optimisation, 40%), Lila Gravellier (en doctorat CIFRE chez Swallis, 33%) et moi-même (33%). Le sujet est basé sur l'analyse des signaux IoT (audio et accéléromètre) issus du collier Swallis Medical en vue de la modélisation de l'efficacité pharyngolaryngée. L'objectif du stage consistait à :
— Caractériser les déglutition mais aussi d'autres activitées pharyngolaryngées (toux, voix, respiration).
— Observer l'évolution du mécanisme de la déglutition au cours du temps.
— Développer un système d'analyse automatique du signal pour y relever tous ces indices.

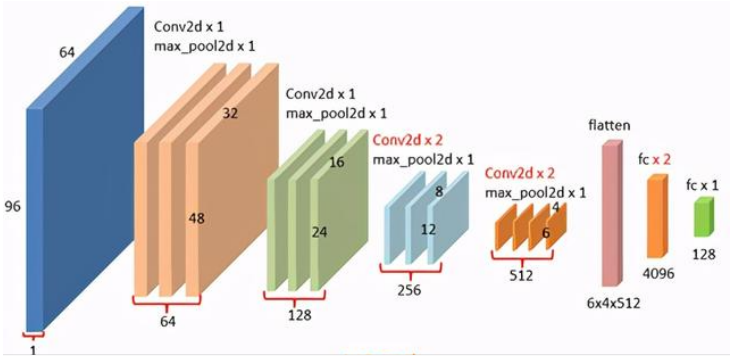


FIGURE 6 – Architecture du système VGGish

Ibrahim, après s'être familiarisé avec les signaux issus du collier, a été amené à produire des annotations pour repérer les événements issus de la déglutition. Il a ensuite mis en œuvre un système de reconnaissance de formes en utilisant une technique de « transfert learning » à partir d'un modèle VGG [Simonyan and Zisserman, 2015, Hershey et al., 2017] à partir du tutoriel « Audio Transfer Learning with Scikit-learn and Tensorflow » [24] [Pons et al., 2018]. Les résultats obtenus sont prometteurs : sur une tâche binaire (détection de 249 déglutition / 304 autres événements), il a été obtenu 91% de bonne détection (F1 score). L'apprentissage avait été réalisé sur peu de données (3h de corpus), ce qui valide la possibilité d'utiliser un modèle appris de des tâches d'événements sonores (audioset) est applicable sur des enregistrements audio de signaux de déglutition, mais également sur les signaux des accéléromètres.

25. **Jean Baqué (14/11/2022 – 02/07/2023)**
Jean Baqué est interne en médecine au service ORL dans le service de chirurgie cervico-faciale du CHU de Tours et effectue un stage « Validation d'un dispositif médical non invasif évaluant l'efficacité pharyngo-laryngée » dans le cadre du Master 2 Biologie santé parcours Sciences chirurgicales de l'Université Paris-Est Créteil (UPEC). Ce stage est effectué à Tours avec le professeur Sylvain Morinière (PU-PH au CHU de Tours), et à l'IRIT avec Jérôme Farinas. La dysphagie oropharyngée est un problème de santé publique, de part sa fréquence et sa détection reposant sur des explorations cliniques et une imagerie peu sensibles et accessibles. Son étude se limite à l'évaluation de l'efficacité des mécanismes de transport, alors que d'autres paramètres physiologiques pourraient permettre de mieux la comprendre. L'objectif du stage est de tester et valider un dispositif médical non invasif évaluant l'efficacité pharyngo-laryngée, à l'aide de capteurs

24. https://github.com/jordipons/sklearn-audio-transfer-learning

enregistrant les signaux physiologiques du fonctionnement du pharyngo-larynx. Ce projet sera réalisé dans le cadre de l'étude multicentrique du projet ANR PHLES-nid (cf. section 5.4.21). Les mesures seront faites dans le service d'ORL du CHU de Tours, sur 20 sujets sains, et 20 à 30 patients atteints de cancers de la tête et du cou traités par radiothérapie. Le dispositif médical enregistrera les signaux physiologiques du fonctionnement du pharyngo-larynx lors de diverses situations avec des capteurs : sonomètre, accéléromètre, électromyogramme de surface, canule nasale et oxymétrie de pouls. Nous aurons ainsi des données sur les quatre fonctions pharyngo-laryngées : la phonation, la déglutition, la respiration, la protection des voies aériennes. Ces signaux seront analysés dans un laboratoire de l'Institut de Recherche en Informatique de Toulouse (UMR 5505), pour repérer les différents évènements de la déglutition, et établir une méthode les analysant automatiquement et modélisant l'efficacité pharyngo-laryngée. Nous nous attendons à montrer que ce dispositif médical apporte des informations analysables, reproductibles et modélisables pour évaluer l'efficacité pharyngo-laryngée, chez les sujets sains et les patients. Ce dispositif médical, de part son analyse automatique et non invasive de l'efficacité pharyngo-laryngée, permettrait de dépister les troubles de la déglutition à grande échelle.

26. **Philippe Allet (27/03/2023 – 25/08/2023)**
Philippe Allet a effectué un stage en Master 2 informatique parcours Intelligence artificielle : fondements et applications (IAFA) de l'Université Toulouse 3, du 27 mars au 25 août 2023. Il a été financé dans le cadre du Domaine d'Application Stratégiques Santé, Autonomie et Bien-Être de l'IRIT. L'encadrement a été assuré par une collaboration entre Sandrine Mouysset (équipe IRIT/APO, 50%) et Jérôme Farinas (équipe IRIT/SAMOVA, 50%). L'intitulé du stage était « Analyse de la déglutition à partir de capteurs non invasifs ».

    L'objectif de son stage était de se familiariser avec les signaux issus du collier SWALLIS DSA® et proposer des modélisations par réseaux de neurones profonds. Les enregistrements du projet PHLES-nid (cf. §5.4.21 page 86) n'étant pas disponibles au début du stage, Philippe a utilisé les données récoltées dans le cadre du doctorat de Lila Gravellier (cf. §6.2.6 page 102).

    Philippe a réussi à mettre en place un système de reconnaissance basé sur YAMNET [Plakal and Ellis, 2020]. Il s'agit d'un extracteur de paramètres pré-appris sur Audioset [Gemmeke et al., 2017, Google, 2017] en utilisant une architecture de convolution séparable en profondeur Mobilenet_v1 [Howard et al., 2017] un corpus d'événements sonores provenant de Youtube®. L'idée était d'évaluer la faisabilité de l'utilisation de modèles appris sur des signaux audio différents afin de pouvoir réaliser des modélisations sur les quatre signaux du capteur. Le modèle avait pour but de différencier les différentes matières ingérées lors des phases de déglutition sur les enregistrements sur des personnes saines. Ceci afin de voir si les matières ingérées se différenciaient à travers une modélisation par réseau de neurones profond.

6.2 Encadrements en Doctorat

6.2.1 José Anibal Arias Aguilar (01/09/2004 – 29/09/2008)

José Anibal Arias Aguilar a réalisé un doctorat sur les « Méthodes spectrales pour le traitement automatique de documents audio ». L'encadrement a été réalisé par Régine André-Obrecht (HDR, 33%), Julien Pinquier (33%) et moi-même (33%). Il a bénéficié d'un financement SFERE-CONACyT [25] du « Mexican National Council of Science and Technology ».

Les méthodes spectrales sont des outils puissants pour la réduction non-linéaire de la dimensionnalité et pour l'apprentissage de variétés. Leur fonctionnement dépend de la diagonalisation de matrices

25. https://www.conacyt.gob.mx

de « similarité » spécialement conçues, très souvent définies à partir des graphes pondérés dont les sommets représentent les données à analyser et dont les arêtes indiquent des relations de voisinage. Ces matrices sont des matrices « noyaux » construites à partir des données d'entrée. Une analyse de différentes techniques a été réalisé : algorithmes PCA, MDS, regroupement spectral, Kernel PCA, Isomap, LLE, GPLVM et Laplacian eigenmaps. Ces méthodes possèdent des propriétés très intéressantes qui ont permis pour projetter les documents acoustiques dans des espaces de faible dimensionnalité [Arias et al., 2008a, Arias et al., 2008b]. Cela a permis de visualiser la structure phonémique de la parole et d'en extraire des algorithmes capables de l'identifier automatiquement [Arias et al., 2006]. À l'aide de cette étude, il est apparu que la parole et la musique partagent le même espace acoustique, mais chaque signal possède un sous-espace spécifique de variabilité.

Une étude sur la dimensionnalité intrinsèque des séquences de parole a été réalisée. Elle est utile pour énoncer certaines observations sur la dimensionnalité de la paramétrisation cepstrale appliquée à la parole et à la musique. Par exemple, il a été mis en évidence que des espaces MFCC de dimension réduite sont envisageables pour caractériser effectivement l'information acoustique. L'analyse de la dimensionnalité intrinsèque indique également que la parole en condition de stress réduit sa variabilité par rapport à la parole spontanée.

L'examen des unités phonétiques qui composent les séquences de parole a été réalisé. Un algorithme non supervisé de segmentation et d'identification automatique de ces unités a été développé [Arias et al., 2008c]. Nous avons obtenu une précision acceptable au niveau de l'identification de ces unités par rapport à un expert humain, surtout pour la classification SCV (grandes classes phonétiques : silence, consonne et voyelle). La précision de l'identification des consonnes voisées et non voisées est moins bonne, mais elle est peut être utile pour certaines applications. L'étiquetage SCV est aussi utilisé pour aligner les projections de faible dimensionnalité des séquences de parole. Dans une approche de « fouille de données », nous sommes arrivés à visualiser les densités de probabilité qui représentent des suites de parole et de musique. Pour la tâche d'identification de locuteurs, une analyse non supervisée de ces projections est proposée pour découvrir le nombre de groupes stables dans l'ensemble, et une analyse supervisée pour étudier le degré de séparabilité de ces groupes et la complexité de leurs frontières.

La soutenance a eu lieu le 29 septembre 2008 à l'Université Paul Sabatier [Arias, 2008]. José est actuellement professeur à l'université technologique de Mixteca au Mexique. Il continue sa recherche sur les méthodes d'apprentissage automatique appliqué aux media audio-vidéo.

6.2.2 Sébastien Ferreira (01/11/2016 – 21/05/2021)

Sébastien Ferreira a obtenu un financement CIFRE avec la société Authôt et l'IRIT. L'encadrement est réalisé de manière conjointe entre Stéphane Robant (directeur technique de l'entreprise, 33%), Julien Pinquier (HDR directeur de thèse, 33%) et moi-même (33%). Authôt est une entreprise spécialisée dans les services de transcription de la parole. Elle propose des services de transcription ou de sous-titrage automatique et d'amélioration par correction manuelle de ces transcriptions. Le client téléverse un document audio ou audio-vidéo et la société met à disposition la transcription.

Le doctorat de Sébastien se place dans ce cadre applicatif. Le sujet est la « Prédiction de la qualité des systèmes de reconnaissance automatique de la parole et stratégies d'adaptation ». L'objectif principal est d'arriver à prédire les performances de transcription de la parole sur des fichiers spécifiques. En effet cette prédiction permettra d'informer l'utilisateur de ces services de la qualité attendue du service automatique et permettra également de pouvoir aiguiller vers le meilleur système de transcription à utiliser pour traiter le fichier. Il ne s'agit pas de faire de la prédiction *a posteriori* mais *a priori*. La production de cette prévision demande une analyse particulière du signal et une identification des conditions de

perturbation de l'enregistrement. Parmi les nombreuses sources de perturbations de l'enregistrement, on retrouve les conditions bruitées et les altérations de type réverbération.

Sébastien a commencé par faire l'inventaire de tout ce qui peut perturber un système de reconnaissance de la parole. Puis il s'est concentré sur la problématique particulière du bruit. Une définition du bruit a été nécessaire du fait sa grande variabilité. Des paramètres pertinents pour caractériser le bruit ont été développés et sélectionnés ce qui a permis de mettre en place une régression pour prédire le taux d'erreur mot de systèmes de reconnaissance automatique [Ferreira et al., 2018]. Un second travail a été mené sur l'impact de la réverbération sur les systèmes de reconnaissance automatique de la parole. Cette perturbation du signal acoustique dégrade fortement les systèmes et présente un réel défi pour la prédiction. La recherche de paramètres appropriés pour la caractériser a été entrepris et une prédiction a été mise en place [Ferreira et al., 2019b].

La dernière partie du travail de recherche concerne l'étude de l'impact de la parole superposée sur la dégradation des résultats des systèmes de reconnaissance automatique de la parole.

Sébastien a soutenu son doctorat « Prédiction a priori de la qualité de la transcription automatique de la parole par l'analyse de l'environnement sonore » le 21 mai 2021 [Ferreira, 2021]. Le jury était composé de François Portet, rapporteur, Jean-François Bonastre, rapporteur et Martine Adda-Decker, examinatrice.

6.2.3 Mathieu Balaguer Navarro (01/09/2018 – 12/10/2021)

Mathieu Balaguer Navarro (ORCID : 0000-0003-1311-4501) a commencé son doctorat en formation continue au CHU de Toulouse, puis s'est placé en détachement en septembre 2019 pour poursuivre sa formation, financé par le projet RUGBI (cf. §5.4.15 page 83). L'encadrement est réalisé par Julien Pinquier (HDR, 33%), Virginie Woisard (HDR, 33%) et moi-même (33%).

Le sujet du doctorat porte sur l'étude de « Impact fonctionnel des troubles de la parole évalués par une mesure automatique sur les actes de communication quotidiens chez les patients traités pour un cancer de la cavité buccale ou de l'oropharynx ». Ce sujet fait suite à celui de son master (cf. §13 page 93).

Les troubles de parole sont une problématique fréquemment rencontrée après traitement d'un cancer de la cavité buccale ou de l'oropharynx. Mais peu d'études s'intéressent à l'heure actuelle aux conséquences de ce trouble sur les capacités de communication des patients ou leur qualité de vie. Or, en contexte clinique, l'optimisation des capacités de communication est un objectif thérapeutique majeur dans le suivi de ces patients. En pratique courante, l'évaluation des troubles de parole donne des scores prédisant mal l'impact de ces troubles sur la communication. L'analyse automatique de la parole, moins variable que l'évaluation perceptive habituellement utilisée, est un axe en plein développement. Dans ce doctorat, Mathieu a cherché à mesurer l'altération de la communication au moyen d'analyses automatiques de la parole spontanée. Trois aspects ont été étudiés : la mesure de l'altération de la communication, l'analyse automatique de la parole spontanée, et la prédiction de l'altération de la communication par les paramètres automatiques. Pour ce faire, un nouveau corpus de parole a été constitué auprès de 25 sujets traités pour un cancer de la cavité buccale ou de l'oropharynx. Il comprend une tâche de parole spontanée enregistrée au cours d'un entretien semi-dirigé, mais aussi des autoquestionnaires autorisant la mesure de la communication et des facteurs associés à la parole et à la communication. Concernant le premier aspect, un score de référence mesurant de façon holistique la communication a été construit. Il permet de combler le manque d'outils disponibles en cancérologie ORL pour cette mesure. Le deuxième aspect concerne l'analyse automatique de la parole. Une revue systématique de littérature a conduit à s'intéresser aux outils applicables à l'analyse de la parole spontanée, qui est le contexte de production le plus proche de la communication quotidienne. Cent quarante-neuf paramètres automatiques issus

des différents niveaux du modèle psycholinguistique de communication de Caron ont été extraits. Puis, un processus de sélection a abouti à retenir 75 paramètres pertinents et non redondants. Enfin, pour le troisième aspect, une modélisation prédictive de l'altération de la communication a été réalisé, au moyen des paramètres automatiques retenus (corrélation de 0,83 entre score prédit et score réel). La corrélation atteint même 0,89 en incluant à la modélisation des facteurs associés (constitution des cercles sociaux, état anxio-dépressif, déficits associés, auto-perception du handicap lié au trouble de parole). L'utilisation de l'analyse automatique de la parole permet donc une prédiction fiable de l'altération de communication ressentie par les patients. Cette étude ouvre de nouvelles perspectives quant à l'utilisation et l'optimisation des systèmes de reconnaissance automatique de parole dans l'évaluation clinique d'une part, et la prise en compte des besoins fonctionnels et psychosociaux exprimés par les patients d'autre part.

Mathieu a soutenu son doctorat « Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé » le 12 octobre 2021 [Balaguer, 2021]. Le jury était composé de : Nathalie Henrich Bernardoni (Directrice de Recherche, Laboratoire GIPSA Lab UMR CNRS, Grenoble-INP et Université de Grenoble-Alpes, rapportrice) ; Emmanuel Babin (Professeur des Universités - Praticien Hospitalier, CHU Caen, Unité INSERM U1086 ANTICIPE « Cancers et prévention », rapporteur) et Rudolph Sock (Professeur des universités, LiLPa Université de Strasbourg, examinateur et président du jury).

6.2.4 Vincent Roger (01/12/2018 – 28/02/2022)

Vincent Roger a débuté son doctorat en décembre 2018 sur un financement de l'Université Fédérale de Toulouse et Région Occitanie pour une collaboration entre l'unversité Jean Jaures (laboratoire Octogone-Lordat représenté par Virginie Woisard) et l'université Paul Sabatier (laboratoire IRIT, équipe SAMOVA). L'encadrement est réalisé par Julien Pinquier (HDR, 45%), Virginie Woisard (HDR, 10%) et moi-même (45%). Son sujet de recherche porte sur la « modélisation de l'intelligibilité de la parole pathologique ».

Vincent a commencé par prendre en main la base de donnée C2SI et la problématique [Woisard et al., 2021]. Il a réalisé des expériences visant à apprendre le concept d'intelligibilité avec des modèles de réseaux profonds classiques qui ont mis en évidence l'inadéquation des approches classiques sur un volume de donnée d'environ 1h. Suite à ce constat, il a réalisé un état de l'art des systèmes de reconnaissance automatique de la parole pour appréhender la contrainte de faible volume d'apprentissage.

L'expérience suivante a été d'utiliser la technique de transfert de connaissances pour représenter les données et classifier l'intelligibilité à partir de ces représentations. Les expérimentations se sont trouvées infructueuses dû au fait que la tâche cible (mesure d'intelligibilité) se trouve être trop éloignée de la tâche initiale du modèle utilisé pour le transfert (transcription de la parole).

Vincent a réalisé un état de l'art sur les techniques dites de « few-shot » (adaptées aux bases de données avec peu d'exemples par classes). Il se trouve que ces techniques sont majoritairement utilisées en image et qu'il faut les adapter pour l'audio. Vincent a publié un article sur l'application de ces techniques pour les signaux audio [Roger et al., 2022a].

Vincent a ensuite proposé une nouvelle méthode pour mesurer l'intelligibilité et la sévérité de la parole en utilisant une mesure entropique. Celle-ci est fondée sur des représentations de la parole auto apprise sur le corpus Librispeech : le modèle PASE+ [Ravanelli et al., 2020], qui est inspiré de l'Inception Score (généralement utilisé en image pour évaluer la qualité des images générées par les modèles). Notre méthode nous permet de produire un score semblable à l'indice de sévérité avec une corrélation de Spearman de 0,87 sur la tâche de lecture du corpus cancer [Roger et al., 2022b]. L'avantage de cette approche est qu'elle ne nécessite pas des données du corpus C2SI-RUGBI pour l'apprentissage.

Ainsi, nous pouvons utiliser l'entièreté du corpus pour l'évaluation de notre système. La qualité de nos résultats nous a permis d'envisager une utilisation en milieu clinique à travers une application sur tablette : des tests sont d'ailleurs en cours à l'hôpital Larrey de Toulouse.

Vincent Roger a soutenu sa thèse « Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond : d'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique » le 29 sptembre 2022 [Roger, 2022]. Le jury était composé de Hervé Glotin (Professeur à l'Université de Toulon, examinateur et président), Cécile Fougeron (Directrice de Recherche CNRS au LPP à Paris, rapportrice) et Jean-François Bonastre (Professeur au LIA à l'université d'Avignon, rapporteur).

6.2.5 Robin Vaysse (01/10/2019 – 31/12/2022)

Robin Vaysse débute sa thèse le 1er octobre 2019 sur un financement fléché par le Bonus Qualité Recherche de l'université Paul Sabatier. La direction est assurée par Corine Astésano (HDR, 50%), et moi-même (50%) (suite à une autorisation à diriger les recherches à titre individuel délivrée par l'Université Paul Sabatier).

Le sujet porte sur la « Mesure de l'intelligibilité de la parole par une modélisation automatique de la prosodie ». Situé dans le cadre du projet ANR RUGBI (cf. §5.4.15 page 83) et de la collaboration entre l'université Jean Jaures (laboratoire Octogone-Lordat) et l'université Paul Sabatier (équipe SAMOVA laboratoire IRIT). J'ai obtenu une habilitation à diriger les recherches à titre individuel afin de pouvoir diriger ce doctorat. La thèse a été soutenue le 21 mars 2023 [Vaysse, 2023].

Dans ce doctorat, nous nous sommes intéressés à l'impact que certaines pathologies peuvent avoir sur la production du rythme de la parole. Plus particulièrement, nous avons étudié deux types de pathologies : la maladie de Parkinson, ainsi que les patients atteints d'un cancer de la cavité buccale ou de l'oropharynx ayant subi un traitement médical. Notre objectif principal a été de proposer une modélisation automatique du rythme de la parole pathologique. Grâce à cette modélisation, nous avons voulu mettre en évidence les régularités rythmiques à différents niveaux prosodiques, dans le but de pouvoir caractériser les stratégies de production de parole mises en jeu chez des personnes atteintes de ces deux pathologies.

Robin a tout d'abord produit une étude visant à sélectionner les meilleurs algorithmes d'extraction de la fréquence fondamentale pour pouvoir traiter des enregistrements audio [Vaysse et al., 2021a, Vaysse et al., 2022a]. Il a ensuite pu réaliser une étude [Vaysse et al., 2022b] sur le corpus de la Maladie de Parkinson dans le cadre du projet RUGBI.

Suite à la constitution d'un état de l'art sur les techniques de modélisation du rythme, nous avons sélectionné celles dont l'implémentation se rapproche au mieux de nos présupposés théoriques. Nous avons alors testé ces méthodes sur un corpus de Slam dans le but de sélectionner les méthodologies qui modélisent au mieux la hiérarchie rythmique de la parole. La modélisation que nous avons retenue se base sur l'analyse des modulations lentes (inférieures à 10 Hz) de l'amplitude du signal de parole. Cette méthode appelée le spectre de modulation d'enveloppe (EMS) permet de caractériser la stratégie de segmentation de la parole des locuteurs. Ainsi, nous avons pu observer dans notre corpus de parole pathologique que les personnes présentant de forts troubles de l'articulation des syllabes ont tendance à favoriser une structuration prosodique très régulière. Au contraire, une personne sans troubles apparents de l'articulation présente une structuration prosodique moins régulière. Nous supposons donc que les patients dont l'intelligibilité est faible à cause de troubles articulatoires se focalisent davantage sur une structuration très régulière de leur parole avec des durées de groupes de mots de longueurs équivalentes. Nous avons par la suite modélisé l'intelligibilité des patients en nous focalisant uniquement sur des indices purement rythmiques issus de l'EMS [Vaysse et al., 2021b]. Cependant, après analyse des

résultats, les indices rythmiques les plus corrélés au score d'intelligibilité de référence estimés par des médecins ORL étaient en réalité fortement dépendants du débit de parole. Nous avons donc proposé de nouvelles caractéristiques du rythme indépendantes du débit de parole. À l'aide de ces nouveaux paramètres, nous avons pu proposer une représentation en deux dimensions de notre corpus de parole pathologique. Cette représentation basée sur les niveaux principaux de régularités de l'EMS nous a permis de caractériser et de regrouper les personnes avec des stratégies de segmentation de la parole particulières.

L'EMS est donc une modélisation pertinente du rythme de la parole qui permet de caractériser efficacement le rythme de la parole au travers d'une représentation de la régularité des niveaux prosodiques à différents niveaux de hiérarchie [Vaysse et al., 2023, Farinas et al., 2023].

Robin Vaysse a soutenu son doctorat « Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson » le 21 mars 2023 [Vaysse, 2023]. Le jury était composé de : Cécile Fougeron (Directrice de Recherche CNRS au LPP à Paris, examinatrice et présidente du jury), François Pellegrino (Directeur de Recherches CNRS au Laboratoire DDL à Lyon, en tant que rapporteur), Elisabeth Delais-Roussarie (Directrice de Recherche CNRS au Laboratoire de Linguistique de Nantes, rapportrice) et Virginie Woisard-Bassols (PU-PH au CHU de Toulouse, examinatrice).

6.2.6 Lila Gravellier (15/07/2021 – 25/04/2024)

Lila Gravellier a obtenu un financement CIFRE avec la société Swallis Medical et l'IRIT pour le doctorat sur les « Analyses multimodales pour une aide au diagnostic non intrusif de la déglutition ». L'encadrement est réalisé de manière conjointe entre Maxime Le Coz (docteur, scientifique des données, 33%), Julien Pinquier (HDR directeur de thèse, 33%) et moi-même (33%).

Ce sujet de thèse s'inscrit dans le cadre du développement de l'outil de diagnostic de la déglutition de la société Swallis Medical. Cet outil contient un collier de mesures enregistrant différents signaux sonores et vibratoires au niveau du cou du patient afin de les visualiser et de les annoter par des spécialistes en vue de l'estimation du bon fonctionnement de la fonction de déglutition. Grace à ce dispositif et à l'analyse vibro-acoustique qui en découle, les utilisateurs peuvent détecter et objectiver des troubles de la déglutition, autrement effectué par l'écoute via un stéthoscope qui demeure une modalité d'écoute subjective. Les dysfonctionnements de la déglutition peuvent avoir de graves conséquences pour un patient (étouffement, infection pulmonaire, malnutrition...) et une détection fiable de ces problèmes permettrait une amélioration de la santé et du confort des patients, notamment par un processus de réhabilitation et une adaptation des textures de l'alimentation.

L'objectif du doctorat est de réaliser un bilan de la déglutition pour un patient en utilisant le dispositif Swallis. Le dispositif utilise des capteurs non invasifs qui diffèrent des méthodes d'acquisition d'informations classiquement utilisées dans les bilans. Afin de pouvoir exploiter ce type de mesures, il est nécessaire de mener une étude complète afin d'appréhender le comportement des capteurs en fonction des différents évènements qui peuvent être amenés à être enregistrés. Ces différents événements sont principalement : la respiration, la déglutition, la parole, et les modes d'expectoration mécanique (toux et raclements de gorge). Dans ce doctorat, nous proposons d'analyser ces situations, en procédant progressivement : en partant de l'analyse de signaux issus d'une seule personne ne présentant pas de pathologies particulières, puis en agissant sur le nombre de personnes étudiées et en variant les conditions d'enregistrement des situations étudiées. La variabilité en fonction des personnes pourra aussi être prise en compte en mettant en place des protocoles d'enregistrements dans les hôpitaux, par exemple lors d'un bilan classique ORL.

Lila a réalisée l'acquisition d'un corpus de plus de 49 personnes (19 hommes et 23 femmes, de 22 à 57 ans, âge moyen de 34 ans). Chaque enregistrement a été réalisé avec la version finale du dispositif Swallis DSA®, et consistait en différentes phases comprenant des déglutitions de différents volumes et textures ainsi que d'autres activités pharyngo-laryngées. Les signaux ont été enregistrés comprennent des annotations de 880 déglutitions contrôlées et 824 déglutitions spontanées.

Lila a également mis en place des modèles permettant la segmentation automatique des signaux et la détection automatique des événements de déglutition [Gravellier et al., 2023b,Gravellier et al., 2023a].

6.2.7 Adrien Lafore (01/10/2023 – 30/09/2026)

Adrien Lafore a obtenu un financement CIFRE avec la société Swallis Medical et l'IRIT pour le doctorat sur les « Détection automatique des états émotionnels dans la parole pour le soutien psychologique ». L'encadrement est réalisé de manière conjointe entre Véronique Moriceau (HDR directrice de thèse, équipe IRIT/MELODI 33%) et moi-même (co-encadrant, équipe IRIT/SAMOVA, 33%) et Marie Françoise Bertrand (encadrante industrielle, présidente de MyFamilyUp, 33%). My Family Up (MFU) est une Start Up toulousaine, labellisée JEI et son programme de recherche METAPSY avec l'IRIT est validé par le Pôle de compétitivité Eurobiomed et le Gérontopôle du CHU de Toulouse. MFU développe des bases de données exclusives et un premier niveau d'Intelligence Artificielle (IA) sur la dématérialisation du diagnostic psychologique (psychologie de l'enfant, psychologie des seniors, etc) afin de proposer des services innovants de détection, de prévention et d'accompagnement des aidants familiaux.

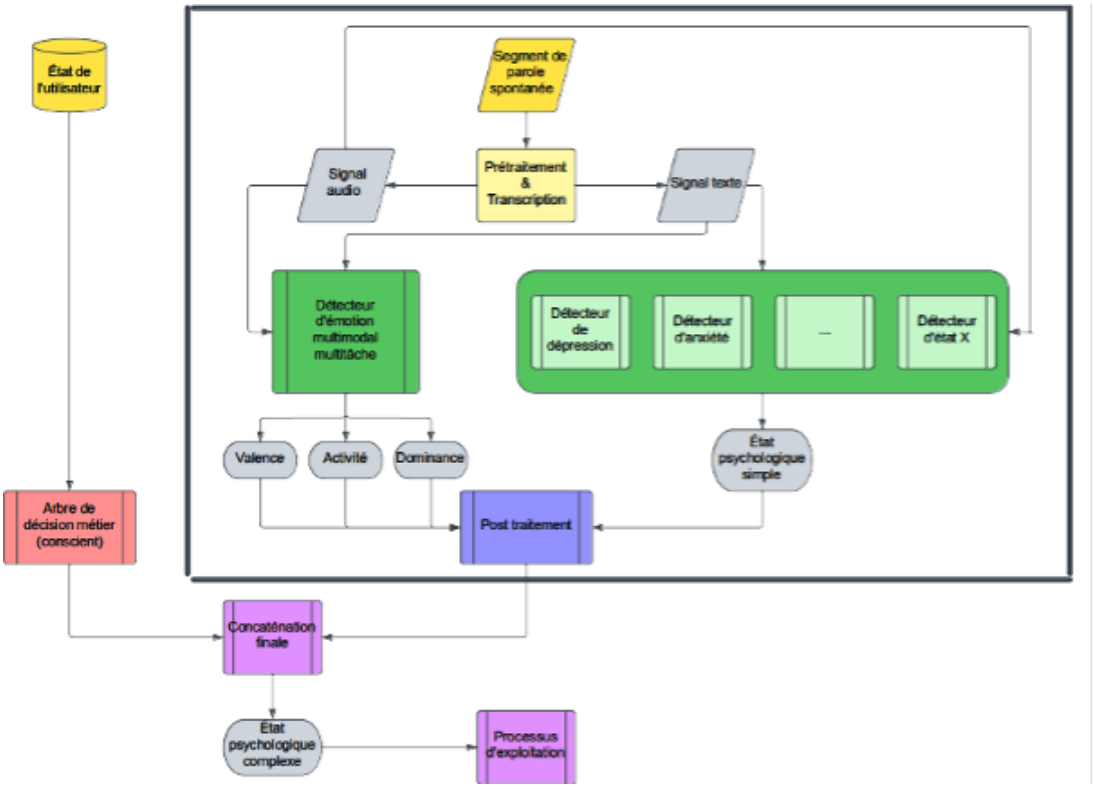


FIGURE 7 – Schéma fonctionnel des systèmes automatiques et du positionnement avec MFU

Un aidant familial ou proche aidant est une « personne qui vient en aide, de manière régulière et fréquente, à titre non professionnel, pour accomplir tout ou partie des actes ou des activités de la vie

quotidienne d'une personne en perte d'autonomie, du fait de l'âge, de la maladie ou d'un handicap » (Loi n° 2015-1776 du 28 décembre 2015, article 51). En 2021, on estime à 11 millions le nombre d'aidants en France, soit un français sur six. D'après un rapport du ministère de l'économie, les principales observations sont :

- Leur âge moyen est de 49 ans et 37% des aidants sont âgés de 50 à 64 ans ; 60% des aidants sont des femmes ;
- 69% des aidants constatent un impact réel sur leur état moral,
- 53% des aidants subissent des effets sur leur propre santé ;
- 50% se sentent parfois seuls, non soutenus moralement ;
- 62% se sont déjà retrouvés dans un état d'épuisement intense.

D'après l'OMS, 70% des jeunes aidants présentent des troubles anxio-dépressifs et 15 à 30% des personnes âgées souffrent de dépression. En 2021, 75% des français estiment qu'il faut du « courage » pour aller voir un psychologue (étude Yougov pour Psychologies) et des études de la CAF et de la DREES ont mis en évidence la demande des citoyens français pour des solutions de soutien psychologique personnalisées, professionnelles et accessibles par internet. La thérapie en ligne est souvent le seul soin possible, mais la majorité des applications de thérapie ne sont pas fiables dans une logique thérapeutique.

Afin de disposer d'un outil d'aide au diagnostic à destination des psychologues de MFU, les modèles développés au cours de la thèse permettront de fournir un ensemble d'informations sur chaque locuteur. La figure 7 détaille une architecture possible. La partie encadrée situe les travaux de thèse par rapport à l'architecture déjà existante chez MFU (en dehors du cadre). Les informations audio permettront de produire des projections sur les axes classiquement utilisés en traitement automatique : valence, activité et dominance. L'analyse textuelle permettra d'identifier des indicateurs émotionnels. La fusion de ces 2 modélisations permettra de classer chaque locuteur selon une ou plusieurs catégories : anxiété, détresse émotionnelle, dépression (classes non mutuellement exclusives). D'autres catégories pourront éventuellement être ajoutées au cours de la thèse en fonction des demandes des psychologues de MFU.

6.2.8 Philippe Allet (01/10/2023 – 30/09/2026)

Philippe Allet va débuter un doctorat le 1er octobre 2023 intitulé « Modélisation de l'efficacité pharyngo-laryngée chez les patients pathologiques avec un dispositif non-invasif » sous la supervision de Jérôme Farinas (direction 40%), Sandrine Mouysset (40%), Julien Pinquier (10%) et Sylvain Morinière (10%), avec le financement du projet PHLES-nid (cf. §5.4.21 page 86).

La dysphagie oropharyngée est un problème majeur de santé publique. Ce trouble de la déglutition touche les personnes âgées (50%) et les patients atteints de troubles neurologiques et de cancers de la tête et du cou (80%), les principales complications étant les infections respiratoires et la dénutrition. Ces complications entraînent de longues hospitalisations, l'utilisation d'antibiotiques et d'aliments industriels modifiés, et une augmentation du taux de mortalité dans ces populations déjà fragiles. Le contrôle de ces complications et la réduction de leurs conséquences reposent sur la détection et le traitement précoces de la dysphagie. Or, à l'heure actuelle, les explorations cliniques de faible sensibilité, l'imagerie diagnostique (vidéofluoroscopie (VFS) et nasofibroscopie de déglutition (FEES)) jugée invasive, chronophage, peu accessible, et de surcroît pratiquée par un nombre limité de spécialistes, rendent le diagnostic et le suivi impossible pour la grande majorité de la population. L'objectif du doctorat est de travailler sur des signaux enregistrés par des moyens non invasifs (et en particulier ceux provenant du collier SWALLIS DSA® : enregistrement par un microphone et des accéléromètres placés au niveau de la gorge). Le travail demandé consistera dans un premier temps à identifier les indicateurs des quatre fonctions pharyngo laryngées (respiration, déglutition, protection des voies aériennes, phonation) en

relation avec le mécanisme de déglutition. Et dans un second temps, une modélisation spécifique sur le rôle fondamental des réflexes pharyngo-laryngés permettra à ces indicateurs de mesurer l'efficacité pharyngo laryngée.

6.3 Encadrements de Post-Doctorants

6.3.1 Eduardo Sanchez-Soto (01/07/2007 – 01/07/2008)

Eduardo Sanchez-Soto a été recruté sur le projet ANR MISTRAL (cf. §5.4.6 page 78). Il a réalisé l'adaptation de la plateforme pour les tâches de vérification de la langue. Il a ensuite participé à la campagne d'évaluation NIST LRE 2007 (cf. §5.5.3 page 88) pour valider le développement [Sanchez-Soto and Farinas, 2007].

6.3.2 Hélène Lachambre (01/02/2013 – 31/06/2013)

Hélène Lachambre a réalisé un état de l'art techniques de mesure de l'intelligibilité sur le projet Archean (cf. §5.4.9 page 79) et a permis de poser les définitions des termes intelligibilité et compréhensibilité.

6.3.3 Lionel Koenig (01/10/2013 – 31/12/2013)

Lionel Koenig a travaillé sur le développement d'un système de reconnaissance de la parole grand vocabulaire à partir du travail qui avait été réalisé sur la campagne d'évaluation ESTER (cf §5.5.2 page 88).

6.3.4 Cynthia Magnen (01/09/2014 – 31/11/2014)

Cyntia Magnen a eu pour mission de la réalisation de l'étiquetage phonétique du corpus de mesure de l'intelligibilité dans le cadre du projet Archean (cf. §5.4.9 page 79). Cela a permis de réaliser les alignements et de permettre l'étude du score de mesure de l'intelligibilité avec les productions automatiques.

6.3.5 Lionel Fontan (01/02/2014 – 31/05/2015)

Lionel Fontan a été recruté comme postdoctorant pour le travail sur la création de mesure automatique de l'intelligibilité sur le projet Archean (cf. §5.4.9 page 79). Il avait réalisé son doctorat « De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication ». Après une formation sur les systèmes automatiques de la parole, il a mené la réalisation du score automatique d'intelligibilité [Fontan et al., 2015b, Fontan et al., 2015a, Fontan et al., 2016, Fontan et al., 2017]. Il a ensuite réalisé le transfert technologique auprès de la société Archean et été recruté dans la société.

6.3.6 Oriol Pont (01/03/2016 – 16/11/2016)

Oriol Pont a été recruté sur le projet C2SI (cf. §5.4.12 page 80). Oriol a mené une étude de caractéristiques pertinentes pour l'intelligibilité de la parole par approches non linéaires. En effet ses précédentes

recherches portaient sur des analyses nouvelles basées sur des formulations microcanonique multi-échelle permettant la description géométrique et statistique des propriétés multi-échelles de la dynamique complexe des signaux de parole [Khanagha et al., 2014]. Oriol a mis en place des paramétrisations basée sur la projection suivant la base de Patrick Hanusse [Hanusse, 2010] permettant la décomposition des signaux anharmoniques.

6.3.7 Imed Laaridth (01/12/2017 – 31/11/2018)

Imed Laaridth a été recruté comme post-doctorant sur le projet PHONICS (cf. §5.4.10 page 79). Il a soutenu son doctorat sur l'« Évaluation de la parole dysarthrique : Apport du traitement automatique de la parole face à l'expertise humaine » à l'université d'Avignon.

Imed a transposé les résultats du projet Archean (cf. §5.4.9 page 79) sur la parole altérée par presbyacousie. Les résultats ont été validés sur les troubles de l'audition liés à des surdités professionnelles [Laaridh et al., 2018b].

Il a réalisé le développement d'un système de reconnaissance automatique de la parole anglaise basée sur la plateforme KALDI [Povey et al., 2011a] et les données Mozilla Common Voice [Mozilla, 2018]. Il a réussi à réaliser une mesure de l'intelligibilté sur l'anglais, bien que le corpus soit très différent de celui utilisé en français et que KALDI ne permette pas d'utiliser des probabilités acoustiques et linguistiques comme les autres systèmes de reconnaissance (tel que CMU Sphinx [Deléglise et al., 2005] ou HTK [Young and Young, 1993]). Les résultats de la tâche sur l'anglais obtiennent de bonne corrélations avec ceux de perception humaine [Laaridh et al., 2018a].

6.3.8 Mathieu Balaguer (01/09/2022 – 31/12/2023)

Mathieu Balaguer travaille a mi-temps pour de l'enseignement dans l'école d'orthophonie de Toulouse et est recruté à mi-temps à l'IRIT pour poursuivre ses travaux de recherche.

Il continue à gérer la supervision du projet DAPADAF-E (cf. §5.4.17 page 84).

Il a réalisé en 2022 les formations auprès des équipes de recherche en linguistique dans le cadre du projet AADI (cf. §5.4.17 page 84).

Il est en ce moment en train d'adapter des systèmes de reconnaissance de la parole grand vocabulaire à la parole de personnes ayant des cancers de la tête et du cou dans le cadre du projet ADAPT (cf. §5.4.22 page 87). Pour cela, avec le support de plusieurs stages d'orthophonie, il augmente les annotations du corpus SpeeComCo, corpus contenant les enregistrements des entretiens qu'il a réalisés pendant son doctorat.

7 Autres activités et responsabilités

- Secrétaire du Collectif de Doctorants Toulousain (association de loi 1901) de 2000 à 2002. J'ai participé activement à la mise en place du Forum des Écoles Doctorales, à la réalisation d'un Guide du Doctorant Toulousain, à l'animation de l'association et à sa représentation au sein de la Confédération des Jeunes Chercheurs.
- Élu au collège B du Conseil d'Administration de l'Université Paul Sabatier (avril 2002 à avril 2004)
- Élu au conseil de l'UFR Mathématique Informatique et Gestion en 2003

- Élu au collège B du conseil de laboratoire IRIT (2010-2014) : participation aux commissions doctorants et bibliothèque, participation au groupe de travail sur la mise à jour du règlement intérieur
- Élu au Conseil d'Administration de l'Association Francophone de la Communication Parlée (AFCP) de 2005 à 2008
- Président de l'association loi 1901 « Les Amis de la Mer » (ADLM) de 2005 à 2007 (association de plongée sous marine en contrat avec le service des sports de l'UT3), Directeur Technique de 2008 à 2009. Moniteur Fédéral 2ème degré (MF2) en juillet 2012 à Hendaye (promotion Guy Poulet).
- Élu au conseil de département informatique de l'université Paul Sabatier de septembre 2016 à septembre 2021
- Responsable avec Sylvain Cussat-Blanc du Domaine d'Application Stratégique (DAS) Santé, Autonomie, Bien-Être de l'IRIT de décembre 2018 à mars 2021 : préparation des conseils mensuels, représentation du laboratoire sur les sujets du DAS, organisation de manifestations, gestion du budget, point de contact de l'IRIT sur les sujets du DAS.
- Élu personnalité extérieure du Conseil du Département des Sciences du langage de l'UFR des Langues, Littératures et Civilisations Étrangères de l'Université Toulouse 2 Jean Jaurès depuis juin 2019
- Correspondant local du GDR-TAL [26] pour la composante traitement automatique de la parole à l'IRIT depuis septembre 2019
- Tournage d'une capsule vidéo intitulée « le traitement automatique de la parole » pour la promotion du département Signaux et Images de l'IRIT le 4 juillet 2022 [Farinas, 2022]
- Correspondant de l'Institut Carnot Cognition depuis 2020. Je suis chargé de diffuser les offres de ressourcement et les Appels à Manifestation d'intérêt. Mais je dois également faire remonter les rapports sur les finances et les publications de l'unité par rapport à la Cognition. J'ai organisé une journée de rencontre entre les chercheurs du CRCA et de l'IRIT. J'ai mobilisé les chercheurs sur les rapports d'activité du Carnot. J'ai tourné pour une capsule vidéo traitant de l'« Évaluation objective de la compréhensibilité d'un patient ayant des difficultés à s'exprimer » le 23 mai 2023 [Farinas, 2023]
- Nommé en 2021 par la direction de l'IRIT en tant que chargé de mission sur les partenariats et les relations industrielles. L'objectif de la mission est d'assurer l'interface entre le monde industriel et le laboratoire. Je suis à cet égard chargé de recevoir les entreprises et les diriger vers les chercheurs les plus appropriés de la structure. Mais la mission concerne également un travail auprès des chercheurs et enseignants-chercheurs dans le but de favoriser les partenariats ou les créations d'entreprises. Je suis amené à intervenir à l'Assemblée des Responsables de Structures, au Conseil Scientifique et au Conseil de Laboratoire. Je me réuni avec la direction tous les deux mois pour organiser un suivi de l'activité. J'ai été amené à rédiger des notices d'information sur le processus CIFRE, et les différents moyens pour réaliser du partenariat. Ma mission m'amène également à rencontrer les différents partenaires régionaux et nationaux (CNRS) en charge du partenariat. La figure 8 détaille la cartographie des contacts développés.
- Membre fondateur du Groupement d'Intérêt Scientifique « Parolothèque » (cf. §5.4.20 page 86). Je participe activement au montage du groupement. Le projet a été signé par tous les établissements en 2022. Ce projet permet de coordonner les projets réunissant les communautés de scientifiques travaillant sur le traitement automatique et les médecins oncologie et spécialistes ORL. L'objectif est de constituer une base de données constituée d'enregistrements de personnes atteintes de cancers et d'enquêtes épidémiologiques pour des personnes ayant été atteints de cancer. Ce projet réunit le CHU de Toulouse et les laboratoires LORDAT/UT2J (Toulouse), IRIT/UT3 (Toulouse),

26. https://gdr-tal.ls2n.fr

LPL (Aix-en-provence), LIA (Avignon). Les retombées concernent à la fois la recherche sur la reconnaissance automatique de la parole pathologique mais également l'avancée épidémiologique concernant le cancer. L'objectif est de fédérer à terme au niveau national et européen la constitution de base de données de personnes cancéreuses.

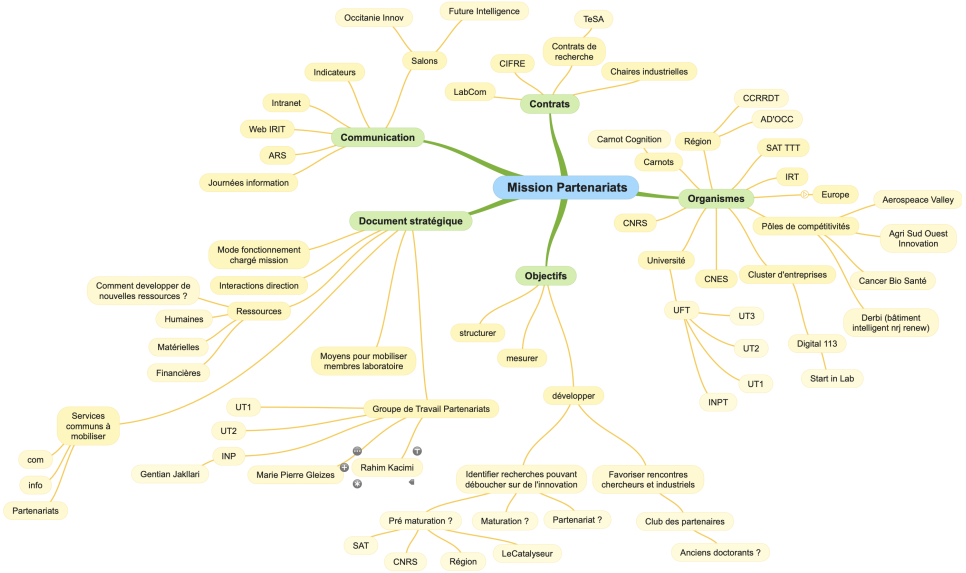


FIGURE 8 – Présentation générale de la mission partenariat de l'IRIT

Annexe A

Publications

Cette annexe reprend la copie des cinq publications principales demandées par la commission des habilitations de l'université Paul Sabatier.

A.1 Article dans le journal Speech Communication

Article intitulé « Rhythmic unit extraction and modelling for automatic language identification » publié dans le numéro 45 du journal Speech Communication en 2005 [Rouas et al., 2005]. Cet article détaille la problématique de la modélisation du rythme développé dans mon doctorat et une application en identification des langues sur le corpus MULTEXT [Ide and Véronis, 1994] et son extension au japonais [Shigeyoshi et al., 2004].

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 436–456
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Rhythmic unit extraction and modelling for automatic language identification

Jean-Luc Rouas [a], Jérôme Farinas [a],
François Pellegrino [b,*], Régine André-Obrecht [a]

[a] *Institut de Recherche en Informatique de Toulouse UMR 5505 CNRS, Université Paul Sabatier, 31062 Toulouse Cedex 9, France*
[b] *Laboratoire Dynamique Du Langage UMR 5596 CNRS, Université Lumière Lyon 2, 14, Avenue Berthelot, 69363 Lyon Cedex 7, France*

Received 23 July 2004; received in revised form 21 April 2005; accepted 26 April 2005

Abstract

This paper deals with an approach to automatic language identification based on rhythmic modelling. Beside phonetics and phonotactics, rhythm is actually one of the most promising features to be considered for language identification, even if its extraction and modelling are not a straightforward issue. Actually, one of the main problems to address is *what* to model. In this paper, an algorithm of rhythm extraction is described: using a vowel detection algorithm, rhythmic units related to syllables are segmented. Several parameters are extracted (consonantal and vowel duration, cluster complexity) and modelled with a Gaussian Mixture. Experiments are performed on read speech for seven languages (English, French, German, Italian, Japanese, Mandarin and Spanish) and results reach up to 86 ± 6% of correct discrimination between stress-timed mora-timed and syllable-timed classes of languages, and to 67 ± 8% of correct language identification on average for the seven languages with utterances of 21 s. These results are commented and compared with those obtained with a standard acoustic Gaussian mixture modelling approach (88 ± 5% of correct identification for the seven languages identification task).

Keywords: Rhythm modelling; Language identification; Rhythm typology; Asian languages; European languages

1. Introduction

Automatic language identification (ALI) has been studied for almost 30 years, but the first competitive systems appeared during the 90s. This recent attention is related to (1) the need for Human–Computer Interfaces and (2) the

* Corresponding author. Tel.: +33 4 72 72 64 94; fax: +33 4 72 72 65 90.
E-mail addresses: rouas@irit.fr (J.-L. Rouas), jerome.farinas@irit.fr (J. Farinas), francois.pellegrino@univ-lyon2.fr (F. Pellegrino), obrecht@irit.fr (R. André-Obrecht).

0167-6393/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.specom.2005.04.012

remarkable expansion of multilingual exchanges. Indeed, in the so-called information society, the stakes of ALI are numerous, both for multilingual Human–Computer Interfaces (Interactive Information Terminal, Speech dictation, etc.) and for Computer-Assisted Communication (Emergency Service, Phone routing services, etc.). Moreover, accessing the overwhelming amount of numeric audio (or multimedia) data available may take advantage from content-based indexing that may include information about the speakers' languages or dialects. Besides, linguistic issues may also be addressed: the notion of linguistic distance has been implicitly present in linguistics typology for almost a century. However, it is still difficult to define, and ALI systems may shed a different light on this notion since correlating automatic, perceptual and linguistic distances may lead to a renewal of the typologies and to a better understanding of the close notions of languages and dialects.

At present, state-of-the-art approaches consider phonetic models as *front-end* providing sequences of discrete phonetic units decoded later in the system, according to language-specific statistical grammars (see Zissman and Berkling, 2001 for a review). The recent NIST 2003 Language Recognition Evaluation (Martin and Przybocki, 2003) has confirmed that this approach is quite effective since the error rate obtained on a language verification task using a set of 12 languages is under 3% for 30-s utterances (Gauvain et al., 2004). However, other systems modelling global acoustic properties of the languages are also very efficient, and yield about 5% error on the same task (Singer et al., 2003). These systems, that take advantage either of speech or speaker recognition techniques, perform quite well. Still, very few systems are trying to use other approaches (e.g. prosodics) and results are much poorer than those obtained with the phonetic approach (for example the combination of the standard OGI "temporal dynamics" system based on a *n*-gram modelling of sequences of segments labelled according their F0 and energy curves yields about 15–20% of equal error rate with three languages of the NIST 2003 campaign task and corpus (Adami and Hermansky, 2003)). However, these alternative approaches may lead to improvements, in terms of robustness in noisy conditions, number of languages recognized or linguistic typology. Further research efforts have to be made to overcome the limitations and to assess the contributions of those alternative approaches.

The motivations of this work are given in Section 2. One of the most important is that prosodic features carry a substantial part of the language identity that may be sufficient for humans to perceptually identify some languages (see Section 2.2). Among these supra-segmental features, rhythm is very promising both for linguistic and automatic processing purposes (Section 2). However, coping with rhythm is a tricky issue, both in terms of theoretical definition and automatic processing (Section 3). For these reasons, the few previous experiments which aimed at language recognition using rhythm were based on hand-labelled data and/or have involved only tasks of language discrimination[1] (Thymé-Gobbel and Hutchins, 1999; Dominey and Ramus, 2000). This paper addresses the issue of automatic rhythm modelling with an approach that requires no phonetically labelled data (Section 4). Using a vowel detection algorithm, rhythmic units somewhat similar to syllables and called *pseudo-syllables* are segmented. For each unit, several parameters are extracted (consonantal and vowel duration, cluster complexity) and modelled with a Gaussian Mixture. This approach is applied to seven languages (English, French, German, Italian Japanese, Mandarin and Spanish) using the MULTEXT corpus of read speech. Descriptive statistics on pseudo-syllables are computed and the relevancy of this modelling is assessed with two experiments aiming at (1) discriminating languages according to their rhythmic classes (stress-timed vs. mora-timed vs. syllable-timed) and (2) identifying the seven languages. This rhythmic approach is then compared to a more standard acoustic approach (Section 5).

From a theoretical point of view, the proposed system focuses on the existence and the modelling of rhythmic units. This approach generates a type of segmentation that is closely related to a syllabic

[1] *Language discrimination* refers to determining to which of two candidate languages L_1–L_2 an unknown utterance belongs to. *Language identification* denotes more complex tasks where the number of candidate languages is more than two.

parsing of the utterances. It leaves aside other components of rhythm related to the sequences of rhythmic units or that span over whole utterances. These considerations are discussed in Section 6.

2. Motivations

Rhythm is involved in many processes of the speech communication. Though it has been neglected for long, several considerations lead to reconsider its role both in understanding and production processes (Section 2.1), and especially in a language identification framework (Section 2.2). Moreover, researchers have tried to take rhythm into consideration for automatic processing purposes for a while, both in speech synthesis and recognition tasks, leading to several rhythm-oriented approaches (Section 2.3). All these considerations emphasize both the potential use of an efficient rhythm model and the difficulty to elaborate it. It leads us to focus on the possible use of rhythmic features for ALI (Sections 3 and 4).

2.1. Linguistic definition and functions of rhythm

Rhythm is a complex phenomenon that has long been said to be a consequence of other characteristics of speech (phonemes, syntax, intonation, etc.). However, an impressive amount of experiments tends to prove that its role may be much more than a mere side effect in the speech communicative process.

According to the Frame/Content theory (MacNeilage, 1998; MacNeilage and Davis, 2000), speech production is based on superimposing a segmental content into a cyclical frame. From an evolutionary point of view, this cycle probably evolved from the ingestive mechanical cycles shared by mammals (e.g. chewing) via intermediate states including visuofacial communication controlled at least by a mandibular movement (lipsmacks, etc.). Moreover, the authors shed light on the status of the syllable both as an interface between segments and suprasegmentals and as the *frame*, a central concept in their theory: convoluting the mandibular cycle with a basic voicing production mechanism results in a sequence of CV syllables composed of a closure and a neutral vowel. Additional experiments on serial ordering errors made by adults or children (e.g. Fromkin, 1973; Berg, 1992) and child babbling (MacNeilage et al., 2000; Kern et al., in press) are also compatible with the idea that the mandibular oscillation provides a rhythmic baseline in which segments accurately controlled by articulators take place.

A huge amount of psycholinguistics studies also draw attention to the importance of the rhythmic units in the complex process of language comprehension. Most of them consider that a rhythmic unit—roughly corresponding to the syllable combined with an optional stress pattern—plays an important role as an intermediate level of perception between the acoustic signal and the word level. The exact role of these syllables or syllable-sized units has still to be clearly identified: whether the important feature is the unit itself (as a recoding unit) or its boundaries (as milestones for the segmentation process) is still in debate. The ones claim that the syllable is the main unit in which the phonetic recoding is performed before lexical access (Mehler et al., 1981). The others propose an alternative hypothesis in which syllables and/or stress provide milestones to parse the acoustic signal into chunks that are correctly aligned with the lexical units (Cutler and Norris, 1988). In this last framework, the boundaries are more salient than the content itself, and no additional hypothesis is made on the size of the units actually used for lexical mapping. Furthermore, recent experiments point out that the main process may consist in locating the onset rather than raw boundary detection (Content et al., 2000, 2001).

These studies show that rhythm plays a key role in the speech communication process. Similarly, several complementary aspects could have been mentioned but they are beyond the scope of this paper.[2] However, several questions regarding the nature of the rhythm phenomenon are still open. First of all and as far as the authors know, an

[2] See Levelt and Wheeldon (1994) for his model of speech production (see also Boysson-Bardies et al., 1992; Mehler et al., 1996; Weissenborn and Höhle, 2001; Nazzi and Ramus, 2003 for the role of rhythm in early acquisition of language).

uncontroversial definition of rhythm does not exist yet even if most researchers may agree on the notion that speech rhythm is related to the existence of a detectable phenomenon that occurs evenly in speech. Crystal proposes to precisely define rhythm as "the regular perception of prominent units in speech" (Crystal, 1990). We prefer not to use the concepts *of perception* and *unit* because they narrow the rhythmic phenomenon with a priori hypotheses: according to Crystal's definition, rhythm can be considered as the alternation of prominent units with less prominent ones, but defining those units is far from straightforward; The alternation of stressed/unstressed syllables results in one kind of rhythm, but the voiced/unvoiced sound sequences may produce another type of rhythm, and so do consonant/vowel alternations or short/long sound sequences, etc. Moreover, rhythm may arise from the even occurrence of punctual *events* and not *units* (like beats superimposed on other instruments in music).

Another question concerns the actual role of the *syllable*. Whether it is a cognitive unit or not is still in debate. Though, several experiments and measures indicate that syllables or syllable-sized units are remarkably salient and may exhibit specific acoustic characteristics. Since the early 1970s, several experiments have indicated that the human auditory system is especially sensitive to time intervals spanning from 150 to 300 ms clearly compatible with average syllable duration.[3] These experiments, based on various protocols (forward and backward masking effect, ear switching speech, shadowing repetition, etc.) showed that this duration roughly corresponds to the size of a human perceptual buffer (see for example Massaro, 1972; Jestead et al., 1982; O'Shaugnessy, 1987). More recently, experiments performed with manipulated spectral envelopes of speech signals showed the salience of the modulation frequencies between 4 and 6 Hz in perception (Drullman et al., 1994). Hence, all these findings support the syllable as a relevant rhythmic unit. In addition, acoustic measurements made on a corpus of English spontaneous speech emphasize also its prominence (Greenberg, 1996, 1998). This study showed that, as far as spectral characteristics are concerned, syllable onsets are in general less variable than nuclei or codas. It also highlights that co-articulation effects are much larger within each syllable than between syllables. Both effects result in the fact that syllable onsets vary less than other parts of the signal and consequently may provide at least reliable anchors for lexical decoding. Besides this search for the intrinsic *nature* of rhythm, perceptual studies may also improve our knowledge of its intrinsic *structure*. Using speech synthesis to simulate speech production, Zellner-Keller (2002) concluded that rhythm structure results from a kind of convolution of a temporal skeleton with several layers, from segmental to phrasal, in a complex manner that can be partially predicted.

One of the main conclusions is that temporal intervals ranging from 150 to 300 ms are involved in speech communication as a relevant level of processing. Moreover, many cues draw attention to this intermediate level between acoustic signal and high level tiers (syntax, lexicon). At this moment, it is not evident to assess if the relevant feature is actually a rhythmic *unit* by itself or a rhythmic *beat*. However syllable-sized units are salient from a perceptual point of view and may have acoustic correlates that facilitate their automatic extraction. Next section deals with the experimental assessment of these correlates in perceptive language identification tasks.

2.2. Rhythm and perceptual language identification

Language identification is an uncommon task for many adult human speakers. It can be viewed as an entertaining activity by the most questioning ones but most adult human beings living in a monolingual country may consider that it is of no interest. However, the situation is quite different in multilingual countries where numerous languages or dialects may be spoken on a narrow geographical area. Furthermore, perceptual language identification is an essential challenge for children who acquire language(s) in that kind of multilingual context: it is then utterly important for them

[3] Greenberg (1998) reports a mean duration of 200 ms for spontaneous discourse on the Switchboard English database.

to distinguish which language is spoken in order to acquire the right language-dependent phonology, syntax, and lexicon. During the last two decades, several experiments have investigated the efficiency of the human being as a language recognizer (see Barkat-Defradas et al., 2003, for a review). Three major types of features may help someone to identify a language: (1) segmental features (the acoustic properties of phonemes and their frequency of occurrence), (2) supra-segmental features (phonotactics, prosody), and (3) high level features (lexicon, morpho-syntax). The exact use made of each set of features is unclear yet and it may actually differ between newborn children and adults.

For example, several experiments have proved that newborns, as early as the very first days, are able to discriminate between their mother tongue and some foreign languages that exhibit differences at the supra-segmental level (see Ramus, 2002a,b, for a review). Whether newborns take advantage from rhythm alone or from both rhythm and intonation is an open issue. It is likely that both levels provide cues that are weighted as function of the experimental conditions (languages, noise, and speech rate) and maybe according to individual strategies. Assessing these adult human capacities to identify foreign languages is a complex challenge since numerous parameters may influence this ability. Among them, the subject's mother tongue and his personal linguistic history seem to be key factors that prove difficult to quantify. Since the end of the 1960s, quite a few studies have tackled this question. Depending on whether they are implemented by automatic speech processing researchers or linguists, the purposes differ. The former intend to use these perceptual experiments as benchmarks for ALI systems, while the latter investigate the cognitive process of human perception. More recently, this kind of experiments has been viewed as a way to investigate the notion *of perceptual distance* among languages. In this framework, the aim is to evaluate the influence of the different levels of linguistic description in the cognitive judgment of language proximity.

From a general point of view all these experiments have shown the noteworthy capacity of human subjects to identify foreign languages after a short period of exposure. For example, one of the experiments reported by Muthusamy et al. (1994) indicates that native English subjects reach a score of 54.2% of correct answers when identifying 6-s excerpts pronounced in nine foreign languages. Performances varied significantly from one language to another, ranging from 26.7% of recognition for Korean to 86.4% of recognition for Spanish. Additionally, subjects were asked to explain which cues they had considered to make their decision. Their answers revealed the use of segmental features (manner and place of articulation, presence of nasal vowels, etc.), supra-segmentals (rhythm, intonation, tones) and "lexical" cues (iteration of the same words or pseudo-words). However these experiments raise numerous questions about the factors influencing the recognition capacity of the subjects: the number of languages that they have been exposed to, the duration of the experimental training, etc. Following Muthusamy, several researchers have tried to quantify these effects. Stockmal, Bond and their colleagues (Stockmal et al., 1996; Stockmal et al., 2000; Bond and Stockmal, 2002) have investigated several socio-linguistic factors (geographical origin of the speakers, languages known by the subjects, etc.) and linguistic factors (especially rhythmic characteristics of languages). In a similar task based on the identification of Arabic dialects our group has shed light on the correlation between the structure of the vocalic system of the dialects and the perceptual distances estimated from the subjects' answers (Barkat-Defradas et al., 2003). The results reported by (Vasilescu et al., 2000) in an experiment of discrimination between romance languages may be interpreted in a similar way. Other studies focus on the salience of supra-segmentals in perceptual language identification. From the first experiments of Ohala and Gilbert (1979) to the recent investigations of Ramus, using both natural and synthesized speech, they prove that listeners may rely on phonotactics, rhythm, and intonation patterns to distinguish or identify languages, even if segmental information is lacking.

Even if the cognitive process leading to language identification is multistream (from segmental acoustics to suprasegmentals and higher level cues), no model of integration has been derived

yet. Moreover, building such a model seems to be still out of range since even the individual mechanisms of perception at each level are still puzzling. At the segmental level, most researchers are working with reference to the motor theory of speech perception (Liberman and Mattingly, 1985) searching arguments that would either confirm or invalidate it. At the suprasegmental level, the perception of rhythm has been mainly studied from a musical point of view, even if comparisons between music and speech perception are also studied (e.g. Todd and Brown, 1994; Besson and Schön, 2001) and if technological applications (e.g. speech synthesis) have lead researchers to evaluate rhythm (see next section).

2.3. Rhythm and syllable-oriented automatic speech processing

Many studies aiming at taking advantage from rhythmic and prosodic features for automatic systems have been developed through the last decades and achieved most of the time disappointing results. Nevertheless several authors consider that this is a consequence of the difficulty to model suprasegmental information and put forward the major role of prosody and temporal aspects in speech communication processes (see for example Zellner-Keller and Keller, 2001 for speech synthesis and Taylor et al., 1997 for speech recognition).

Beside its role in the parsing of sentence into words (Cutler and Norris, 1988; Cutler, 1996), prosody constitutes sometimes the only means to disambiguate sentences, and it often carries additional information (mood of the speaker, etc.). Even when focusing on the acoustic–phonetic decoding, suprasegmentals may be relevant at two levels: first of all, segmental and suprasegmental features are not independent, and thus, the suprasegmental level may help to disambiguate the segmental level (e.g. see the correlation between stress accent and pronunciation variation in American English (Greenberg et al., 2002)). Moreover, as it has been argued above, suprasegmentals and especially rhythm, may be a salient level of treatment as itself for humans and probably for computational models. Speech synthesis is an evident domain where perceptual experiments have shown the interest of syllable-length units for the naturalness of synthesized speech (Keller and Zellner, 1997). Additionally, rhythm and rhythmic units may play a major role in Automatic Speech Recognition: from the proposal of the syllable as a unit for speech recognition (Fujimura, 1975) to the summer workshop on "Syllable Based Speech Recognition" sponsored by the Johns Hopkins University (Ganapathiraju, 1999), attempts to use rhythmic units in automatic speech recognition and understanding have been numerous. Disappointingly, most of them failed to improve the standard speech recognition approach based on context-dependent phone modelling (for a review, see Wu, 1998). However, the definitive conclusion is not that suprasegmentals are useless, but instead, that the phonemic level may not be the suitable time scale to integrate them and that larger scales may be more efficient. We have already mentioned that co-articulation effects are much greater *within* each syllable than *between* syllables in a given corpus of American English spontaneous speech (Greenberg, 1996). Context-dependent phones are well-known to efficiently handle this co-articulation. However, their training needs a big amount of data, and consequently they cannot be used when few data are available (this happens especially in multilingual situations): the state-of-the-art systems of ALI are based on context-independent phones (Singer et al., 2003; Gauvain et al., 2004). Thus, syllable-sized models are a promising alternative with limited variability at the boundaries. However, several unsolved problems limit the performance of the current syllable-based recognition systems and the main problem may be that syllable boundaries are not easy to identify, especially in spontaneous speech (e.g. Content et al., 2000 for a discussion on ambisyllabicity and resyllabification). Thus, combining phoneme-oriented and syllable-oriented models in order to take several time scales into account may be a successful approach to overcome the specific limits of each scale (Wu, 1998). Finally, syllable-oriented studies are less common in the fields of speaker and language identification. Among them, we can however distinguish approaches adapted from standard phonetic or phonotactic approaches to syllable-sized units (Li, 1994 and

more recently Antoine et al., 2004 for a "syllabotactic" approach and Nagarajan and Murthy, 2004 for a syllabic Hidden Markov Modelling) from those trying to model the underlying rhythmic structure (see Section 4.1).

This section showed that: (1) Rhythm is an important mechanism of speech communication involved in comprehension and production processes; (2) It is difficult to define, to handle, and most of all, to efficiently model; (3) Syllable or syllable-like units may play an important role in the structure of rhythm. Furthermore, experiments reported above clearly demonstrate that different languages may be different from the rhythmic perspective and that these differences may be perceived and used in a perceptual language identification task. Next section deals with these differences, both in terms of linguistic diversity and its underlying acoustic parameters.

3. The rhythm typology and its acoustic correlates

Languages can be labelled according to a rhythm typology proposed by linguists. However, rhythm is complex and some languages do not perfectly match this typology and the search for acoustic correlates has been proposed to evaluate this linguistic classification.

Experiments reported here focus on five European languages (English, French, German, Italian and Spanish) and two Asian languages (Mandarin and Japanese). According to the linguistic literature, French, Spanish and Italian are syllable-timed languages while English and German are stress-timed languages. Regarding Mandarin, classification is not definitive but recent works tend to affirm that it is a stress-timed language (Komatsu et al., 2004). The case of Japanese is different since it is the prototype of a third rhythmic class, namely the mora-timed languages for which timing is related to the frequency of morae.[4] These three categories are related to the notion of isochrony and they emerged from the theory of rhythm classes introduced by Pike, developed by Abercrombie (1967) and enhanced with mora-timed class by Ladefoged (1975). More recent works, based on the measurement of the duration of inter-stress intervals in both stress-timed and syllable-timed languages provide an alternative framework in which these discrete categories are replaced by a continuum (Dauer, 1983) where rhythmic differences among languages are mostly related to their syllable structure and the presence (or absence) of vowel reduction.

The syllable structure is closely related to the phonotactics and to the accentuation strategy of the language. While some languages will allow only simple syllabic patterns (CV or CVC), other will permit much more complex structures for the onset, the coda or both (e.g. syllables with up to six consonants in the coda[5] are encountered in German). Table 1, adapted from (Greenberg, 1998) displays a comparison of the syllabic forms from spontaneous speech corpora in Japanese and American English.

The most striking statement is that in both languages, the CV and CVC forms stand for nearly 70% of the encountered syllables. However, the other forms reveal significant differences in the syllabic structure. On the one hand, consonantal clusters are rather common in American English (11.7% of the syllables) while they are almost absent from the Japanese corpus. On the other hand, VV transitions are present in 14.8% of the Japanese syllables while they could only occur by resyllabification at word boundaries in English. These observations roughly correspond with our knowledge of the phonological structure of the words in those two languages. However, the nature of the corpora (spontaneous speech) widely influences the relative distribution of each structure. With read speech (narrative texts), Delattre and Olsen (1969) found fairly different patterns for British English: CVC (30.1%), CV (29.7%), VC (12.6%), V (7.4%) and CVCC (7%). CCV that occurs 5.1% in the Switchboard corpus represents

[4] Morae can consist of a V, CV or C. For instance, [kakemono] (scroll) and [nipponŋ] (Japan) must both be divided in four morae: [ka ke mo no] and [ni p po ŋ] (Ladefoged, 1975, p. 224).

[5] For instance, "you shrink it" will be translated *du schrumpfst's* [duː ʃrʊmpfsts]. This example is taken from (Möbius, 1998).

Table 1
The 10 most common syllabic forms and their frequency of occurrence in Japanese and English

Japanese		English	
Form	% of occurrence	Form	% of occurrence
CV	60.4	**CV**	47.2
CVC	17.9	**CVC**	22.1
CVV	11.7	**V**	11.2
V	2.9	**CCV**	5.1
CCV	1.7	**VC**	4.8
CVVC	1.3	CVVC	2.9
CCVV	1.3	**CCVC**	2.5
VC	1.2	VCC	0.5
VV	0.5	CCVCC	0.4
CCVC	0.4	CCCV	0.3
Other	0.7	Other	3.0

Frequencies are computed on two spontaneous speech corpora. Form in bold are encountered in both languages (adapted from Greenberg, 1998).

only 0.49% of the syllables in the Delattre and Olson corpus. However, statistics calculated on the Switchboard corpus show that 5000 different syllables are necessary to cover 95% of the vocabulary[6] (Greenberg, 1997) and thus that inter-language differences are not restricted to high-frequency syllabic structures. These broad phonotactic differences explain at least partially the mora-time vs. stress-time opposition. Still, studying the temporal properties of languages is necessary to determine whether the rhythm is totally characterized by syllable structures or not.

Beyond the debate on the existence of rhythmic classes (opposed to a rhythmic continuum), the measurement of the acoustic correlates of rhythm is essential for automatic language identification systems based on rhythm. The first statistics made by Ramus, Nespor and Mehler with an ad hoc multilingual corpus of eight languages led to a renewal of interest for these studies (Ramus et al., 1999). Following Dauer, they searched for duration measurements that could be correlated with vowel reduction (resulting in a wide range of duration for vowels) and with the syllable structure. They came up with two reliable parameters: (1) the percentage of vocalic duration %V and (2) the standard deviation of the duration of the consonant intervals ΔC both estimated over a whole utterance. They provided a 2-dimension space in which languages are clustered according to their rhythm class.[7] These results are very promising and prove that in nearly ideal conditions (manual labelling, homogeneous speech rates, etc.), it is possible to find acoustic parameters that cluster languages into explainable categories. The extension of this approach to ALI necessitates the evaluation of these parameters with more languages and less constrained conditions. This raises several problems that can be summarized as follows:

– Adding speakers and languages will add inter-speaker variability. Would it result in an overlap of the language-specific distributions?
– Which part of the duration variation observed in rhythmic unit is due to language-specific rhythm and which part is related to speaker-specific speech rate?
– Is it possible to take these acoustic correlates into account for ALI?

A recent study (Grabe and Low, 2002) answers partially to the first question. Considering 18 languages and relaxing constraints on the speech rate, Grabe and Low have found that the studied languages spread widely without visible clustering effect in a 2-dimension space somewhat related to the %V/ΔC space. However, in their study, each language is represented by only 1 speaker, which prevents from drawing definite conclusion on the discrete or continuous nature of the rhythm space. Addressing the variability issue between speakers, dialects and languages, similar experiments focusing on dialects are in progress in our group (Hamdi et al., 2004; Ferragne and Pellegrino, 2004). Though it is beyond the scope of this paper, the second question is essential. Speech rate involves computing a number of certain *units* per second; choosing the appropriate unit(s) remains controversial (syllable, phoneme or morpheme)

[6] This number falls to 2000 syllables necessary to cover 95% of the corpus (i.e. taking into account the frequency of occurrence of each word of the vocabulary).

[7] Actually, the clustering seems to be maximum along one dimension derived from a linear combination of ΔC and %V.

and so is the interpretation of the measured rate: few units per second means long units, but does it mean that the units are intrinsically long or is the speaker an especially slow speaker? Moreover, the *variation* of speech rate within an utterance is also relevant: the speaking rate of a hesitating speaker may switch from local high values to very low values during disfluencies (silent or filled pauses, etc.) along a single utterance. Consequently, fast variations may be masked according to the time span used for the estimation and the overall speech rate estimation may not be relevant. Besides, the estimation of speech rate is also relevant for automatic speech recognition, since recognizers' performances usually decrease when they come to dealing with especially fast or slow speakers (Mirghafori et al., 1995). For this reason, algorithms exist to estimate either phone rate or syllable rate (e.g. Verhasselt and Martens, 1996; Pfau and Ruske, 1998). However, the subsequent normalization is always applied in a monolingual context, and no risk of masking language specific variation can occur. At present, the effect of this kind of normalization in a multilingual framework has not been studied extensively though it will be essential for ALI purposes. Our group has elsewhere addressed this issue in a study of inter-languages differences of speech rate in terms either of syllables per second or phonemes per second (Pellegrino et al., 2004; Rouas et al., 2004).

The last question is the main issue addressed in this paper. This section assesses the existence of acoustic correlates of the linguistic rhythmic structure. However, whether they are detectable and reliable enough to perform ALI or not is to be tackle. The following sections thoroughly focus on this issue.

4. Rhythm modelling for ALI

4.1. Overview of related works

The controversies about the status of rhythm illustrate the difficulty to segment speech into meaningful rhythmic units and emphasize that a global multilingual model of rhythm is a long range challenge. As a matter of fact, even if correlates between speech signal and linguistic rhythm exist, developing a relevant representation of it and selecting an appropriate modelling paradigm is still at stakes.

Among others, Thymé-Gobbel and Hutchins (1999) have emphasized the importance of rhythmic information in language identification systems. They developed a system based on likelihood ratio computation from the statistical distribution of numerous parameters related to rhythm and based on syllable timing, syllable duration and amplitude (224 parameters are considered). They obtained significant results, and proved that mere prosodic cues can distinguish between some language pairs of the telephone speech OGI-MLTS corpus with results comparable to some non-prosodic systems (depending on the language pairs, correct discrimination rates range from chance to 93%). Cummins et al. (1999) have combined the delta-F0 curve and the first difference of the band-limited amplitude envelope with neural network models. The experiments were also conducted on the OGI-MLTS corpus, using pairwise language discrimination for which they obtained up to 70% of correct identification. The conclusions were that F0 was a more effective discriminant variable than the amplitude envelope modulation and that discrimination is better across prosodic family languages than in the same family.

Ramus and colleagues have proposed several studies (Ramus et al., 1999; Ramus and Mehler, 1999; Ramus, 2002a,b) based on the use of rhythm for language identification. This approach has been furthermore implemented in a semi-automatic modelling task (Dominey and Ramus, 2000). Their experiment aimed at assessing whether an artificial neural network may extract rhythm characteristics from sentences *manually labelled* in terms of consonants and vowels or not. Using the RMN "Ramus, Nespor, Mehler" corpus (1999), they reached significant discrimination results between languages belonging to different rhythm categories (78% for English/Japanese pair) and chance level for languages belonging to the same rhythm category. They concluded that those consonant/vowel sequences carry a significant part of the rhythmic

patterns of the languages and that they can be modelled. Interestingly, Galves et al. (2002) have reached similar results with no need for hand labelling: Using the RMN data, they automatically derived two criteria from a sonority factor. These two criteria (the mean value $\overline{S}$ and the mean value of the derivate δS of the sonority factor S) lead to a clustering of the languages closely related to the one obtained by Ramus and colleagues. Moreover, δS exhibits a linear correlation with ΔC and $\overline{S}$ is correlated to %V, tending to prove the consistency between the two approaches.

This quick overview of the rhythmic approaches to automatic language identification shows that several approaches, directly exploiting acoustic parameters without explicit unit modelling (e.g. Hidden Markov Model), may significantly discriminate some language pairs. Consequently, rhythm may be relevant for automatic discrimination or identification of the rhythm category of several languages. However, the fact that all these automatic systems exhibit results from "simple" pairwise discrimination emphasizes that using rhythm in a more complex identification task (with more than two languages) is not straightforward.

4.2. Rhythm unit modelling

The main purpose of this study is to provide an automatic segmentation of the signal into rhythmic units relevant for the identification of languages and to model their temporal properties in an efficient way. To this end, we use an algorithm formerly designed to model vowel systems in a language identification task (Pellegrino and André-Obrecht, 2000). The main features of this system are reviewed hereunder. This model does not pretend to integrate all the complex properties of linguistic rhythm and more specifically, hence it provides by no way a linguistic analysis of the prosodic systems of languages; the temporal properties observed and statistically modelled result from the interaction of several suprasegmental properties and an accurate analysis of this interaction is not yet possible.

Fig. 1 displays the synopsis of the system. A language-independent processing parses the signal into vowel and non-vowel segments. Parameters related to the temporal structure of the rhythm units are then computed and language-specific rhythmic models are estimated. During the test phase, the same processing is performed and the most likely language is determined following the Maximum Likelihood rule (see Section 5.2 for more details).

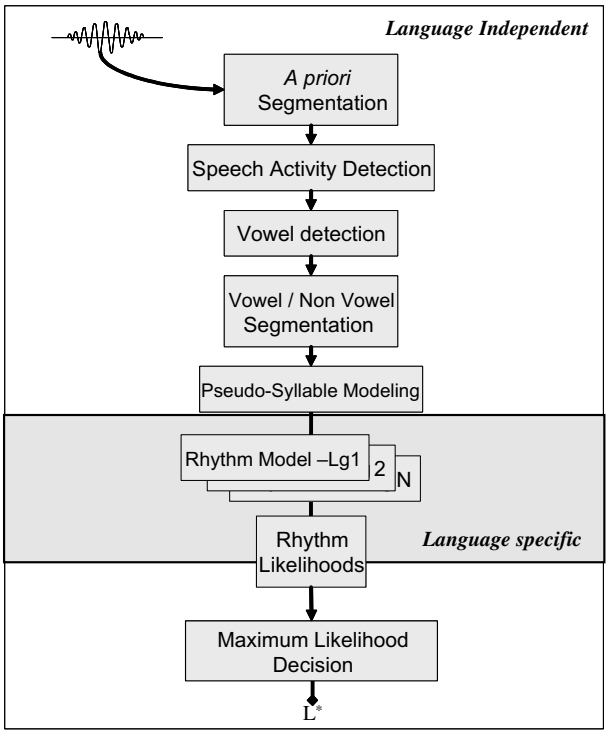


Fig. 1. Synopsis of the implemented system.

In order to extract features related to the potential consonant cluster (number and duration of consonants), a statistical segmentation based on the "Forward–Backward Divergence" algorithm is applied. Interested readers are referred to (André-Obrecht, 1988) for a comprehensive and detailed description of this algorithm. It identifies boundaries corresponding with abrupt changes in the wave spectrum resulting in two main categories of segments: short segments (bursts, but also transient parts of voiced sounds) and longer segments (steady parts of sounds).

A segmental speech activity detection (SAD) is performed to discard long pauses (not related to rhythm), and, finally, the vowel detection

algorithm locates sounds matching a vocalic structure via a spectral analysis of the signal. The SAD detects the less intense segment of the utterance (in term of energy) and the others segments are classified as Silence or Speech according to an adaptive threshold; Vowel detection is based on a dynamic spectral analysis of the signal in Mel frequency filters (both algorithms are detailed in (Pellegrino and André-Obrecht, 2000)). An example of the vowel/non-vowel parsing is provided in Fig. 2 (vertical lines).

The algorithm is applied in a language- and speaker-independent way without any manual adaptation phase. It is evaluated with the vowel error rate metric (VER) defined as follows:

$$\mathrm{VER} = 100 \cdot \left[\frac{N_{\mathrm{del}} + N_{\mathrm{ins}}}{N_{\mathrm{vow}}}\right]\% \qquad (1)$$

where N_{del} and N_{ins} are respectively the number of deleted vowels and inserted vowels, and N_{vow} is the actual number of vowels in the corpus.

Table 2 displays the performance of the algorithm for spontaneous speech, compared to other systems. The average value reached on five languages (22.9% of VER) is as good as the best systems optimized for a given language. The algorithm may be expected to perform better with read speech. However, no phonetically hand-labelled multilingual corpus of read speech was available to the authors to confirm this assumption.

The processing provides a segmentation of the speech signal in pause, non-vowel and vowel segments (see Fig. 2). Due to the intrinsic properties of the algorithm (and especially the fact that transient and steady parts of a phoneme may be separated), it is somewhat incorrect to consider that this segmentation is exactly a consonant/vowel segmentation since by nature, segments are shorter than phonemes. More specifically, vowel duration is on average underestimated since attacks and damping are often segmented as transient segments. Fig. 2 displays also examples of over-segmentation problems with consonants: the final /fnə/ sequence is segmented into eight segments

Table 2
Comparison of different algorithms of vowel detection

Reference	Corpus	Language	VER (%)
Pfitzinger et al. (1996)[a]	PhonDatII (read speech)	German	12.9
	Verbmobil (spontaneous speech)	German	21.0
Fakotakis et al. (1997)	TIMIT (read speech)	English	32.0
Pfau and Ruske (1998)	Verbmobil (spontaneous speech)	German	22.7
Howitt (2000)	TIMIT (read speech)	English	29.5
Pellegrino and André-Obrecht (2000)	OGI MLTS (spontaneous speech)	French	19.5
		Japanese	16.3
		Korean	28.5
		Spanish	19.2
		Vietnamese	31.1
		Average	22.9

The formula of the vowel error rate (VER) is given in the text of the paper.

[a] In this study, the error rate is estimated according to syllable nuclei and not explicitly vowels.

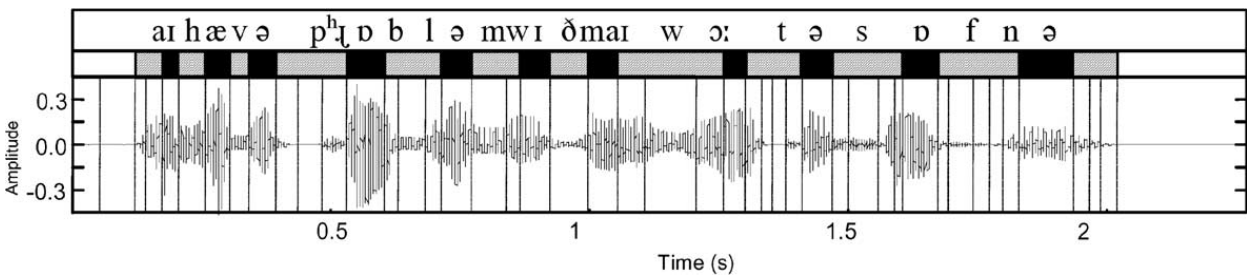


Fig. 2. Example of the automatic vowel/non-vowel labelling. The utterance is "I have a problem with my water softener…". The first tier gives the phonetic transcription. The second tier displays the result of the automatic algorithm (white = pause; dashed = non-vowel and black = vowel). Vertical lines displays the result of the a priori segmentation.

(four for the consonantal cluster, one for the vowel steady part and three for the final damping). However, our hypothesis is that this sequence is significantly correlated to the rhythmic structure of the speech sound; and the correlation already mentioned between actual syllabic rhythm and its estimation using vowel detection (Pellegrino et al., 2004) confirms this. Our assumption is that this correlation enables a statistical model to discriminate languages according to their rhythmic structure.

Even if the optimal rhythmic units may be language-specific (syllable, mora, etc.), the syllable may be considered as a good compromise. However, the segmentation of speech into syllables seems to be a language-specific mechanism even if universal rules related to sonority and if acoustic correlates of the syllable boundaries exist (see Content et al., 2000). Thus no language-independent algorithm can be derived at this moment, and even language-specific algorithms are uncommon (Kopecek, 1999; Shastri et al., 1999).

For these reasons, we introduce the notion of pseudo-syllables (PS) derived from the most frequent syllable structure in the world, namely the CV structure (Vallée et al., 2000). Using the vowel segments as milestones, the speech signal is parsed into patterns matching the structure: .C^{n}V. (with n an integer that may be zero).

For example, the parsing of the sentence displayed in Fig. 2 results in the following sequence of 11 pseudo-syllables:

(CCV.CV.CV.CCCV.CCCV.CCV.CV.CCCV.
CCCCV.CCCCV.CCCCCV)

roughly corresponding to the following phonetic segmentation:

(aɪ.hæ.və.pʰ.ɹɒ.blə.mwɪ.ðmaɪ.wɔː.tə.sɒ.fnə)

As said before, the segments labelled in the PS sequence are shorter than phonemes; consequently the length of the consonantal cluster is to a large extent biased to higher values than those given by a phonemic segmentation. We are aware of the limits of such a basic rhythmic parsing, but it provides an attempt to model rhythm that may be subsequently improved. However, it has the considerable advantage that neither hand-labelled data nor extensive knowledge of the language rhythmic structure is required.

A pseudo-syllable is described as a sequence of segments characterized by their duration and their binary category (consonant or vowel). This way, each pseudo-syllable is described by a variable length matrix. For example, a .CCV. pseudo-syllable will give:

$$P_{.\mathrm{CCV}.} = \begin{bmatrix} \mathrm{C} & \mathrm{C} & \mathrm{V} \\ D_{\mathrm{C1}} & D_{\mathrm{C2}} & D_{\mathrm{V1}} \end{bmatrix} \quad (2)$$

where C and V are binary labels and D_X is the duration of the segment X.

This variable length description is the most accurate, but it is not appropriate for Gaussian Mixture Modelling (GMM). For this reason, another description resulting in a constant length description for each pseudo-syllable has been derived. For each pseudo-syllable, three parameters are computed, corresponding respectively with total consonant cluster duration, total vowel duration and complexity of the consonantal cluster. With the same .CCV. example, the description becomes:

$$P'_{.\mathrm{CCV}.} = \{(D_{\mathrm{C1}} + D_{\mathrm{C2}})\ D_{\mathrm{V}}\ N_{\mathrm{C}}\} \quad (3)$$

where N_{C} is the number of segments in the consonantal cluster (here, $N_{\mathrm{C}} = 2$).

Even if this description is clearly not optimal since the individual information on the consonant segments is lost, it takes a part of the complexity of the consonant cluster into account.

5. Language identification task

5.1. Corpus description and statistics

Experiments are performed on the MULTEXT multilingual corpus (Campione and Véronis, 1998), extended with Japanese (Kitazawa, 2002) and Mandarin (Komatsu et al., 2004). This database thus contains recordings of seven languages (French, English, Italian, German, Japanese, Mandarin and Spanish), pronounced by 70 different speakers (five male and five female per language).

The MULTEXT data consist of read passages that may be pronounced by several speakers.

Table 3
The MULTEXT Corpus (from Campione and Véronis, 1998)

Language	Passages per speaker	Total duration (min)	Average duration per passage (s)	Training (min)	Test (min)
English	15	44	17.6	24	6
French	10	36	21.9	29	7
German	20	73	21.9	29	7
Italian	15	54	21.7	30	7
Mandarin	15	58	20.0	26	11
Japanese	40	124	31	39	6
Spanish	15	52	20.9	27	8

Table 4
Estimation of D_C as a predictor of N_C

Language	R^2	Equation	N_B PS
EN	0.83	$100\ \widehat{N}_C = 3.68D_C + 22$	11 741
FR	0.78	$100\ \widehat{N}_C = 3.25D_C + 15$	9307
GE	0.82	$100\ \widehat{N}_C = 3.43D_C + 56$	19 296
IT	0.81	$100\ \widehat{N}_C = 3.27D_C + 34$	14 867
JA	0.80	$100\ \widehat{N}_C = 3.27D_C + 56$	28 913
MA	0.79	$100\ \widehat{N}_C = 2.95D_C + 86$	14 583
SP	0.80	$100\ \widehat{N}_C = 3.76D_C + 20$	15 005

Results of a linear regression in least-squares sense. N_B PS is the number of pseudo-syllables from which the regression was performed for each language. R^2 is the squared correlation coefficient (according to Spearman rank order estimation). All correlations are highly significant ($p < 0001$).

Despite the relative small amount of data and to avoid possible text dependency, the following experiments are performed with two subsets of the corpus defining no-overlapping training and test sets in terms of speakers and texts (see Table 3). The training corpus is supposed to be representative of each language syllabic inventory. For instance, the mean duration of each passage for the French data is 98 syllables (±20 syllables) and the overall number of syllable tokens in the French corpus is about 11 700.[8] Even if the syllable inventory is not exhaustive in this corpus, it is reasonable to assume that a statistical model derived from these data will be statistically representative of most of the syllable diversity for each language.

In the classical rhythm typology, French, Italian and Spanish are known as syllable-timed languages while English, German and Mandarin are stress-timed. Japanese is the only mora-timed language of the corpus. Whether this typology is correct or results from an artefact of a rhythmic continuum, our approach should be able to capture features linked to the rhythm structure of these languages.

Intuitively, the duration of consonantal clusters is supposed to be correlated to the number of segments constituting the cluster. Table 4 gives the results of a linear regression with D_C (in seconds) as a predictor of N_C. For each language, a significant positive correlation is achieved and R^2 values range from 0.78 for French to 0.83 for English (see Fig. 3 for the scatter plot of English data). In term of slope, values range from 2.95 for Mandarin to 3.76 for Spanish meaning that the relation between N_C and D_C is to some extend language dependent. For this reason, both parameters have been taken into account in the following experiments.

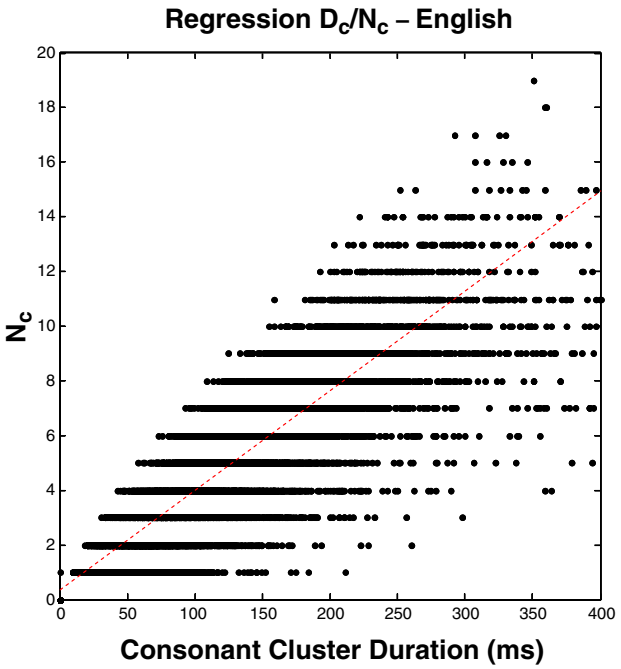


Fig. 3. Evaluation of D_C as a predictor of N_C for English. Dots are measured values and the solid line is the best linear fit estimated in the least-squares sense.

[8] This number takes the number of repetitions of each passage into account. Considering each passage once, the number of syllables is 3900.

In order to test hypotheses on language specific differences in the distribution of the parameters, a Jarque-Bera test of normality was performed. It confirms that the distributions are clearly non normal ($p < .0001$; $j > 10^3$ for D_C, D_V and N_C, for all languages). Consequently, a non-parametric Kruskal–Wallis test was performed for each parameter to evaluate the differences among the languages. They reveal a highly significant global effect of the language for D_V ($p < .0001$; df = 6; chi-square = 2248), D_C ($p < .0001$; df = 6; chi-square = 1061) and N_C ($p < .0001$; df = 6; chi-square = 2839). The results of the Kruskal–Wallis test have then been used in a multiple comparison procedure using Tukey criterion of significant difference.

Table 5–7 gives the results of the pairwise comparison. In order to make the interpretation easier, a graphical representation is drawn from the values (Fig. 4). Regarding consonant duration, a cluster grouping the stress-timed languages is clearly identified. This cluster is coherent with the complex onsets and coda present in these languages, either in number of phonemes (English and German) or intrinsic complexity of the consonants (aspirated, retroflex, etc. for Mandarin) The other languages spread along the D_C dimension and Japanese and Italian are intermediate between the most prototypical syllable-timed languages (Spanish and French) and the stress-timed languages cluster.

The situation revealed by D_V is quite different: English, Japanese, German and Italian cluster together (though significant differences exist between Italian on one side, and English, Japanese and German on the other side) while Mandarin and French are distant. Spanish is also individualized at this opposite extreme of this dimension. N_C distributions exhibit important diversity among languages since English and Mandarin are the only cluster for which no significant difference is observed.

Table 5
Significancy of the differences among the distributions of D_C (multiple comparisons from the Kruskal–Wallis analysis)

	EN	FR	GE	IT	JA	MA	SP
EN		*	*	*	*	n.s.	*
FR			*	*	*	*	*
GE				*	*	n.s.	*
IT					n.s.	*	*
JA						*	*
MA							*

n.s. is not significant and * is significant or highly significant.

Table 6
Significancy of the differences among the distributions of D_V (multiple comparisons from the Kruskal–Wallis analysis)

	EN	FR	GE	IT	JA	MA	SP
EN		*	n.s.	*	n.s.	*	*
FR			*	*	*	n.s.	*
GE				*	n.s.	*	*
IT					*	*	*
JA						*	*
MA							n.s.

n.s. is not significant and * is significant or highly significant.

Table 7
Significancy of the differences among the distributions of N_C (multiple comparisons from the Kruskal–Wallis analysis)

	EN	FR	GE	IT	JA	MA	SP
EN		*	*	*	*	n.s.	*
FR			*	*	*	*	*
GE				*	*	*	*
IT					*	*	*
JA						*	*
MA							*

n.s. is not significant and * is significant or highly significant.

5.2. GMM modelling for identification

GMM (Gaussian Mixture Models) are used to model the pseudo-syllables which are represented in the three-dimensional space described in the previous section. They are estimated using the EM (Expectation–Maximization) algorithm initialized with the LBG algorithm (Reynolds, 1995; Linde et al., 1980).

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the training set and $\Pi = \{(\alpha_i, \mu_i, \Sigma_i),\ 1 \leqslant i \leqslant Q\}$ the parameter set that defines a mixture of Q p-dimensional Gaussian pdfs. The model that maximizes the overall likelihood of the data is given by:

$$\Pi^* = \arg\max_{\Pi} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \times \exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] \right\} \quad (4)$$

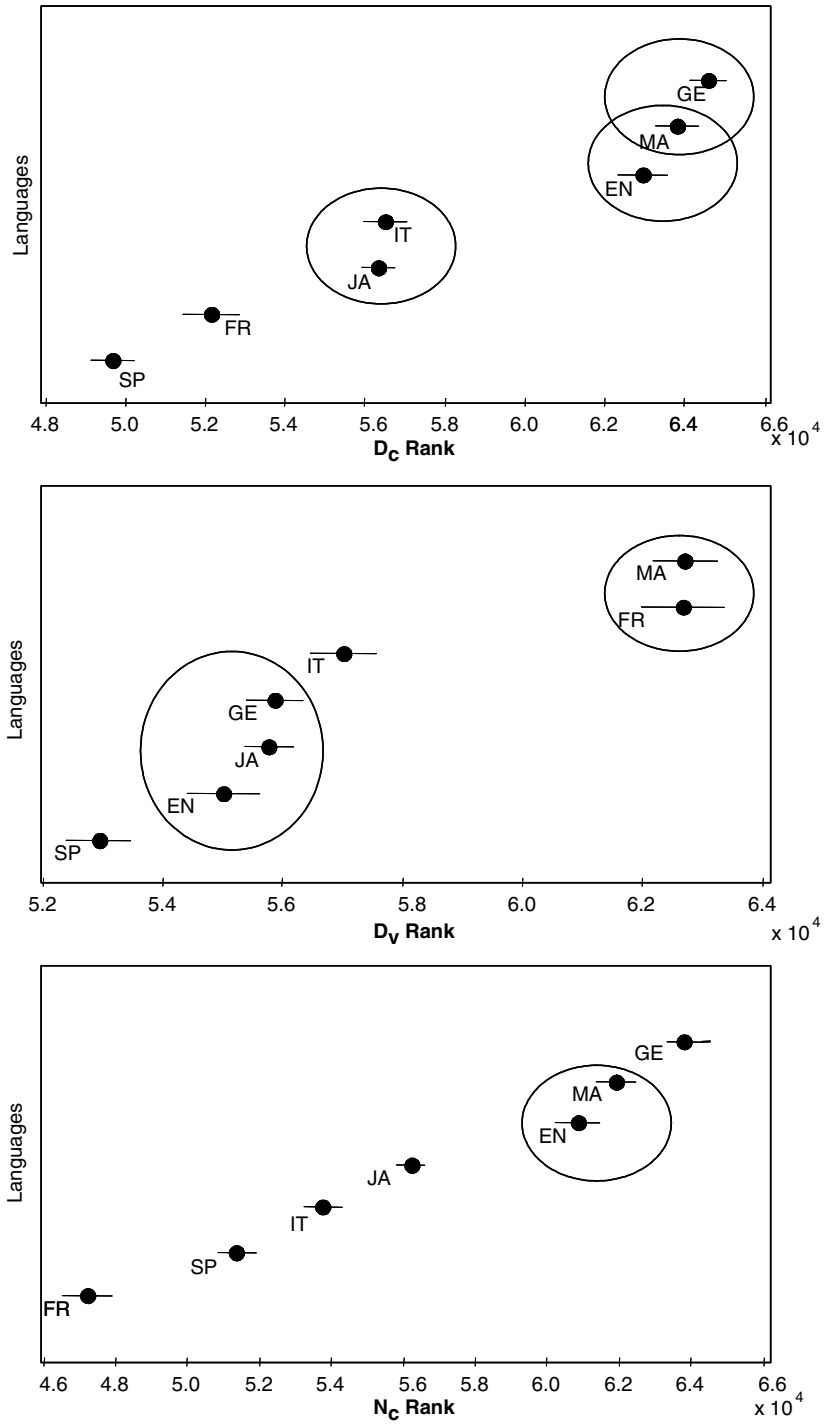


Fig. 4. Estimated rank for each language for the D_C distribution above, the D_V distribution (middle) and the N_C distribution below. Lines spanning across the dots give the 95% confidence interval. Ellipses cluster languages for which the multiple comparisons show no significant differences.

where α_k is the mixing weight of the kth Gaussian term.

The maximum likelihood parameters Π^* are obtained using the EM algorithm. This algorithm

presupposes that the number of components Q and the initial values are given for each Gaussian pdf. Since these values greatly affect the performances of the EM algorithm, a vector quantization (VQ) is applied to the training corpus to optimize them.

The LBG algorithm (Linde et al., 1980) is applied to provide roots for the EM algorithm; it performs an iterated clustering of the learning data into codewords optimized according to the nearest neighbor rule. The splitting procedure may be stopped either when the variation of the data distortion drops under a given threshold or when a given number of codewords is reached (this option is used here).

During the identification phase, all the PS detected in the test utterance are gathered and parameterized. The likelihood of this set of segments $Y = \{y_1, y_2, \ldots, y_N\}$ according to each model (denoted L_i) is given by:

$$\Pr(Y|L_i) = \sum_{j=1}^{N} \Pr(y_j|L_i) \tag{5}$$

where $\Pr(y_j|L_i)$ denotes the likelihood of each segment that is given by:

$$\Pr(y_j|L_i) = \sum_{k=1}^{Q_i} \frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \times \exp\left[-\frac{1}{2}(y_j - \mu_k^i)^T \Sigma_k^{-1}(y_j - \mu_k^i)\right] \tag{6}$$

Furthermore, hypothesizing under the *winner takes all* (WTA) assumption (Nowlan, 1991), the expression (7) is then approximated by:

$$\Pr(y_j|L_i) = \max_{1\leqslant k\leqslant Q_i} \left[\frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \times \exp\left[-\frac{1}{2}(y_j - \mu_k^i)^T \Sigma_k^{-1}(y_j - \mu_k^i)\right]\right] \tag{7}$$

5.3. Automatic identification results

Pseudo-syllable segmentation has been conceived to be related to language rhythm. In order to assess whether this is actually verified or not, a first experiment aiming at discriminating between the three rhythmic classes is performed; a language identification experiment with the seven languages is then achieved. At last, a standard acoustic approach is implemented and tested with the same task to provide a comparison.

The first experiment aims at identifying to which rhythmic group belongs the language spoken by an unknown speaker of the MULTEXT corpus. The stress-timed language group gather English, German and Mandarin. French, Italian and Spanish define the syllable-timed language group. The mora-timed language group consists only of Japanese. The number of Gaussian components is fixed to 16 using the training set as a development set to optimize the number of Gaussian components of the GMM. The overall results are presented in Table 8 in a confusion matrix. 119 from 139 files of the test set are correctly identified. The mean identification rate is 86 ± 6% of correct identification (chance level is 33%) and scores range from 80% for syllable- and mora-timed languages to 92% for stress-timed languages. These first results show that the PS approach is able to model temporal features that are relevant for rhythmic group identification.

The second experiment aims at identifying which of the seven languages is spoken by an unknown speaker of the MULTEXT corpus. The number of Gaussian components is fixed to 8, using the training set as a development set to optimize the number of Gaussian components of the GMM. The overall results are presented in Table 9 in a confusion matrix. 93 from the 139 files of the test set are correctly identified. The mean identification score thus reaches 67 ± 8% of correct

Table 8
Results for the rhythmic group identification task (16 Gaussian components per GMM)

Rhythmic group	Model		
	Stress-timed	Syllable-timed	Mora-timed
Stress-timed	**55**	5	–
Syllable-timed	10	**48**	1
Mora-timed	2	2	**16**

Overall score is 86 ± 6% (119/139 files).

Table 9
Results for the seven language identification task (eight Gaussian components per GMM)

Language	Model						
	EN	GE	MA	FR	IT	SP	JA
English	**16**	1	1	–	1	1	–
German	5	**14**	1	–	–	–	–
Mandarin	4	3	**11**	–	1	–	1
French	–	–	–	**19**	–	–	–
Italian	6	1	1	–	**11**	–	1
Spanish	–	–	–	8	2	**6**	4
Japanese	2	–	–	–	2	–	**16**

Overall score is 67 ± 8% (93/139 files).

identification (chance level is 14%). Since the test corpus is very limited, the confidence interval is pretty wide.

Scores broadly vary and range from 30% for Spanish to 100% for French. Actually, Spanish is massively confused with French; Italian is also fairly misclassified (55% of correct decision) and especially with English. Bad classification is also observed for Mandarin which is confused with both German and English (55% of correct identification). It thus tends to confirm that the classification of Mandarin as a stress-timed language is consistent with the acoustic measurements performed here and for which the Mandarin PS distributions are not significantly different from either German or English distributions.

The wide range of variation observed for the scores may be partially explained studying the speaking rate variability. As for rhythm, speaking or speaker rate is difficult to define but it may be evaluated in term of syllable or phoneme per second. Counting the number of vowels detected per second may provide a first approximation of the speaking rate (see Pellegrino et al., 2004, for a discussion about the speaking rate measurement). Table 10 displays for each language of the database the mean and standard deviation of the number of vowels detected per second among the speakers of the database.

This rate ranges from 5.05 for Mandarin to 6.94 for Spanish and these variations may be due to both socio-linguistic factors and rhythmic factors related to the structure of the syllable in those languages. Spanish and Italian exhibit the greatest standard deviations (resp. 0.59 and 0.64) of their rate. It means that their models are probably less robust than the others since the parameter distributions are wider. On the opposite, French dispersion is the smallest (0.33) and consistently has the better language identification rate. This hypothesis is supported by a correlation test (Spearman rank order estimation) between the language identification score and speaking rate standard deviation ($\rho = -0.77$, $p = 0.05$). This shortcoming points out that, at this moment, no normalization is performed on the D_C and D_V durations. This limitation prevents our model from being adapted to spontaneous speech and this major bottleneck must be tackled in a near future.

At last, the same data and task have been used with an acoustic GMM classifier in order to compare the results of the purely rhythmic approach proposed in this paper with those obtained with a standard approach. The parameters are computed on each segment issued from the automatic segmentation (Section 4). The features consist of 8 Mel Frequency Cepstral Coefficients, their derivatives, and energy, computed on each segment. The number of Gaussian components is fixed to 16 using the training set as a development set to optimize the number of Gaussian components of the GMM. Increasing the number of components does not result in better performances; this may be due to the limited size of the training set both in terms of duration and number of speakers (only eight speakers per language, except for Japanese: four speakers). The overall results are presented in Table 11 in a confusion matrix. 122 from 139

Table 10
Speaking rate approximated by the number of vowels detected per second for the seven languages

	English	French	German	Italian	Japanese	Mandarin	Spanish
Mean	5.39	6.37	5.06	5.71	5.29	5.05	6.94
Std. deviation	0.52	0.33	0.45	0.64	0.51	0.52	0.59

Table 11
Results for the seven language identification task (standard acoustic approach, 16 Gaussian components per GMM)

Language	Model						
	EN	GE	MA	FR	IT	SP	JA
English	**15**	–	–	–	5	–	–
German	–	**20**	–	–	–	–	–
Mandarin	–	–	**20**	–	–	–	–
French	–	–	–	**17**	–	2	–
Italian	2	–	2	–	**13**	1	2
Spanish	1	–	–	2	–	**17**	–
Japanese	–	–	–	–	–	–	**20**

Overall score is 88 ± 5% (122/139 files).

files of the test set are correctly identified. The mean identification rate is 88 ± 5% of correct identification.

German, Mandarin and Japanese are perfectly identified. The worst results are reached for Italian (65%). Noteworthy is that Mandarin is well discriminated from English and German, contrary to what was observed with rhythmic models. This suggests that the two approaches may be efficiently combined to improve the performances. However, the fact that the acoustic approach reaches significantly better results than the rhythmic approach implies that further improvement are necessary before designing an efficient merging architecture.

6. Conclusion and perspectives

While most of the systems developed nowadays for language identification purposes are based on phonetic and/or phonotactic features, we believe that using other kinds of information may be complementary and widen the field of interest of these systems, for example by tackling linguistic typological or cognitive issues about language processing. We propose one of the first approaches dedicated to language identification based on *rhythm* modelling that is tested on a task more complex than pairwise discrimination. Our system makes use of an automatic segmentation into vowel and non-vowel segments leading to a parsing of the speech signal into pseudo-syllabic patterns. Statistical tests performed on the language-specific distributions of the pseudo-syllable parameters show that significant differences exist among the seven languages of this study (English, French, German, Italian, Japanese, Mandarin and Spanish). A first assessment of the validity of this approach is given by the results of a rhythmic class identification task: The system reaches 86 ± 6% of correct discrimination when three statistical models are trained with data from stress-timed languages (English, German and Mandarin), from syllable-timed languages (French, Italian and Spanish) and from Japanese (the only mora-timed language of this study). This experiment shows that the traditional stress-timed vs. syllable-timed vs. mora-timed opposition is assessed with the seven languages we have tested, or more precisely, that the three language groups (English + German + Mandarin vs. French + Italian + Spanish vs. Japanese) exhibit significant differences according to the temporal parameters we propose.

A second experiment done with the seven language identification task produces relatively good results (67 ± 8% correct identification rate for 21-s utterances). Once again, confusions occur more frequently *within* rhythmic classes than *across* rhythmic classes. Among the seven languages, three are identified with high scores (more than 80%) and can be qualified as "prototypical" from the rhythmic groups (English for stress-timing, French for syllable-timing and Japanese for mora-timing). It is thus interesting to point out that the pseudo-syllable modelling may also manage to identify languages that belong to the same rhythmic family (e.g. French and Italian are not confused), showing that the temporal structure of the pseudo-syllables is quite language-specific. To summarize, even if the pseudo-syllable segmentation is rough and not able to take the language-specific syllable structures into consideration, it captures at least a part of the rhythmic structure of each language.

However, rhythm cannot be reduced to a raw temporal sequence of consonants and vowels, and, as pointed out by Zellner-Keller (2002) its multilayer nature should be taken into account to correctly characterize languages. Among many parameters, those linked to tones or to the stress phenomenon may be pretty salient. For instance, Mandarin, which is fairly confused with other languages in the present study may be well recognized

with other suprasegmental features due to its tonal system. Consequently, taking energy or pitch features into account may lead to significant improvement in the language identification performance. However, these physical characteristics lay at the interface between segmental and supra-segmental levels and their values and variations thus result from a complex interaction, increasingly complicating their correct handling.

Besides, the algorithm of pseudo-syllable segmentation may also be enhanced. An additional distinction between voiced and voiceless consonants may be performed to add another rhythmic parameter, and moreover, more complex pseudo-syllables including codas (hence with a C^mVC^n structure) may be obtained by applying segmentation rules based on sonority (see Galves et al., 2002 for a related approach).

Last, the major future challenge will be to tackle the speaking rate variability (shown in Section 5 to be correlated to the identification performance) and to propose an efficient normalizing or modelling that will allow us to adapt this approach to spontaneous speech corpora and to a larger set of languages. Very preliminary experiments performed on the OGI MLTS corpus are reported in (Rouas et al., 2003).

Acknowledgments

The authors would like especially to thank Brigitte Zellner-Keller for her helpful comments and advices and Emmanuel Ferragne for his careful proofreading of the draft of this paper. The authors are very grateful to the reviewers for their constructive suggestions and comments.

This research has been supported by the EMERGENCE program of the Région Rhône-Alpes (2001–2003) and the French Ministère de la Recherche (program ACI "Jeunes Chercheurs"—2001–2004).

References

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.

Adami, A.G., Hermansky, H., 2003. Segmentation of speech for speaker and language recognition. In: Proc. Eurospeech, Geneva, pp. 841–844.

André-Obrecht, R., 1988. A new statistical approach for automatic speech segmentation. IEEE Trans. Acoust. Speech Signal Process. 36 (1).

Antoine, F., Zhu, D., Boula de Mareüil, P., Adda-Decker, M., 2004. Approches Segmentales multilingues pour 1'identification automatique de la langue: phones et syllabes. In: Proc. Journées d'Etude de la Parole, Fes, Morocco.

Barkat-Defradas, M., Vasilescu, I., Pellegrino, F., 2003. Stratégies perceptuelles et identification automatique des langues. Revue PArole, 25/26, 1–37.

Berg, T., 1992. Productive and perceptual constraints on speech error correction. Psychol. Res. 54, 114–126.

Bernd Möbius, 1998. Word and syllable models for German text-to-speech synthesis. In: Proceedings of the Third International Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 59–64.

Besson, M., Schön, D., 2001. Comparison between language and music. In: Zatorre, R., Peretz, I. (Eds.), "The biological foundations of music". Annals of The New York Academy of Sciences, Vol. 930.

Bond, Z.S., Stockmal, V., 2002. Distinguishing samples of spoken Korean from rhythmic and regional competitors. Lang. Sci. 24, 175–185.

Boysson-Bardies, B., Vihman, M.M., Roug-Hellichius, L., Durand, C., Landberg, I., Arao, F., 1992. Material evidence of infant selection from the target language: A cross-linguistic study. In: Ferguson, C., Menn, L., Stoel-Gammon, C. (Eds.), Phonological Development: Models, Research, Implications. York Press, Timonium, MD.

Campione, E., Véronis, J., 1998. A multilingual prosodic database. In: Proc. ICSLP'98, Sydney, Australia.

Content, A., Dumay, N., Frauenfelder, U.H., 2000. The role of syllable structure in lexical segmentation in French. In: Proc. Workshop on Spoken Word Access Processes. Nijmegen, The Netherlands.

Content, A., Kearns, R.K., Frauenfelder, U.H., 2001. Boundaries versus onsets in syllabic segmentation. J. Memory Lang. 45 (2).

Crystal, D., 1990. A Dictionary of Linguistics and Phonetics, third ed. Blackwell, London.

Cummins, F., Gers, F., Schmidhuber, J., 1999. Language identification from prosody without explicit features. In: Proc. EUROSPEECH'99.

Cutler, A., 1996. Prosody and the word boundary problem. In: Morgan, Demuth (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition. Lawrence Erlbaum Associates, Mahwah, NJ.

Cutler, A., Norris, D., 1988. The role of strong syllables in segmentation for lexical access. J. Exp. Psychol.: Human Perception Perform., 14.

Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. J. Phonet. 11.

Delattre, P., Olsen, C., 1969. Syllabic features and phonic impression in english, german, french and spanish. Lingua 22, 160–175.

Dominey, P.F., Ramus, F., 2000. Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure in the infant. Lang. Cognitive Process. 15 (1).

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulation on speech reception. JASA 95 (5).

Fakotakis, N., Georgila, K., Tsopanoglou, A., 1997. Continuous HMM text-independent speaker recognition system based on vowel spotting. In: 5th European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece, September 1997, vol. 5, pp. 2247–2250.

Ferragne, E., Pellegrino, F. Rhythm in read British english: Interdialect variability. In: Proc. INTERSPEECH/ICSLP 2004, Jeju, Korea, October 2004.

Fromkin, V. (Ed.), 1973. Speech Errors as Linguistic Evidence. Mouton Publishers, The Hague.

Fujimura, O., 1975. Syllable as a unit of speech recognition. IEEE Trans. on ASSP ASSP-23 (1), 82–87, 02/1975.

Galves, A., Garcia, J., Duarte, D., Galves, C., 2002. Sonority as a basis for rhythmic class discrimination. In: Proc. Speech Prosody 2002 Conference, 11–13 April.

Ganapathiraju, A., 1999. The webpage of the Syllable Based Speech Recognition Group, Available from: <http://www.clsp.jhu.edu/ws97/syllable/>, last visited July 2002.

Gauvain, J.-L., Messaoudi, A., Schwenk, H., 2004. Language recognition using phone lattices. In: Proc. International Conference on Spoken Language Processing, Jeju island, Korea.

Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis, Papers in Laboratory Phonology 7, Mouton.

Greenberg, S., 1996. Understanding speech understanding—towards a unified theory of speech perception. In: Proc. ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Keele, England.

Greenberg, S., 1997. On the origins of speech intelligibility in the real world. In: Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-á-Mousson, France.

Greenberg, S., 1998. Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. In: Proc. ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition, Kekrade, The Netherlands.

Greenberg, S., Carvey, H.M., Hitchcock, L., 2002. The relation of stress accent to pronunciation variation in spontaneous American English discourse. In: Proc. 2001 ISCA Workshop Prosody and Speech Processing, Red Bank, NJ, USA, 2002, pp. 53–56.

Hamdi, R., Barkat-Defradas, M., Ferragne, E., Pellegrino, F. Speech timing and rhythmic structure in Arabic dialects: A comparison of two approaches. In: Proc. INTERSPEECH/ICSLP 2004, Jeju, Korea, 2004.

Howitt, A.W., 2000. Vowel landmark detection. In: 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Jestead, W., Bacon, S.P., Lehman, J.R., 1982. Forward masking as a function of frequency, masker level and signal delay. JASA 74 (4).

Keller, E., Zellner, B., 1997. Output requirements for a high-quality speech synthesis system: The case of disambiguation. In: Proc. MIDDIM-96, 12–14 August 96, pp. 300–308.

Kern, S., Davis, B.L., Koçbas D., Kuntay A., Zink I. Crosslinguistic "universals" and differences in babbling. In: OMLL—Evolution of language and languages, European Science Foundation, in press.

Kitazawa, S., 2002. Periodicity of Japanese accent in continuous speech. In: Speech Prosody, Aix en Provence, France, April 2002.

Komatsu, M., Arai, T., Sugawara, T., 2004. Perceptual discrimination of prosodic types. In: Proc. Speech Prosody, Nara, Japan, 2004, pp. 725–728.

Kopecek, I., 1999. Speech recognition and syllable segments. In: Proc. Workshop on Text, Speech and Dialogue—TSD'99Lectures Notes in Artificial Intelligence 1692. Springer-Verlag.

Ladefoged, P., 1975. A Course in Phonetics. Harcourt Brace Jovanovich, New York, p. 296.

Levelt, W., Wheeldon, L., 1994. Do speakers have access to a mental syllabary. Cognition, 50.

Li, K.P., 1994. Automatic language identification using syllabic spectral features. In: Proc. IEEE ICASSP'94, Adelaide, Australia.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition, 21.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer. IEEE Trans. Comm. 28 (January).

MacNeilage, P., 1998. The frame/content theory of evolution of speech production. Brain Behavior. Sci. 21, 499–546.

MacNeilage, P.P., Davis, B.L., 2000. Evolution of speech: The relation between ontogeny and phytogeny. In: Hurford, J.R., Knight, C., Studdert-Kennedy, M.G. (Eds.), The Evolutionary Emergence of Language. Cambridge University Press, Cambridge, pp. 146–160.

MacNeilage, P.P., Davis, B.L., Kinney, A., Maryear, C.L., 2000. The motor core of speech: A comparison of serial organization patterns in infants and languages. Child Develop. 71, 153–163.

Martin, A.F., Przybocki., M.A., 2003. NIST 2003 language recognition evaluation. In: Proc. Eurospeech, Geneva, pp. 1341–1344.

Massaro, D.W., 1972. Preperceptual images, processing time and perceptual units in auditory perception. Psychol. Rev. 79 (2).

Mehler, J., Dommergues, J.Y., Frauenfelder, U., Segui, J., 1981. The syllable's role in speech segmentation. J. Verbal Learning Verbal Behavior, 20.

Mehler, J., Dupoux, E., Nazzi, T., Dehaene-Lambertz, G., 1996. Coping with linguistic diversity: The infant's viewpoint. In: Morgan, Demuth (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition. Lawrence Erlbaum Associates, Mahwah, NJ.

Mirghafori, N., Fosler, E., Morgan, N., 1995. Fast speakers in large vocabulary continuous speech recognition: Analysis & antidotes. In: Proc. Eurospeech'95, Madrid, Spain.

Muthusamy, Y.K., Jain, N., Cole, R.A., 1994. Perceptual benchmarks for automatic language identification. In: Proc. IEEE ICASSP'94, Adelaide, Australia.

Nagarajan, T., Murthy, H.A., 2004. Language identification using parallel syllable-like unit recognition. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Canada, pp. 401–404.

Nazzi, T., Ramus, F., 2003. Perception and acquisition of linguistic rhythm by infants. Speech Comm. 41 (1–2), 233–243.

Nowlan, S., 1991. Soft Competitive Adaptation: Neural Network Learning Algorithm based on fitting Statistical Mixtures, PhD Thesis, School of Computer Science, Carnegie Mellon Univ.

Ohala, J.J., Gilbert, B., 1979. On listeners' ability to identify languages by their prosody. In: Leon & Rossi (Eds.), Problémes de prosodie, Vol. 2, Hurtubise HMH.

O'Shaugnessy, D., 1987. Speech Communication. Human and Machine. Addison Wesley, Reading, MA, USA.

Pellegrino, F., André-Obrecht, R., 2000. Automatic language identification: An alternative approach to phonetic modelling. Signal Process. 80 (7), 1231–1244.

Pellegrino, F., Farinas, J., Rouas, J-.L., 2004. Autom atic estimation of speaking rate in multilingual spontaneous speech. In: Proc. Speech Prosody 2004, Nara, Japan, March 2004.

Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. In: Proc. IEEE ICASSP'98, Seattle, WA, USA.

Pfitzinger, H., Burger, S., Heid, S., 1996. Syllable detection in read and spontaneous speech. In: 4th International Conference on Spoken Language Processing, Philadelphia, vol. 2, pp. 1261–1264.

Ramus, F., 2002a. Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. Ann. Rev. Lang. Acquis. 2, 85–115.

Ramus, F., 2002b. Acoustic correlates of linguistic rhythm: Perspectives. In: Proc. Speech Prosody 2002, Aix-en-Provence, France.

Ramus, F., Mehler, J., 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. J. Acoust. Soc. Amer. 105 (1).

Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73 (3).

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17 (1–2), 91–108.

Rouas, J-.L., Farinas, J., Pellegrino, F., Regine André-Obrecht, 2003. Modeling prosody for language identification on read and spontaneous speech. In: Proc. ICASSP'2003, Hong Kong, China, pp. 40–43.

Rouas, J-.L., Farinas, J., Pellegrino, F., Regine André-Obrecht, 2004. Evaluation automatique du débit de la parole sur des données multilingues spontanées. In: actes des XXVémes JEP, Fés, Maroc, April 2004.

Shastri, L., Chang, S., Greenberg, S., 1999. Syllable detection and segmentation using temporal flow neural networks. In: Proc. ICPhS'99, San Francisco, CA, USA.

Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A., 2003. Acoustic, phonetic, and discriminative approaches to automatic language identification. In: Proc. Eurospeech, Geneva, pp. 1345–1348.

Stockmal, V., Muljani, D., Bond, Z.S., 1996. Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling. In: Proc. ICSLP, Philadelphia, pp. 1748–1751.

Stockmal, V., Moates, D., Bond, Z.S., 2000. Same talker, different language. Appl. Psycholinguistics 21, 383–393.

Taylor, P.A., King, S., Isard, S.D., Wright, H., Kowtko, J., 1997. Using intonation to constrain language models in speech recognition. In: Proc. Eurospeech 97, Rhodes, Greece.

Thymé-Gobbel, A., Hutchins, S.E., 1999. Prosodic features in automatic language identification reflect language typology. In: Proc. ICPhS'99, San Francisco, CA, USA.

Todd, N.P., Brown, G.J., 1994. A computational model of prosody perception. In: Proc. ICSLP'94, Yokohama, Japan.

Vallée, N., Boë, L.J., Maddieson, I., Rousset, L., 2000. Des lexiques aux syllabes des langues du monde—Typologies et structures. In: Proc. JEP 2000, Aussois, France.

Vasilescu, I., Pellegrino, F., Hombert, J., 2000. Perceptual features for the identification of romance languages. In: Proc. ICSLP'2000, Beijing.

Verhasselt, J.P., Martens, J-.P., 1996. A fast and reliable rate of speech detector. In: Proc. ISCLP'96, Philadelphia, PA, USA.

Weissenborn, J., Höhle, B. (Eds.), 2001. Approaches to Bootstrapping. Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition, Vol. 1, Acquisition and Language Disorders 23. John Benjamins Publishing Company, p. 299.

Wu, S-.L., 1998. Incorporating information from syllable-length time scales into automatic speech recognition, Report TR-98-014 of the International Computer Science Institute, Berkeley, CA, USA.

Zellner Keller, B., 2002. Revisiting the status of speech rhythm. In: Bernard Bel, Isabelle Marlien (Eds.), Proc. Speech Prosody 2002 Conf., 11–13 April 2002, pp. 727–730.

Zellner Keller, B., Keller, E., 2001. Representing speech rhythm. In: Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (Eds.), Improvements in Speech Synthesis. John Wiley, Chichester.

Zissman, M.A., Berkling, K.M., 2001. Automatic language identification. Speech Comm. 35 (1–2), 115–124.

A.2 Article dans le journal JSHR

Article intitulé « Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss » paru le 18 septembre 2017 dans le Journal of Speech Language and Hearing Research [Fontan et al., 2017].

Cette communication détaille le corpus qui a été produit par le projet régional AGILE IT (cf. section 5.4.9 page 79) et les travaux qui en ont découlé. Cette collaboration entre le laboratoire CLLE de l'université Jean Jaures, la société Archean et l'équipe SAMOVA de l'université Paul Sabatier a permis la mise en place d'une mesure automatique de l'intellgibilité, à partir de tests perceptifs humains.

Ces résultats de recherche ont pour but de faciliter le réglage de prothèses auditives afin d'en faciliter l'adoption par les personnes atteintes de presbyacousie.

JSLHR

Research Article

Automatic Speech Recognition Predicts Speech Intelligibility and Comprehension for Listeners With Simulated Age-Related Hearing Loss

Lionel Fontan,[a,b] Isabelle Ferrané,[b] Jérôme Farinas,[b] Julien Pinquier,[b] Julien Tardieu,[c] Cynthia Magnen,[c] Pascal Gaillard,[d] Xavier Aumont,[a] and Christian Füllgrabe[e]

Purpose: The purpose of this article is to assess speech processing for listeners with simulated age-related hearing loss (ARHL) and to investigate whether the observed performance can be replicated using an automatic speech recognition (ASR) system. The long-term goal of this research is to develop a system that will assist audiologists/hearing-aid dispensers in the fine-tuning of hearing aids.
Method: Sixty young participants with normal hearing listened to speech materials mimicking the perceptual consequences of ARHL at different levels of severity. Two intelligibility tests (repetition of words and sentences) and 1 comprehension test (responding to oral commands by moving virtual objects) were administered. Several language models were developed and used by the ASR system in order to fit human performances.
Results: Strong significant positive correlations were observed between human and ASR scores, with coefficients up to .99. However, the spectral smearing used to simulate losses in frequency selectivity caused larger declines in ASR performance than in human performance.
Conclusion: Both intelligibility and comprehension scores for listeners with simulated ARHL are highly correlated with the performances of an ASR-based system. In the future, it needs to be determined if the ASR system is similarly successful in predicting speech processing in noise and by older people with ARHL.

Age-related hearing loss (ARHL)—the progressive decline with increasing age of hearing sensitivity, as measured by an audiometric assessment—affects more than 45% of the population over the age of 48 years (Cruickshanks et al., 1998). The most common complaint of listeners with ARHL is the difficulty to understand speech, especially in noisy environments (e.g., CHABA, 1988). In part, this difficulty results directly from the loss in audibility of the speech signal, but it is also due to additional deficits in suprathreshold auditory processing, such as loudness recruitment, loss in frequency selectivity (e.g., Nejime & Moore, 1997), and reduced sensitivity to temporal–fine structure and temporal-envelope information (e.g., Füllgrabe, 2013; Füllgrabe, Moore, & Stone, 2015). When left uncorrected, speech-perception difficulties can compromise interindividual communication, resulting in various negative consequences for the affected person, such as social isolation (e.g., Strawbridge, Wallhagen, Shema, & Kaplan, 2000), depression (e.g., Gopinath et al., 2009), and accelerated cognitive decline (e.g., Lin et al., 2013).

Currently, the standard treatment for ARHL is digital hearing aids (HAs), providing amplification in a number of frequency channels in order to restore the audibility of sounds. However, up to 40% of the listeners fitted with HAs never or rarely use their devices (Knudsen, Öberg, Nielsen, Naylor, & Kramer, 2010). One explanation for this high rejection rate might be the quality of HA fitting, resulting in suboptimal speech-intelligibility benefits.

[a]Archean Technologies, Montauban, France
[b]IRIT - Université de Toulouse, France
[c]MSHS-T (USR 3414), Université de Toulouse, CNRS, France
[d]CLLE (UMR 5263), Université de Toulouse, CNRS, France
[e]Medical Research Council Institute of Hearing Research, School of Medicine, The University of Nottingham, Nottinghamshire, UK
Correspondence to Lionel Fontan: lfontan@archean.fr

Editor: Julie Liss
Associate Editor: Bharath Chandrasekaran
Received June 24, 2016
Revision received February 10, 2017
Accepted March 14, 2017
https://doi.org/10.1044/2017_JSLHR-S-16-0269

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

Traditionally, speech perception is measured by determining the percentage of speech items (e.g., words) that are correctly identified by the listener. However, a growing number of authors distinguish between speech intelligibility and speech comprehension; intelligibility tests focus on the perception of speech units, whereas speech comprehension tests aim at quantifying the degree to which listeners can interpret the meaning of spoken messages in a communication context (e.g., Fontan, Tardieu, Gaillard, Woisard, & Ruiz, 2015; Hustad, 2008; Wilson & Spaulding, 2010). Intelligibility and comprehension measures may be thought of as complementary because they provide different insights into speech communication. Intelligibility tests yield sensitive and reproducible scores that are mainly dependent on the integrity of the acoustic information present in speech signals. In contrast, the contextual information present in comprehension tests allows listeners to compensate for losses of acoustic information in the speech signal through top-down cognitive processes. Therefore, comprehension scores are less sensitive to small degradations of the speech signal (Lindblom, 1990). They also show a better external validity because they involve processes that are used in everyday communication (Fontan, 2012; Fontan, Gaillard, & Woisard, 2013; Fontan, Tardieu, et al., 2015). Thus, the two kinds of measures might be relevant in professional contexts for which the evaluation of both speech signal transfer and communicative performance is needed. Indeed, it has been shown that performance on one test does not strongly predict performance on the other (Fontan, Tardieu, et al., 2015; Smith, 1992; Smith & Nelson, 2008).

From a practical perspective, both intelligibility and comprehension tests are fairly time consuming, which might limit their clinical use, for example, for the fitting of HAs. In France, audiologists/HA dispensers generally establish speech-processing abilities of their patients/clients by asking them to repeat lists of words (such as those developed by Fournier, 1951). Either a global intelligibility score, corresponding to the percentage of correctly identified words, or the Speech Reception Threshold (SRT), corresponding to the speech level required to obtain 50% correctly identified words, is calculated. In order to establish those HA settings yielding optimal speech intelligibility, the word identification task has to be repeated for each combination of HA settings. Such prolonged testing can result in increased levels of fatigue in the generally older patients/clients, leading to lower identification performance.

Moreover, it has been shown that speech intelligibility scores depend on the listener's familiarity with the speech material (e.g., Hustad & Cahill, 2003). Hence, in theory, speech material should only be used once with the same patient/client, thus limiting the number of combinations of HA settings that can be tested.

To overcome these issues, automatic speech recognition (ASR) could be used to predict speech-processing performance. Indeed, ASR systems have been shown to yield good predictions of human intelligibility and comprehension performance of disordered speech by listeners with normal hearing (e.g., Fontan, Pellegrini, Olcoz, & Abad, 2015; Maier et al., 2009; Schuster et al., 2006). However, to the best of the authors' knowledge, this is the first time that an ASR system is used to predict speech intelligibility and comprehension performance in listeners with simulated ARHL.

Context and Objective of the Present Study

This study is part of a larger research project bringing together language and computer scientists and ear, nose, and throat specialists.[1] The long-term goal of this project is to develop a clinical tool allowing audiologists/HA dispensers to predict speech intelligibility and comprehension for listeners experiencing ARHL in order to facilitate and improve HA fitting. More specifically, the system to be developed would allow recording speech stimuli (e.g., lists of words) in the patient's ear canal near the eardrum, both when wearing a HA and without a HA. The recorded speech is then processed in order to mimic the perceptual consequences of the hearing loss experienced by the patient/client, based on his/her auditory data (e.g., audiogram). The resulting audio signals are then fed to an ASR system that tries to recognize the original speech stimuli. The ASR results (e.g., word error rate [WER]; phonological distances between stimuli and ASR results) and the associated confidence scores are used to predict the intelligibility and comprehension scores human listeners would obtain in the same conditions.

As a first step to reach this goal, several experiments were conducted to study the ability of an ASR system to predict intelligibility and comprehension observed in young participants with normal hearing who are listening to speech processed to simulate ARHL at various levels of severity. This experimental design allowed the same stimuli to be presented to the human listeners and to the ASR system. In addition, it was reasoned that the use of young listeners would reduce the influence of individual differences in cognitive functions (such as working memory) on speech perception (Füllgrabe & Rosen, 2016b) that has been found in older listeners (e.g., Füllgrabe et al., 2015; Füllgrabe & Rosen, 2016a).

Method

Assessment of Human Intelligibility and Comprehension Scores

Speech Material

Word and sentence materials. The material used for the intelligibility tests consisted of 60 words (six lists of 10 words), taken from the intelligibility test developed by Fournier (1951) and widely used by French audiologists/HA dispensers, and of 60 sentences (three lists of 20 sentences),

[1]French Occitanie/Pyrénées-Méditerranée region's AGILE-IT project Grant 12052648 (2012–2015): "Mesure de la compréhension de la parole: équipement électronique intelligent de mesure de la compréhension de la parole basé sur une approche cognitive sur l'exemple de la compréhension humaine."

taken from the French version of the Hearing in Noise Test (HINT; Vaillancourt et al., 2005). Word lists exclusively contained disyllabic masculine nouns preceded by the French definite article "*le*." Sentences consisted of various but rather simple syntactic structures forming single assertive clauses.

The material used for the comprehension test consisted of 30 imperative sentences asking the listener to move virtual objects presented on a computer screen via a click-and-drag action with the computer mouse. These oral commands all matched the following lexico-syntactic pattern:

Mettez [Object 1] [position] [Object 2]

(Move [Object 1] [position] [Object 2]),

where [Object 1] and [Object 2] corresponded to mono- or polysyllabic nouns, and [position] referred to one out of the four spatial prepositions: *à droite de* ("to the right of"), *à gauche de* ("to the left of"), *au-dessus de* ("above"), and *au-dessous de* ("under"). Example sentences are: *Mettez la feuille à gauche du chapeau* ("Move the leaf to the left of the hat") or *Mettez la loupe au-dessus du slip* ("Move the magnifying glass above the underpants").

For each intelligibility and comprehension test, 10 additional items were presented prior to data collection for training purposes.

Speech recordings. Recordings took place in an audiometric booth[2] using an omnidirectional Sennheiser MD46 microphone (Sennheiser, Wedemark, Germany) and a TASCAM DM-3200 mixing console (TEAC Corporation, Tokyo, Japan). Each of three native French speakers (a 12-year-old girl, a 46-year-old man, and a 47-year-old woman) produced all 70 words and 110 sentences. Hence, the entire corpus comprised 210 words and 330 sentences and had a total duration of 12 min. In order to equalize the loudness of the recorded words and sentences, three of the authors adjusted the level of each word and sentence relative to a reference item, and the mean gain values of the adjustments were applied to the stimuli.

Simulation of ARHL. The algorithms described by Nejime & Moore (1997) were used to simulate some of the perceptual consequences of ARHL, using a custom-written MATLAB program (MathWorks, 2015). The program uses the audiometric thresholds as input data. Here, nine levels of hearing-loss severity were simulated. This was done by using the mean audiograms observed for the 3,753 older participants of the Beaver Dam study, grouped into nine age groups with mean ages ranging between 60 and 110 years (Cruickshanks et al., 1998). Figure 1 shows the best polynomial regression curves obtained for fitting mean hearing thresholds at 15 frequencies ranging from 125 Hz to 16 kHz.

Based on the audiometric input data, the program simulates three effects associated with ARHL: (a) reduced audibility (by filtering several frequency bands according to the audiogram values given as an input); (b) reduced frequency selectivity (by spectrally smearing the speech signal; Baer & Moore, 1993); (c) loudness recruitment (by raising the signal envelope; Moore & Glasberg, 1993).

To simulate the reduction of frequency selectivity, the program first defines the degree of hearing loss (mild, moderate, or severe) based on the pure-tone average for audiometric frequencies between 2 and 8 kHz. Three different degrees of spectral smearing are then applied, depending on the degree of hearing loss.

In this study, only one simulation of loudness recruitment was used; the envelope of the speech signal was raised in order to simulate the effect of moderate loudness recruitment (Moore & Glasberg, 1993). This choice made loudness recruitment subject to a high intersubject variability (Moore, 2007); therefore, the level of recruitment cannot be predicted on the sole basis of age and auditory thresholds.

The nine conditions of simulated thresholds, loudness recruitment, and loss of frequency selectivity are shown in Table 1, with corresponding theoretical ages.

The speech corpus was processed to simulate the effects of ARHL at nine levels of severity, resulting in 1,890 word stimuli and 2,970 sentence stimuli (approximate total corpus duration: 115 min). These levels are referred to as ARHL-simulation conditions 1 to 9 in the remainder of the article.

Participants

Sixty university students (34 women, 26 men) ages 18 to 30 years (mean age = 21.3 years; standard deviation = 2.2) took part in this study in exchange for monetary compensation. All were native French speakers and had hearing thresholds ≤ 15 dB HL at 250, 500, and 1000 Hz, and average audiometric thresholds < 15 dB HL for frequencies of 2, 4, and 8 kHz (the latter complied with the definition of no ARHL in the program used to simulate the effects of ARHL). None of the participants reported any uncorrected vision problem.

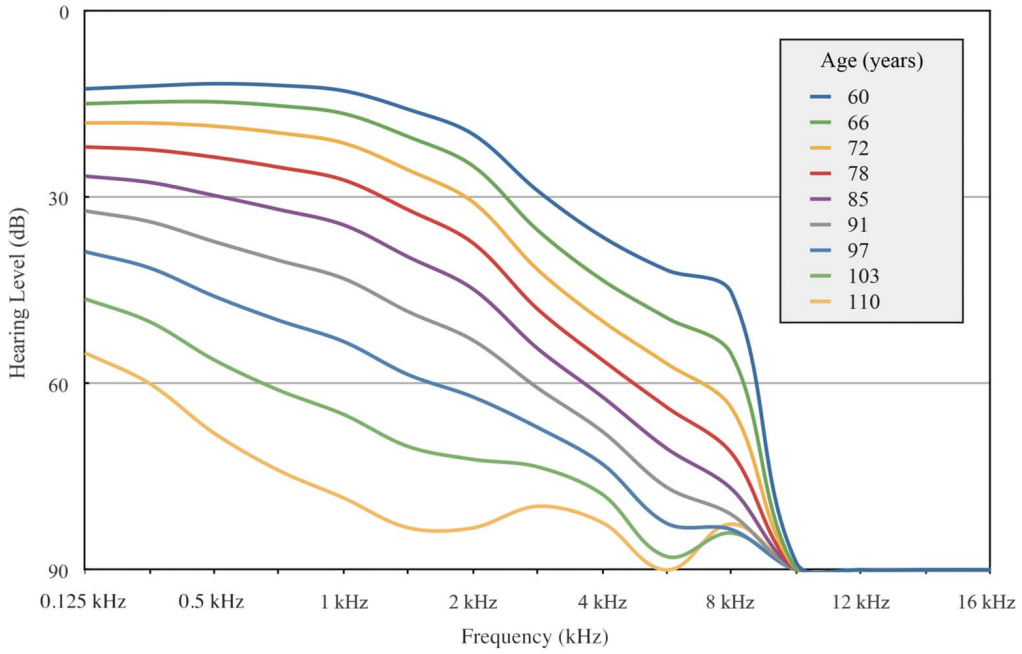
Procedure

Participants completed two intelligibility tests (referred to as IT1 and IT2) and one comprehension test (referred to as CT). They were separated in two groups: Half of the participants completed IT1 and CT, whereas the other half completed IT2 and CT. Consequently, both IT1 and IT2 were completed by 30 participants, whereas CT was completed by all 60 participants. For both participant groups, the order between intelligibility and comprehension tests was counterbalanced. The nine ARHL-simulation conditions were also counterbalanced.

Each participant completed the tests individually in a double-walled sound-attenuating booth (ambient noise level: 28-dB, A-weighted). The participant was seated in front of two Tannoy Precision 6D loudspeakers (Tannoy Ltd., Coatbridge, Scotland, UK), placed at a distance of approximately one meter and at ± 30° azimuth relative to the listener. Speech level was calibrated so that the level of the unprocessed speech stimuli (i.e., stimuli that had not undergone the ARHL-simulation process) reached, on average, 60 dB (A-weighted) at the participant's ear.

[2]http://petra.univ-tlse2.fr

Figure 1. Regression curves fitting mean hearing thresholds for nine mean ages as a function of tone frequency, according to Cruickshanks et al. (1998).



Intelligibility Tests

Participants were asked to repeat what they had heard (words in IT1 and sentences in IT2) into a microphone positioned in front of them. They were encouraged to guess in case they were unsure about what had been presented. Responses were recorded to be transcribed and scored offline by three of the authors. For IT1, a response was judged as "correct" if it matched every phoneme of the target word, and "incorrect" otherwise (binary score). For IT2, the score for each sentence was calculated as the number of correctly identified words divided by the total number of words in the sentence. The final intelligibility scores were calculated as the total percentage of correct responses for each test.

Comprehension Test

Participants were seated facing a 20-in. color monitor displaying six images (evenly distributed into two rows

Table 1. Experimental conditions of age-related hearing loss simulated for ages between 60 and 110 years and associated absolute thresholds, level of loudness recruitment, and reduction of frequency selectivity.

	Condition no.								
Parameter	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Theoretical age (years)	60	66	72	78	85	91	97	103	110
Loudness recruitment	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Reduction of frequency selectivity	Mild	Moderate	Moderate	Moderate	Severe	Severe	Severe	Severe	Severe
					Simulated thresholds (dB HL)				
0.125 kHz	12	15	18	22	27	32	39	46	55
0.250 kHz	12	15	18	22	28	34	41	50	60
0.500 kHz	12	15	19	24	30	37	46	56	68
0.750 kHz	12	15	20	25	32	40	50	61	74
1.000 kHz	13	16	21	27	34	43	53	65	78
1.500 kHz	16	20	26	32	40	48	59	70	83
2.000 kHz	20	25	31	37	45	53	62	72	83
3.000 kHz	29	35	42	48	54	61	67	73	80
4.000 kHz	36	43	50	56	62	68	73	78	82
6.000 kHz	42	49	57	64	70	77	83	88	90
8.000 kHz	45	55	64	71	77	81	84	84	83

and three columns) for each sentence that was played (see Figure 2 in Fontan, Tardieu, et al., 2015). They were asked to respond to each command by selecting and then dragging a target image (Object 1) above/below/to the right of/to the left of another target image (Object 2). A computer mouse was used to perform these actions. Each sentence was considered as understood if the three key elements (Target Image 1, position, Target Image 2) were correctly identified by the listener, based on the actions that were performed on the computer screen. Final CT scores were calculated as the percentage of sentences correctly understood by the listeners in each of the nine ARHL-simulation conditions.

Computing Automatic Intelligibility and Comprehension Scores

In contrast to most research in the field of ASR that has the goal to design the best possible system (in terms of WER), the present work aimed at developing an ASR system that would simulate as closely as possible human behavior, even if that resulted in suboptimal performance. To achieve this, an ASR system based on SPHINX-3 (distributed by Carnegie Mellon University; Seymore et al., 1998) and French acoustic models adapted to the three voices were used. The ESTER2[3] audio and text corpus (Galliano, Gravier, & Chaubard, 2009) was used for the creation of the acoustic models and of a trigram language model that together constituted the baseline ASR system. Different configurations for the lexicon and language models were then developed in order to best fit the human scores.

Description of the ASR System

Acoustic Models

The French acoustic models used in this study came from the Laboratoire d'Informatique de l'Université du Maine (France; Deléglise, Estève, Meignier, & Merlin, 2005; Estève, 2009). They included 35 phones and five kinds of pauses and were designed to process 16-kHz audio samples based on a Perceptual Linear Predictive feature extraction (Hermansky, 1990). Acoustic models were trained upon the basis of French radio broadcasts (Galliano et al., 2009) and were formed by 5,725 context-dependent states (senones) with 22 Gaussian mixtures per state.

Speaker Adaptation

The acoustic models were trained on recordings that included more male than female voices (Galliano et al., 2006). This caused the system to show a better WER for the male than for the female and child speaker. As a consequence, a first step of the present work was to adapt the ASR system to the voices of the speakers in order to obtain ASR performances as similar as possible (in terms of WER) for the adult-male, adult-female, and child speech. To this end, the vocal-tract-length normalization technique (VTLN; Wegmann, McAllaster, Orloff, & Peskin, 1996) was used. The VTLN technique is based on the assumption that there is a direct linear relationship between speech formant areas and the vocal tract length of the speaker. It consists of determining the best warping factor λ that maximizes the likelihood of the phones produced by a specific speaker:

$$\lambda = \arg\max P(0|X, \lambda_k), \quad (1)$$

where X is the observation and k the index of the kth frequency warping factor considered. In order to find the best λ value for each of the three speakers, the inverse linear function $y = x/\lambda$ was used and tested on the IT1 subcorpus. Results showed that the adult-male speaker recordings did not need any adaptation ($\lambda = 1.0$), whereas the adult-female speaker and the girl speaker recordings needed VTLN with optimal λ values of 1.84 and 2.24, respectively. These two warping factors were systematically used for the extraction of female and child speech features during automatic speech recognition.

Lexicon and Language Models

The lexicon and language modeling constituted the main stage in the fitting of the ASR scores to human scores. Based on the assumption that human performance is greatly influenced by top-down expectations at lexical, syntactic, and contextual levels, the goal was to feed the system with similar lexical and syntactic cues to constrain its behavior and thus to get a better linear correlation with human data.

To this end, different ways of modeling the lexicon and syntax were explored for each of the three tests, and the results compared to those obtained with the baseline model. The latter contained very low constraints upon the lexicon and syntactic structures to be recognized; it is a trigram language model calculated on the basis of the ESTER2 corpus (Deléglise et al., 2005; Galliano et al., 2009) associated with a 62,000-word lexicon and is henceforth referred to as BM.

For IT1, one additional language model was designed in order to reflect the syntactic, lexical, and phonological properties of the test stimuli. This model, referred to as IT1M, is a bigram language model based on the syntactic structure [Det + Noun], and its lexicon contains only disyllabic masculine nouns beginning with a consonant (15,000 forms). To reflect the frequency of these forms in oral French, the frequency values defined by New, Brysbaert, Veronis, and Pallier (2007), based on movie subtitles and available in the database Lexique 3.8,[4] were used.

For IT2, four additional language models were designed:

- IT2M1 is a trigram language model, based on a subcorpus of ESTER2; this subcorpus consists of the ESTER2 utterances containing at least one word occurring in IT2 sentences.

[3]Évaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques (Evaluation of Broadcast News Enriched Transcription Systems)

[4]http://www.lexique.org

- IT2M2 is the same trigram model as BM, but with a lexicon restricted to the words constituting the 260 sentences of the complete French version of the HINT.
- IT2M3 is the same trigram model as BM, but with a lexicon restricted to the words constituting the 60 HINT sentences included in the IT2 test.
- IT2M4 is an finite-state grammar (FSG), allowing the generation of the 60 HINT sentences included in the IT2 test.

Finally, two additional language models were designed for CT:

- CTM1 is an FSG, allowing the generation of all the possible combinations in the CT test, that is, around 50,000 sentences (112 objects × 4 positions × 111 objects);
- CTM2 is the same FSG as CTM1, but associated with a dynamic lexicon; for each sentence processed by the ASR system, only the nouns corresponding to the six images actually presented to the listeners during CT were included in the lexicon.

Table 2 summarizes the different language models used for the automatic recognition.

ASR-Score Calculation. For the recognition of the intelligibility-test items (words and sentences for IT1 and IT2, respectively) only the percentage of correct words was considered as an outcome measure. The determinant "le" was not taken into account for IT1 items in order to follow the scoring procedure used with human participants. For the recognition of CT items, only the recognition of the three main keywords of each sentence (i.e., Object 1, position, Object 2) was considered and the ASR score corresponds to the percentage of sentences for which all keywords were recognized by the system.

Table 2. Language models used for the automatic recognition of Intelligibility Test 1 (IT1), Intelligibility Test 2 (IT2), and Comprehension Test (CT) stimuli.

Model	Description
Baseline model (BM)	Trigrams (ESTER2 corpus) and 62,000-word lexicon
IT1	
IT1M	Bigrams (Det + N) with word frequencies based on Lexique 3.8
IT2	
IT2M1	Trigrams calculated on a subcorpus of ESTER2
IT2M2	BM trigrams with a lexicon restricted to that of the 260 HINT sentences
IT2M3	BM trigrams with a lexicon restricted to that of the 60 IT2 sentences
IT2M4	FSG allowing the generation of the 60 IT2 sentences
CT	
CTM1	FSG allowing the generation of the ~50,000 sentences possibly combined in CT
CTM2	Same FSG as CTM1 associated with a dynamic lexicon

Note. HINT = Hearing in Noise Test; FSG = finite-state grammar.

Results

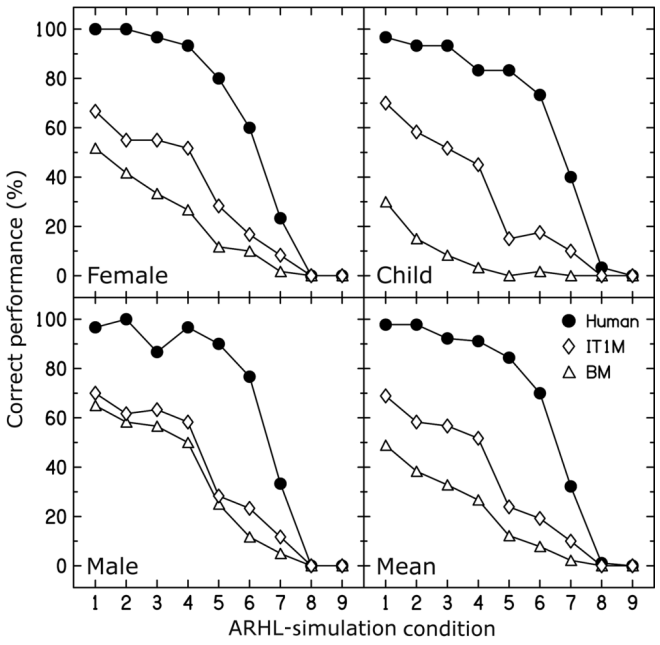
Word Intelligibility Test (IT1)

Word-identification performance for the human listeners and the ASR system using two different language models are presented in Figure 2 as a function of simulated-ARHL condition. The different panels show average results for each of the three speakers and the grand average.

Human word identification declines sigmoidally with increasing severity of the simulated ARHL condition, with, on average, Conditions 1 and 2 and Conditions 8 and 9 yielding ceiling and floor effects, respectively.

Machine scores are generally lower than human scores (mean: 63%), with IT1M yielding higher scores (mean: 32.1%) than BM (mean: 18.8%), but also followed a downward trend with increasing ARHL-simulation condition. However, the shape of the performance functions for the two language systems differs from that for human listeners (e.g., it is linear for BM/female and concave for BM/child). Marked decreases in performance can be seen

Figure 2. Word intelligibility for Intelligibility Test 1 (IT1) as a function of the condition of simulated age-related hearing loss (ARHL) for human listeners (filled circles) and the automatic speech recognition (ASR) system, using different language models (see open symbols in the figure legend). The different panels show results for each of the three speakers (female, child, and male) and averaged across speakers (mean). BM = baseline model.



between Conditions 1 and 2 and Conditions 4 and 5. On average, the highest machine scores are obtained for the male speaker ($mean_{BM}$: 30.2%, $mean_{IT1M}$: 35.2%), then the female speaker ($mean_{BM}$: 19.6%, $mean_{IT1M}$: 31.3%), and finally the child speaker ($mean_{BM}$: 6.5%, $mean_{IT1M}$: 29.7%).

To establish the goodness-of-fit of the machine scores for human word intelligibility as a function of ARHL-simulation condition, Pearson's linear correlation coefficient was computed for each combination of language model and speaker condition (see Table 3). Given the existence of floor and ceiling effects, scores were first transformed into rational-arcsine units (RAUs; Studebaker, 1985). As expected based on the visual inspection of the results, all correlations were positive, strong (ranging from .71 to .99), and significant (all $p \leq .032$, two tailed). Comparing correlation coefficients for the ASR system using the two different language models, after applying Fisher's r-to-z transformation, revealed a significant difference (i.e., improvement in the strength of the correlation) between BM and IT1M only for the child speaker ($z = -2.95$, $p = .002$, one tailed; a one-tailed test was used because it was assumed that BM would yield lower correlation coefficients than the more "sophisticated" models).[5] The left panel of Figure 3 shows the scatterplot relating mean RAU-transformed human and machine scores for the BM and the best linear fit. Because ASR performance is lower than human performance for all conditions, all data points (except for the most severe ARHL-simulation condition yielding the lowest possible RAU score for listeners and the ASR system) fall above the diagonal that indicates identical performance for both the human and machine listener.

Table 3. Pearson's correlation coefficients for Intelligibility Test 1 (IT1) between human intelligibility scores and automatic speech recognition (ASR) scores, using different language models and for all speakers combined (mean) or individual speakers.

	Speaker			
Model	**Mean**	**Male**	**Female**	**Child**
BM	.94 (.000)	.93 (.000)	.97 (.002)	.71 (.032)
IT1M1	.97 (.000)	.95 (.000)	.99 (.000)	.93 (.000)**

Note. p values for two-tailed t tests are given in parentheses. Asterisks indicate the p values for one-tailed tests of the difference between the correlation coefficient for the baseline model (BM) and each of the other language models. All tests were uncorrected for multiple comparisons.

**$p < .01$; one-tailed test re: BM.

Sentence Intelligibility Test (IT2)

Sentence-identification performance for the human listeners and the ASR system using five different language models are presented as a function of simulated-ARHL condition in Figure 4.

As for word identification, human sentence identification declines sigmoidally with ARHL-simulation condition, with the ceiling effect extending up to and including Condition 4. On average, identification for sentences (mean: 64.6%) was very similar to that for words.

The ASR system generally yields lower than human scores, independently of the language models used (mean performance ranges from 30.9% to 51.8%), which, consistent with the human results, decline with ARHL-simulation condition. However, the shape of these functions differs from that of human performance; marked decreases are again observed between Conditions 4 and 5 for all language models, and, for IT2M4, performance for Conditions 8 and 9 is actually better than human performance. Across all ARHL simulations and language models, the male speaker yields the highest machine scores (mean: 46.6%), the female speaker the second highest (mean: 40.2%), and the child speaker the lowest (mean: 35%).

Table 4 indicates Pearson's linear correlation coefficients between RAU-transformed human and machine scores for the different language models and speakers. All correlations were positive, strong (ranging from .90 to .99), and significant (all $p \leq .001$, two tailed). Compared to the correlation coefficient obtained for the BM—and when averaging performance across speakers—the four other language models yield significantly stronger correlations (all $z \geq -2.37$, all $p \leq .009$, one tailed). However, when considering performances for each speaker individually, some models did not yield a significant improvement in the strength of the correlation (IT2M1 with the male speaker and IT2M2 and IT2M4 with the female speaker).

The middle panel of Figure 3 shows the scatterplot relating mean RAU-transformed human and machine scores for the BM and IT2M4 (which yielded significantly higher correlation compared to the BM) and the best linear fits. The regression lines for the BM and IT2M4 have comparable slopes; however, the y-intercept of the regression line is higher with the BM than with IT2M4, showing that the difference between human and ASR scores is reduced for IT2M4. For both BM and IT2M4, the regression lines present degrees of incline > 45 degrees, showing that the differences between human and ASR performances tend to increase with the elevation of scores.

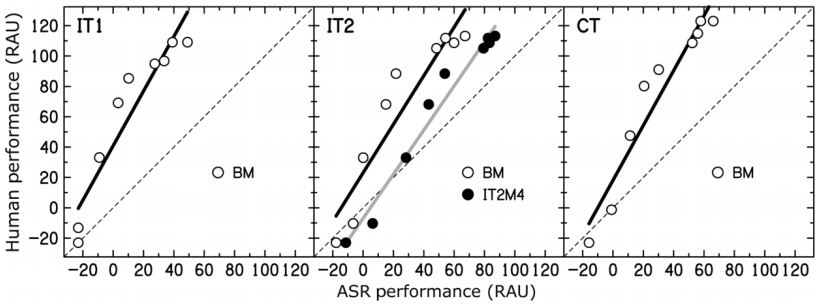
Comprehension Test (CT)

Comprehension performance for the human listeners and the ASR system using three different language models are presented as a function of simulated-ARHL condition in Figure 5.

Human speech comprehension (mean: 68.7%) is only slightly better than word and sentence identification but shows a similar dependence on ARHL-simulation condition and extent of the ceiling effect.

[5]The calculation of the significance of the difference between the two correlations was checked according to Steiger (1980) on the online software made available by Lee and Preacher (2013).

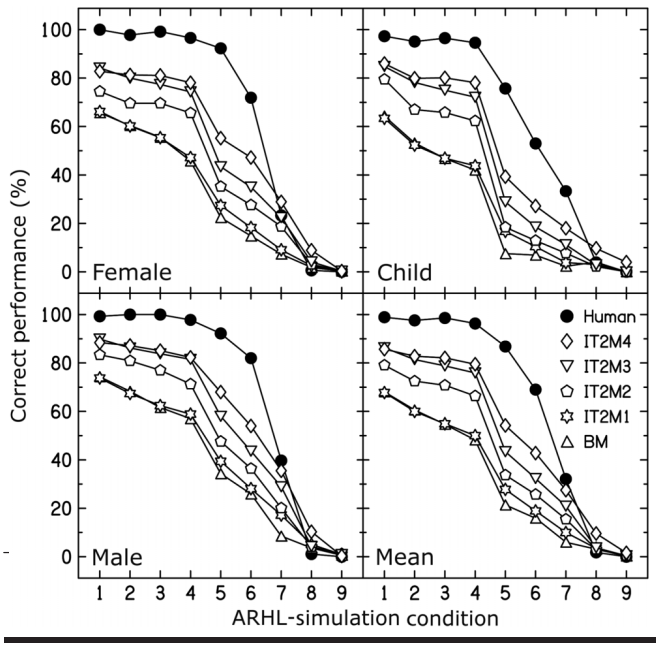
Figure 3. Scatterplots relating the mean rational-arcsine unit (RAU)-transformed human scores to RAU-transformed machine scores obtained in each of the three speech tests, Intelligibility Test 1 (IT1), Intelligibility Test 2 (IT2), and Comprehension Test (CT). Results for the baseline model (BM) and a more sophisticated language model (IT2M4) showed a significant improvement in the strength of the correlation. ASR = automatic speech recognition.



Machine scores increased from BM (mean: 33.6%) over CTM2 (mean: 54.7%) to CTM1 (mean: 68%) but are lower than human scores. Once again, a marked decline in performance is observed between ARHL-simulation conditions 4 and 5. Interestingly, both more sophisticated language models (especially CTM1) yield performance for the most severe ARHL conditions that exceeds that observed in human listeners. Across all ARHL simulations and language models, highest machine performance was observed for the male speaker (mean: 57.3%), then the child speaker (mean: 51.2%), and finally the female speaker (mean: 47.8%).

Pearson's correlation coefficients between RAU-transformed human and machine scores, presented in Table 5, were positive, strong (ranging from .91 to .98), and significant (all $p \leq .001$, two tailed). Compared to BM, both more sophisticated models CTM1 and CTM2 do not yield significantly stronger correlations with human scores for any of the speaker conditions. The right panel of Figure 3 shows the scatterplot relating mean RAU-transformed human and machine scores for the BM and the best linear fit. The *y*-intercept of the regression line is almost situated on the diagonal indicating identical performance for both the human and machine listener. However, for higher scores, the difference between human and ASR performance tends to increase.

Figure 4. Sentence intelligibility for Intelligibility Test 2 (IT2) as a function of the condition of simulated age-related hearing loss (ARHL) for human listeners (filled circles) and the automatic speech recognition (ASR) system, using different language models (see open symbols in the figure legend). The different panels show results for each of the three speakers (female, child, and male) and averaged across speakers (mean). BM = baseline model.



Discussion

The long-term goal of this research work is to develop an ASR-based system able to predict human speech-processing performance, with the purpose of facilitating HA fitting for people with ARHL. The current study constituted the first step toward this goal by comparing ASR results to speech intelligibility and comprehension scores of young participants with normal hearing who are listening to speech processed to simulate ARHL.

The observed strong correlations for all three tests and all speaker conditions indicate that the ARHL simulation, representing the different levels of severities of the perceptual consequence of hearing loss associated with ages 60 to 110 years, had comparable effects on human and ASR scores. Thus, it appears that ASR systems could be used to predict trends in human speech intelligibility and comprehension with increasing level of ARHL.

However, weaker machine scores were generally found when processing words and sentences uttered by the female and child speakers than those uttered by the male speaker. This is probably due to the acoustic models used in this study. They were trained on speech recordings consisting of two thirds male voices (Galliano et al., 2006).

Table 4. Pearson's correlation coefficients and associated *p* values (in parentheses) for Intelligibility Test 2 (IT2) performance between human intelligibility scores and automatic speech recognition (ASR) scores, using different language models and for all speakers combined (mean) or individual speakers.

	Speaker			
Model	**Mean**	**Male**	**Female**	**Child**
BM	.94 (.000)	.95 (.000)	.94 (.000)	.90 (.001)
IT2M1	.96 (.000)**	.97 (.000)	.96 (.000)*	.94 (.000)**
IT2M2	.97 (.000)**	.97 (.000)**	.97 (.000)	.95 (.000)***
IT2M3	.98 (.000)**	.98 (.000)*	.97 (.000)*	.97 (.000)***
IT2M4	.99 (.000)***	.99 (.000)*	.98 (.000)	.96 (.000)***

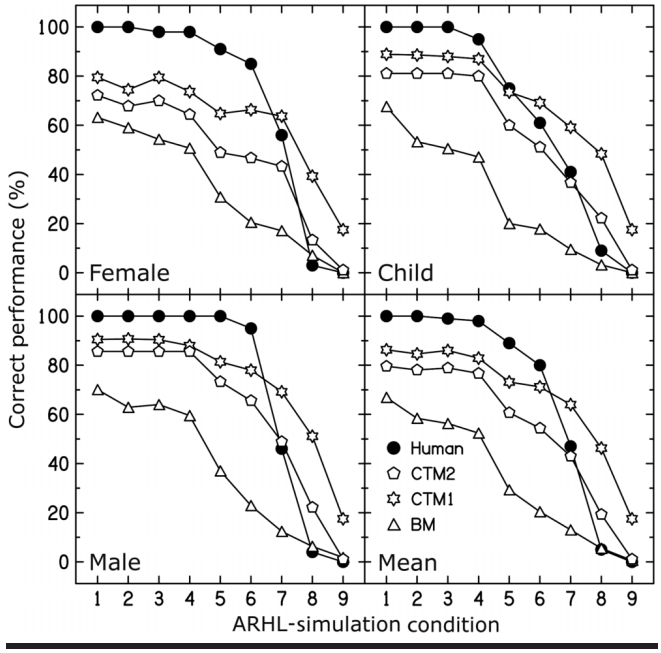
Note. BM = baseline model.

*$p \leq .05$; **$p < .01$; ***$p \leq .001$; one-tailed test re: BM.

Although the VTLN technique (Wegmann et al., 1996) was used to adapt the ASR system to the different voices, the results indicate that this did not fully eliminate interspeaker differences.

In order to obtain the strongest associations possible between ASR scores and human performance, several language models were designed by taking into account the phonological, lexical, syntactic, and extralinguistic cues that underlie human performance in IT1, IT2, and CT. A model with very low constraints on the lexicon and syntactic structures (BM) was used to generate machine baseline performance against which improvements due to the different, more sophisticated language models could be measured.

Figure 5. Speech comprehension (for the Comprehension Test [CT]) as a function of the condition of simulated age-related hearing loss (ARHL) for human listeners (filled circles) and the automatic speech recognition (ASR) system, using different language models (see open symbols in the figure legend). The different panels show results for each of the three speakers (female, child, and male) and averaged across speakers (mean). BM = baseline model.



Apart from IT2, model sophistication did not yield significant improvements in the strength of the association between machine and human scores. A possible partial explanation for this finding is that the correlation between human and BM scores was already very high (≥ .90 except for IT1 and the child speaker), and this left little room for improvements. When looking at the conditions yielding the lowest correlations with BM (e.g., child speech conditions in IT1 and IT2), highly significant improvements were observed when using more sophisticated models.

Floor and ceiling effects were observed in the human performance for all three speech tests. However, the number of ARHL-simulation conditions that were affected varied with the test used, with CT > IT2 > IT1 as regards ceiling effects, and IT1 > IT2 > CT as regards floor effects. This observation is consistent with the assumption that contextual effects have an increasingly beneficial effect on human performance as speech processing goes from word identification over sentence identification to comprehension (Lindblom, 1990; Fontan, Tardieu, et al., 2015). The aim of the present study was to explore the effect of the full range of levels of ARHL (from mild to severe) on speech processing in quiet. The existence of floor and ceiling effects was therefore unavoidable. The application of a RAU transformation to the raw data allowed us to overcome the problems associated with such effects, at least to some extent.

To enhance the performance of the prediction system and to extend its applicability to other groups of listeners

Table 5. Pearson's correlation coefficients and associated *p* values (in parentheses) for Comprehension Test (CT) performance between human intelligibility scores and automatic speech recognition (ASR) scores, using different language models and for all speakers combined (mean) or individual speakers.

	Speaker			
Model	**Mean**	**Male**	**Female**	**Child**
BM	.96 (.000)	.91 (.001)	.95 (.000)	.98 (.000)
CTM1	.97 (.000)	.95 (.000)	.96 (.000)	.98 (.000)
CTM2	.98 (.000)	.97 (.000)	.98 (.000)	.98 (.000)

Note. BM = baseline model.

and to other listening conditions, the following lines of research will be pursued in the future.

1. Acoustic models will be trained on speech corpora containing a balance of male and female adult and child voices. Also, because the main difficulty of people experiencing ARHL is to understand speech in the presence of interfering sounds, the applicability of these models to the recognition of speech in adverse listening conditions (e.g., background noise, reverberation) needs to be assessed. Whether our findings can be generalized to other languages than French also needs to be established.
2. In the present study, the goodness-of-fit of the prediction by machine scores of the observed trends in average human performance for a range of simulated ARHL conditions was quantified. Because the ultimate aim of the system is to help audiologists/HA dispensers with the fitting of HAs to individual patients/clients, the system's predictive power of individual cases of ARHL needs to be tested. Also, the present study used young listeners with normal hearing for whom ARHL was simulated. This choice was made so that the same stimuli could be presented to the participants and to the ASR system. However, the participants tested in the present study differed considerably from the population for which the prediction system was originally designed, namely people with ARHL who are generally older. Performance in many cognitive abilities declines with age (for example, in working memory, attention, processing speed, inhibition; e.g., Füllgrabe et al., 2015; Park et al., 2002). In addition, age-related changes in supraliminary auditory processing are observed in older listeners, for example in the processing of temporal–fine structure (Füllgrabe, 2013; Füllgrabe et al., 2015; Grose & Mamo, 2010; Moore, Glasberg, Stoev, Füllgrabe, & Hopkins, 2012) and temporal-envelope information (Füllgrabe, Meyer, & Lorenzi, 2003; Füllgrabe et al., 2015; He, Mills, Ahlstrom, & Dubno, 2008). As both temporal processing and cognitive abilities are associated with speech-in-noise perception (e.g., Füllgrabe et al., 2015), it is probably necessary to consider these abilities in future versions of the ASR-based prediction system to yield accurate predictions of speech-perception performance for all listeners across the adult lifespan.
3. The simulations of ARHL used to produce the degraded input to the prediction system might have to be refined. For example, in the current study, the simulation of loudness recruitment was fixed at one severity condition corresponding to a moderate level of loudness recruitment. This choice was made because the severity of loudness recruitment is highly variable among people experiencing ARHL (Moore, 2007). However, this restriction has a potential impact on the ability of the system to accurately predict speech intelligibility and comprehension for older people presenting different degrees of loudness recruitment. In a next step, the level of loudness recruitment should be manipulated in order to verify that its effect on ASR scores is comparable to its effects on human speech recognition performance. In addition, strong decreases in ASR scores were observed between ARHL-simulation conditions 4 and 5, and, to a lesser extent, between ARHL-simulation conditions 1 and 2. These decreases in performance are probably caused by the spectral smearing, which simulated the consequences of losses in frequency selectivity on speech perception. As shown in Table 1, going from ARHL-simulation conditions 1 to 2 and from 4 to 5 coincides with a change in the simulated severity of the loss of frequency selectivity. No comparable effect was observed in the human scores, consistent with results reported by Baer and Moore (1993), showing that the spectral-smearing algorithm had a significant effect on speech intelligibility only when stimuli were heard in noisy conditions. This matter warrants further investigation on the use of an ASR-based system to predict human processing performance for spectrally smeared speech presented in noise.
4. To further address the role of top-down effects on speech processing, future work will investigate different test materials and tasks from those used in the present study. For example, nonsense or low-predictable sentences such as those used in the Matrix test (Vlaming et al., 2011) could be used to eliminate the syntactic and semantic predictability present in the sentences used in IT2 and CT. Also, using sentences requiring the listener to process other linguistic cues to interpret the meaning of the sentences (e.g., thematic roles) could be relevant.
5. Finally, this research concentrated on correlations between machine and humans scores, that is, on the predictability of the trends observed in human speech intelligibility and comprehension. Further research is needed to assess the ability of the ASR system to predict actual intelligibility and comprehension scores. Also, in addition to the purely quantitative prediction, it would be of interest to investigate whether qualitative aspects of human speech processing can be predicted by the ASR system. For example, the analysis of the phonetic confusions predicted by the ASR system might provide insights into which acoustic features are misperceived by human listeners (Fontan, Ferrané, Farinas, Pinquier, & Aumont, 2016; e.g., predicting that with a specific HA setting the listener will tend to perceive stop consonants as their constrictive counterparts). This information might prove very helpful to audiologists/HA dispensers for the fine-tuning of HAs.

Taken together, our results indicate that ASR-based prediction systems are able to provide good estimates of the trends in human speech-processing abilities that can

be seen with increasing levels of ARHL. In the future, this might help to optimize the fitting process of HAs in terms of the time necessary to find optimal HA settings and the amount of benefit derived from HAs, and therefore reduce their rejection by people experiencing ARHL.

Acknowledgments

This research was supported by the European Regional Development Fund within the framework of the Occitanie/Pyrénées-Méditerranée French region's AGILE-IT 2012 project grant. The project, for which a European patent (Aumont & Wilhem-Jaureguiberry, 2009) has been filed, is led by Archean Technologies (France).

The project won the grand prize of the 2016 French Aviva Community Fund "La Fabrique" in the category of health-related innovations.

Parts of this study were presented at the INTERSPEECH 2015 conference (Fontan, Farinas, Ferrané, Pinquier, & Aumont, 2015) in Dresden (Germany) and at the INTERSPEECH 2016 conference (Fontan, Ferrané, Farinas, Pinquier, & Aumont, 2016) in San Francisco (USA).

The Medical Research Council Institute of Hearing Research is supported by the Medical Research Council (Grant U135097130).

References

Aumont, X., & Wilhem-Jaureguiberry, A. (2009). *European Patent No. 2136359—Method and Device for Measuring the Intelligibility of a Sound Distribution System*. Courbevoie, France: Institut National de la Propriété Industrielle.

Baer, T., & Moore, B. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *Journal of the Acoustical Society of America, 94*(3), 1229–1241. https://dx.doi.org/10.1121/1.408176

CHABA. (1988). Speech understanding and aging. *Journal of the Acoustical Society of America, 83*(3), 859–895. https://dx.doi.org/10.1121/1.395965

Cruickshanks, K., Wiley, T., Tweed, T., Klein, B., Klein, R., Mares-Perlman, J., & Nondahl, D. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin: The epidemiology of hearing loss study. *American Journal of Epidemiology, 148*(9), 879–886. https://dx.doi.org/10.1093/oxfordjournals.aje.a009713

Deléglise, P., Estève, Y., Meignier, S., & Merlin, T. (2005). The LIUM speech transcription system: A CMU Sphinx III-based system for French broadcast news. In *Proceedings of Interspeech '05* (pp. 1653–1656). Lisbon, Portugal: International Speech and Communication Association.

Estève, Y. (2009). *Traitement automatique de la parole: Contributions (Automatic speech processing: Contributions)*. (Thesis for the *Habilitation à Diriger des Recherches* authorization). Le Mans (France): Université du Maine.

Fontan, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication* (Doctoral dissertation). University of Toulouse, Toulouse, France.

Fontan, L., Farinas, J., Ferrané, I., Pinquier, J., & Aumont, X. (2015). Automatic intelligibility measures applied to speech signals simulating age-related hearing loss. In *Proceedings of Interspeech '15* (pp. 663–667). Dresden, Germany: International Speech and Communication Association.

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., & Aumont, X. (2016). Using phonologically weighted Levenshtein distances for the prediction of microscopic intelligibility. In *Proceedings of Interspeech '16* (pp. 650–654). San Francisco, CA: International Speech and Communication Association.

Fontan, L., Gaillard, P., & Woisard, V. (2013). Comprendre et agir: Les tests pragmatiques de compréhension de la parole et EloKanz. In R. Sock, B. Vaxelaire, & C. Fauth (Eds.), *La voix et la parole perturbées* (pp. 131–144). Mons, Belgium: CIPA.

Fontan, L., Pellegrini, T., Olcoz, J., & Abad, A. (2015). Predicting disordered speech comprehensibility from goodness of pronunciation scores. In *Sixth Workshop on Speech and Language Processing for Assistive Technologies: SLPAT 2015—Satellite workshop of Interspeech '15,* Dresden, Germany. Retrieved from http://www.slpat.org/slpat2015/papers/fontan-pellegrini-olcoz-abad.pdf

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., & Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research, 58*(3), 977–986. https://dx.doi.org/10.1044/2015_jslhr-h-13-0335

Fournier, J. (1951). *Audiométrie vocale*. Paris, France: Maloine.

Füllgrabe, C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. *American Journal of Audiology, 22*(2), 313–315. https://dx.doi.org/10.1044/1059-0889(2013/12-0070)

Füllgrabe, C., Meyer, B., & Lorenzi, C. (2003). Effect of cochlear damage on the detection of complex temporal envelopes. *Hearing Research, 178*(1-2), 35–43.

Füllgrabe, C., Moore, B., & Stone, M. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience, 6*(347), 1–25. https://dx.doi.org/10.3389/fnagi.2014.00347

Füllgrabe, C., & Rosen, S. (2016a). Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing. *Advances in Experimental Medicine and Biology, 894,* 29–36. https://dx.doi.org/10.1007/978-3-319-25474-6_4

Füllgrabe, C., & Rosen, S. (2016b). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology, 7,* 1268. https://doi.org/10.3389/fpsyg.2016.01268

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. In *Proceedings of the 5th international Conference on Language Resources and Evaluation: LREC 2006* (pp. 315–320). Genova, Italy: European Language Resources Association.

Galliano, S., Gravier, G., & Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech '09* (pp. 2583–2586). Brighton, United Kingdom: International Speech and Communication Association.

Gopinath, B., Wang, J. J., Schneider, J., Burlutsky, G., Snowdon, J., McMahon, C. M., … Mitchell, P. (2009). Depressive symptoms in older adults with hearing impairments: The Blue Mountains study. *Journal of the American Geriatrics Society, 57*(7), 1306–1308. https://dx.doi.org/10.1111/j.1532-5415.2009.02317.x

Grose, J. H., & Mamo, S. K. (2010). Processing of temporal fine structure as a function of age. *Ear & Hearing, 31*(6), 755–760. https://dx.doi.org/10.1097/AUD.0b013e3181e627e7

He, N. J., Mills, J. H., Ahlstrom, J. B., & Dubno, J. R. (2008). Age-related differences in the temporal modulation transfer function with pure-tone carriers. *Journal of the Acoustical Society of America, 124*(6), 3841–3849. https://dx.doi.org/10.1121/1.2998779

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America, 87*(4), 1738–1752. https://dx.doi.org/10.1121/1.399423

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51,* 562–573.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*(2),198–208. https://dx.doi.org/10.1044/1058-0360(2003/066)

Knudsen, L. V., Öberg, M., Nielsen, C., Naylor, G., & Kramer, S. E. (2010). Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids: A review of the literature. *Trends in Amplification, 14*(3), 127–154. https://doi.org/10.1177/1084713810385712

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Retrieved from http://www.quantpsy.org.

Lin, F. R., Yaffe, K., Xia, J., Xue, Q. L., Harris, T. B., Purchase-Helzner, E., . . . Health ABC Study Group FT. (2013). Hearing loss and cognitive decline in older adults. *JAMA Internal Medecine, 173*(4), 293–299. https://dx.doi.org/10.1001/jamainternmed.2013.1868

Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech. *Augmentative and Alternative Communication, 6,* 220–230. https://dx.doi.org/10.1080/07434619012331275504

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Nöth, E. (2009) PEAKS—A system for the automatic evaluation of voice and speech disorders. *Speech Communication, 51*(5), 425–437.

MathWorks. (2015). MATLAB [Computer software]. *MATLAB and statistics toolbox release*. Natick, MA: Author.

Moore, B. C. J. (2007). *Cochlear hearing loss: Physiological, psychological and technical issues*. Chichester, United Kingdom: Wiley.

Moore, B. C. J., & Glasberg, B. R. (1993). Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech. *Journal of the Acoustical Society of America, 94*(4), 2050–2062. https://dx.doi.org/10.1121/1.407478

Moore, B. C. J., Glasberg, B. R., Stoev, M., Füllgrabe, C., & Hopkins, K. (2012). The influence of age and high-frequency hearing loss on sensitivity to temporal fine structure at low frequencies (L). *Journal of the Acoustical Society of America, 131*(2), 1003–1006. https://dx.doi.org/10.1121/1.3672808

Nejime, Y., & Moore, B. C. J. (1997). Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. *Journal of the Acoustical Society of America, 102*(1), 603–615.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*(4), 661–677.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17*(2), 299–320.

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., . . . Nöth, E. (2006). Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology, 70*(10), 1741–1747. https://dx.doi.org/10.1016/j.ijporl.2006.05.016

Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., . . . Thayer, E. (1998). The 1997 CMU Sphinx-3 English broadcast news transcription system. In *Proceedings of the 1998 DARPA Speech Recognition Workshop* (pp. 55–59). Lansdowne, VA: Morgan Kaufmann Publishers.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Karhru (Ed.), *The other tongue: English across cultures* (pp. 75–90). Urbana, IL: University of Illinois Press.

Smith, L. E., & Nelson, C. L. (2008). World Englishes and issues of intelligibility. In B. Kachru, Y. Kachru, & C. L. Nelson (Eds.), *The handbook of world Englishes* (pp. 428–435). Malden: Blackwell Publishing.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Strawbridge, W. J., Wallhagen, M. I., Shema, S. J., & Kaplan, G. A. (2000). Negative consequences of hearing impairment in old age: A longitudinal analysis. *Gerontologist, 40*(3), 320–326.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research, 28,* 455–462.

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., . . . Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian francophone populations. *International Journal of Audiology, 44*(6), 358–361. https://dx.doi.org/10.1080/14992020500060875

Vlaming, M. S. M. G., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B., Mohammadh, Y., & Houtgast, T. (2011). HearCom: Hearing in the communication society. *Acta Acustica United With Acustica, 97*(2), 175–192. https://doi.org/10.3813/AAA.918397

Wegmann, S., McAllaster, D., Orloff, J., & Peskin, B. (1996). Speaker normalization on conversational telephone speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 339–341). Atlanta, GA: IEEE Signal Processing Society.

Wilson, E. O., & Spaulding, T. J. (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *Journal of Speech, Language, and Hearing Research, 53,* 1543–1554.

A.3 Article dans le journal Language Ressources and Evaluation

L'article intitulé « C2SI corpus : a Database of Speech Disorder Productions to Assess Intelligibility and Quality of Life in Head and Neck Cancers » soumis au journal Language Ressource and Evaluation [Woisard et al., 2021] détaille le corpus de voix pathologiques de patients atteints de cancers ORL issu des projets C2SI (cf. section 5.4.12 page 80) et RUGBI (cf. section 5.4.15 page 83). La collecte d'enregistrements audio, des metadonées et des enrichissements produits sont décrits et les différents choix de conception décrits. Ce corpus est le socle des nombreux travaux de recherche qui en ont découlé et constitue également la première pierre du GIS PAROLOTHEQUE (cf. section 5.4.20 page 86).

Lang Resources & Evaluation (2021) 55:173–190
https://doi.org/10.1007/s10579-020-09496-3

ORIGINAL PAPER

C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers

Virginie Woisard[1] · Corine Astésano[2] · Mathieu Balaguer[1] · Jérôme Farinas[5] · Corinne Fredouille[4] · Pascal Gaillard[2] · Alain Ghio[3] · Laurence Giusti[3] · Imed Laaridh[5] · Muriel Lalain[3] · Benoît Lepage[1] · Julie Mauclair[5] · Olivier Nocaudie[2] · Julien Pinquier[5] · Gilles Pouchoulin[3] · Michèle Puech[1] · Danièle Robert[3] · Vincent Roger[5]

Published online: 15 June 2020

Abstract Within the framework of the Carcinologic Speech Severity Index (C2SI) INCa Project, we collected a large database of French speech recordings aiming at validating Disorder Severity Indexes. Such a database will be useful for measuring the impact of oral and pharyngeal cavity cancer on speech production. It will permit to assess patients' quality of life after treatment. The database is composed of audio recordings from 134 sessions and associated metadata. Several intelligibility and comprehensibility levels of speech functions have been evaluated. Acoustics and prosody have been assessed. Perceptual evaluation rates from both naive and expert juries are being produced. Automatic analyzes are being carried out. It is intended to provide speech therapists and physicians with objective tools, which take into account the intelligibility and comprehensibility of patients which received cancer

✉ Jérôme Farinas
jerome.farinas@irit.fr

Virginie Woisard
woisard.v@chu-toulouse.fr

Corine Astésano
astesano@univ-tlse2.fr

Corinne Fredouille
corinne.fredouille@univ-avignon.fr

Alain Ghio
alain.ghio@lpl-aix.fr

[1] Toulouse University Hospital Centre, Toulouse, France

[2] Jean Jaures University, Toulouse, France

[3] CNRS UMR 7309, LPL, Aix-Marseille University, Aix-en-Provence, France

[4] Avignon University, LIA, Avignon, France

[5] CNRS UMR 5505 IRIT, Toulouse University, Toulouse, France

treatment (surgery and/or radiotherapy and/or chemotherapy). The aim of this paper is to justify the necessity of such a corpus and to present its data collection. This C2SI corpus will be available to the scientific community through the Scientific Interest Group Parolothèque.

Keywords Speech intelligibility and comprehensibility · Quality of life assessment · Speech corpus · Pathological speech

1 Introduction

The decreasing mortality in cancerology brings to light the necessity to reduce the impact of treatments on the quality of life (QoL) after cancer. That particularly concerns head and neck cancers (HNC), because their treatment can be mutilating and disabling. However, the usual tools for assessing QoL are not relevant for measuring the impact of the treatment on the main functions affected by the sequelae. And, there is a clear lack of uniform methods for assessing functional outcomes. Measuring the impact on one or several of the most altered functions after therapeutic care of a given tumoral localization would allow for: (1) completing the expression of the therapeutic outcomes by functional forecast index, (2) adjusting the treatment in order to reduce its functional consequences. For the HNC, it is mainly about impacts of (oral) communication and feeding (swallowing) Mlynarek et al. (2008). QoL research has, to date, failed to provide health care professionals with clinically relevant and interpretable information that can guide treatment decisions. This has led researchers to attempt to make commonly used research tools more accessible to the clinicians. Health-related quality of life (HRQoL) questionnaires reflect the disease impact or functional deficits on general well-being Cardol et al. (1999) by developing questions modules dedicated to the specific consequences of this disease or physiological function. But validated tools to measure the functional outcomes of carcinologic treatment are still missing, in particular for speech disorders. Some assessments are available for voice disorders in laryngeal cancer, but they are based on very poor tools for oral and pharyngeal cancers, dwelling on the articulation of speech rather than on the voice. Given that the usual tools to assess QoL are not relevant to measure the impact of the treatment on the main functions affected by the sequelae, and given that automatic speech processing tools are necessary for unbiased and objective assessments of communication deficiency caused by a speech disorder, we set out to develop a severity index of speech disorders describing the outcomes of therapeutic protocols supplementing the commonly used survival rates. The aim is to perform an audio recording of the patient's speech and to compute the intelligibility of the utterances produced with the aim to get a score. Middag presented a new method that predicts running speech intelligibility in a robust way Middag et al. (2014). This method is text-independent and robust to differences in regional variations of Dutch/Flemish, hence robustly applicable to patients treated for HNC. Therefore, our hypothesis is that an automatic assessment technique can measure the impact of speech disorders on the communication abilities, by giving a severity index of speech for patients

treated for HNC, more particularly for oral and pharyngeal cancers. We will name this index the Carcinologic Speech Severity Index (C2SI). Speech intelligibility is the usual way to quantify the severity of neurologic speech disorders. But this measure is not valid in clinical practice because of several difficulties, such as the familiarity effect experienced by clinicians with their patients' speech disorder, and the poor inter-judge reproducibility. Moreover, the scores do not accurately reflect listeners' comprehension. In order to develop and evaluate this C2SI, a project has been funded from 2014 to 2018 by the French National Cancer Institute (Grant INCa SHS 2014-135) including the following partners: (1) University Hospital Toulouse, (2) LPL laboratory from Aix-Marseille University, (3) Octogone-Lordat from Toulouse University, (4) LIA laboratory from Avignon University, (5) IRIT laboratory from Toulouse University. This C2SI project aims to create a speech corpus and to determine an automatic intelligibility measure. The C2SI corpus is presented in this paper. The structure and the list of tasks performed by each speaker are presented in Sect. 2. Section 3 presents the available material, and some statistics run on the corpus are reported in Sect. 4.

2 Why did we build this corpus?

To cover the broad spectrum of intelligibility and comprehensibility aspects, we wanted to analyze speech distortions at different levels, involving several speaking tasks in order to apply complementary assessment methods. We also needed individual information through quality of life questionnaires.

2.1 Distortions during speech production

2.1.1 Voice signal

In general, low intelligibility is seen as a consequence of poor speech articulation, leading to the belief that there is a weak correlation between voice production and speech intelligibility. However, results from Pyo (2007) indicate that "Patients with severe voice disorders showed very low intelligibility in spite of their intact articulation and prosody." A confirmation can be found in Porcaro et al. (2019). For instance, the capacity to hold a vowel more than 5 s in one breath is a minimal condition for a correct speech production. Recording such a sustained vowel (**AAA**) is a basic task linked to the aerodynamic/acoustic source performance of the speaker. This can also give indications on the speaker's breathing capacity. However, whereas measuring the voice level is really important in the case of laryngeal disease like, for instance, laryngeal cancers, this may not be the case with oral cavity cancers. If the relation "bad voice equal poor intelligibility" seems to be true, the reciprocal is false. Another way to state this is to say that a good voice is a necessary but not sufficient condition for good intelligibility.

2.1.2 Articulation quality

As a first proposal, we can give a definition of intelligibility of a speaker as *the performance by a listener to recognize the words and/or the sounds of the speech produced by the speaker*. We are close to the concept of articulation quality, and the idea is to take into account the accuracy of the phonetic realization. Sadly, intelligibility tests are performed with sentences or words extracted from a restricted list of items. The limitation of this type of test is the ability for listeners to restore the distorted sequences after some time of exposure to the same stimuli. This effect is emphasized when auditors have a strong knowledge of the words used in the test, and when these words are unambiguous and therefore strongly predictable, as in the FDA tests proposed by Enderby (1983) and Enderby and Palmer (2008). These restoration effects are clearly observable in speech-language pathologists who make such an extensive use of these lists that they eventually know them by heart. The bias associated with this knowledge, and therefore with the strong influence of the top-down perceptual mechanisms, results in an overvalued intelligibility score because the phonemic restoration of the listener makes production distortions opaque Warren and Warren (1970) and Samuel (1981). The solution we adopted consists in using large quantities of pseudo-words complying with the frequent phonotactic structures of the speakers native language, in order to completely neutralize the effects of lexicality, familiarization and learning of the items by listeners Ghio (2018). Finally, listeners are confronted with a task that is similar to Acoustic-Phonetic Decoding (**DAP**) followed by a written transcription. The closer the transcription of the pseudo word to the target form, the better its intelligibility. We can make a quick calculation by simply counting the number of correctly recognized phonemes. We can also refine the method by counting the number of different phonetic features between the phonemes of the expected form and those of the transcribed form.

2.1.3 Continuous speech

In order to evaluate speech comprehension, it is important to go beyond the simple tests on isolated words.

(1) We introduced a Sentence Verification Tasks (**SVT**) in order to assess the global comprehension of running speech. In this task, speakers read a set of sentences. The semantic content of each sentence can be true (ex: “January is a winter month”) or false (ex: “January is a summer month”). In the perception evaluation, participants are presented with a variety of utterances across several knowledge domains and have to decide as fast as possible if these statements are true or false Pisoni et al. (1987). The accuracy score and the response time are both used as indicators of the comprehension process. Indeed, when auditors need to understand the linguistic content of a message and perform an appropriate response [True or False], the quality of the

acoustic-phonetic information of the speech signal plays an important role both in the speed and accuracy of the answer provided.

(2) We also used a very common task, whereby speakers had to read a Short Text (**LEC**). This type of spoken communication is very useful because it integrates most of the linguistic levels (phonetic, lexical, syntactic, semantic) in a comparable way between speakers. It makes it possible to produce automatic phonetic alignments, even if the speech production is very altered. Speech rate, prosody, consonant and vowel precision, pauses and other speech features may be easily extracted and compared between the normal and patient groups.

2.1.4 Prosodic specific level

In spoken language, prosodic cues are at the interface of other linguistic levels and fulfill various functions in both message encoding and decoding. Prosody helps structuring utterances by indicating linguistic units' boundaries, thus fulfilling a syntactic function. It also serves the purpose of indicating sentence modality (assertion, interrogation, order...) by precise variations of the intonation contours. Prosodic devices such as focalization (strong accentual marking) are also used to highlight the central information of a message. Beyond these communicative functions, the coherent production of prosodic cues partakes in the fluency of utterances and their temporal organization. Prosody's multiple functions in speech, as well as its interaction with all levels of linguistic structuring, makes it an essential, indispensable feature of speech comprehensibility. Some models describe prosody as a tool for compensatory/palliative strategies to segmental alterations. In other words, speakers would tend to amplify the prosodic cues in their productions to make up for the loss of intelligibility/comprehensibility, may it be due to ambient noise or pronunciation difficulties (theory of speech adaptability, for communication optimization purposes Lindblom (1990)). The patients we focus on in this project have undergone treatment at the supra-laryngeal level of their anatomy (glossectomy, mandibulectomy for example). Theoretically, these treatments are not expected to impact on the production of the laryngeal flow or on prosodic indices, even though some types of treatment may lead to a stiffening of the tissues beyond the treated area and, consequently, to a functional impairment of the larynx. We thus hypothesize that these speakers will obtain satisfactory results in the perceptual assessments with regard to the preservation of prosodic functions despite these peripheral risks: it will be difficult to distinguish between the patient and control groups solely on the basis of their scores, particularly during these tasks where the segmental information is negligible (syntax and modality). However, we believe that the articulatory damages on patients will have an impact on the response time of listeners' perception. Indeed, the alteration of patients' speech should result in increased difficulties of listeners' comprehension. The three prosodic tasks that we propose are designed to evaluate which structural functions of prosody are most affected by these types of cancer:

(1) Modality Function (**MOD**): prosodic marking of assertion, question and injunction, by intonation contour shapes and directions.
(2) Pragmatic Focus (**FOC**): this task required speakers to mark the pragmatic focus by highlighting the important information of an utterance by sole prosodic cues.
(3) Syntactic Disambiguation (**SYN**): speakers had to solve syntactic ambiguity by prosodic means in syntagms composed of two nouns and an adjective, where the adjective either applied to both nouns (high syntactic attachment) or to the last noun (low syntactic attachment).

These tasks are taken from Aura (2012) who adapted Magne et al. (2005) and Astésano et al. (2007) for clinical use. Speakers' capacity to properly use prosodic cues in these different tasks is then used for perceptual evaluation tests on naive, healthy listeners Nocaudie et al. (2018).

2.1.5 Spontaneous speech

In everyday speech, top-down effects are used to decode continuous speech. This is why spontaneous speech is very often used for assessing intelligibility Woisard et al. (2013). In order to reduce speech predictability, we recorded patients and controls in a picture description task (**DES**), as well as in a free task where they had to spontaneously comment on a text they had read just before (**SPO**). Indeed, recording spontaneous speech can also be interesting to assess the comprehensibility of a text. But the evaluation of an index based on these recordings is not easy because semantics may widely vary. However, this task could also be analyzed in order to confirm the other indexes used in the perceptual analyzes.

2.2 Self assessment questionnaires

Self-assessment questionnaires are used in practice to evaluate QoL in its several dimensions. The main generic quality of life questionnaire is the MOS-SF36 Ware and Sherbourne (1992). It is validated in all kinds of illnesses and explores physical as well as mental health disorders. Handicap self-assessment questionnaires were proposed for various functions of the upper aerodigestive tract (UADT). The Speech Handicap Index (SHI) for speech Rinkel et al. (2008) is validated for HNC. The Phonation Handicap Index (PHI) is a similar tool for French, which is however validated for all kinds of speech production disorders Fichaux-Bourin et al. (2009). The relationships between QoL questionnaires and Handicap questionnaires have often been analyzed, with the former being used to validate the contents of latter. Strong correlations (0.7 to 0.9) were computed between the SHI and the speech domain of the QoL questionnaire. This correlation is much lower, if not absent, regarding the other domains Borggreven et al. (2007); Dwivedi et al. (2011) and Thomas et al. (2009). Because using the Handicap questionnaire targeting a specific function is well correlated to the domains of the QoL questionnaires, we selected the generic QoL questionnaire (SF36) and the specific speech related handicap questionnaire (SHI and PHI) in order to integrate the communication dimension.

Table 1 Example of a pseudo words list for DAP

spofo	stoumo	vurtant	muja	teilli	charou
chubra	blania	leba	quermant	neji	jarant
quindu	yainzou	yopta	froubin	finto	joucant
danfin	psitrin	squigu	tichu	ranto	pridi
sonin	gorquin	crazu	zouilla	vougou	grispi
trufru	plirbi	dinli	lanchin	banant	soublou
brussa	cliflu	glepa	zacrin	ruvo	pampou
floniant	drapro	manlo	nemou	nioucou	
poglu	bimpsi	guevant	finsi	nianscou	

3 Speech tasks

3.1 Sustained vowel AAA

These recordings consist in the production of 3 sustained /a/. A sustained vowel gives information about voice level, phonation time, stability, harmonics contents, noise, unvoiced segments, etc.

3.2 Pseudo-words DAP

Each speaker had to pronounce 52 pseudo-words. The pseudo-words have the following phonotactic structure: $C(C)_1V_1C(C)_2V_2$, where $C(C)_i$ was an isolated consonant or a consonant cluster. Such a combinatorial method made it possible to generate around 90,000 pseudo-words. Each list contained the same amount of phonemes in C_1, V_1, C_2 and V_2 position, but in different combinations for each speaker Ghio (2018). Real words have been removed from the dictionary. See Table 1 for a list example with these constraints.

To facilitate the production of pseudo-words by speakers, we used the software Perceval Lancelot.[1] The speaker was placed in front of a screen, and the pseudo-words were automatically displayed while a sound version was produced synchronously. This dual modality, visual and auditory, made it possible to limit reading errors. Given the large size of the corpus (89,346 possible forms), the sound versions was computed using the Voxygen synthesis.[2] The recordings were then segmented semi-automatically, and each pseudo-word was automatically extracted in a separate audio file.

3.3 Sentences SVT

A list of 150 pairs of true/false sentences were created from Pisoni and Dedina (1986), others from Zumbiehl (2010), while the remaining sentence pairs were created by the members of the C2SI project. These sentences have a specific syntactic-semantic structure, whereby the true or false property can be checked only

[1] http://www.lpl-aix.fr/∼lpldev/perceval/.

[2] http://voxygen.fr/.

when the last lexical unit was produced (e.g. “Paris is the capital of France” vs. “Paris is the capital of Germany”). Consequently, it is necessary to decode and understand the whole sentence before coming up with the answer.

A set of 50 sentences selected from the list of 300 sentences was produced by each speaker.

3.4 Text LEC

The first paragraph of “La chèvre de M. Seguin”, a tale by Alphonse Daudet, was read by the speakers. This text was chosen because it is long enough and it encompasses all French phonemes. It is also well known and widespread in clinical phonetics in France Ghio et al. (2012). Here is the full plain text: “*Monsieur Seguin n’avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s’en allaient dans la montagne, et là-haut le loup les mangeait. Ni les caresses de leur maitre, ni la peur du loup rien ne les retenait. C’était parait-il des chèvres indépendantes voulant à tout prix le grand air et la liberté.*”

3.5 Prosodic tasks

*3.5.1 Modality function **MOD***

The modality task consists in the production of ten identical sentences with 3 different modalities: assertion, question and injunction (*You eat pastas ?/./!*). Each speaker recorded 10 different scripts uttered with the 3 modalities. Each script was presented on a computer screen, with the expected prosodic modality indicated by either of the 3 punctuation marks (’.’ ’?’ ’!’).

*3.5.2 Focus function **FOC***

In the focus task Aura (2012), speakers had to resolve a paradigmatic opposition (contrastive focus) between two words given in an auditorily presented sentence so as to prosodically highlight the relevant word (“*Did you see a duck or a pig in the garden?*” with the written answer: “*I saw a DUCK in the garden*”). Each speaker recorded the same set of 20 sentences, for which they had to produce the proper focus as scripted, following the audio presentation of the question.

*3.5.3 Syntactic function **SYN***

The syntactic task Aura (2012) and Astésano et al. (2007) consists of similar written scripts that only prosody can disambiguate. For example, in the sentence “*les chevaux et les poneys blancs*” (eg. “*White horses and poneys*” : note that the adjective in French is at the end of the sentence), the adjective “*blancs*” (eg. “*white*”) can either apply to the second noun only (narrow scope) or to the two nouns (broad scope): prosodic cues such as final lengthening, pause and f0

excursions can give the proper syntactic parsing (either "*les chevaux//et les poneys blancs*" or "*les chevaux et les poneys//blancs*").

Each speaker recorded 13 scripts with two syntactic conditions (narrow vs. broad scope of adjective). The sentences were written on a computer screen, with the expected syntactic grouping indicated visually by vertical bars.

3.6 Picture description DES

The subject was asked to choose one amoung several pictures representing a similar scenery (the sea with boats). Each subject had to describe the picture to the examiner so that the latter could redraw it just on the basis of the oral explanations.

3.7 Spontaneous speech SPO

The patient had to give his/her opinion on the questionnaire that he/she has to fill out before the recording session. He/she had to speak for at least 3 min. This task allows us to collect spontaneous speech recordings with no constraint on the sentences.

4 Corpus description

4.1 Population

The number of patients to recruit in this database was estimated statistically on the following constraints: We expected the correlation between the automatic index and the perceived index given by the jury to be as high as 0.86 correlation, similar to the one achieved in the work done by the University of Ghent Middag et al. (2014).

The size of the sample influences the precision of this estimation, a bigger sample bringing a bigger precision (characterized by a narrower confidence interval). To obtain a 95% confidence interval, the width of which is not superior to 0.15 around a coefficient of 0.8, it is necessary to recruit 94 patients. In September 2017, we recorded 134 sessions, that represents 87 patients and 40 control speakers. 7 patients were recorded twice. That is superior to the corpus used in Middag et al. (2014), which contained recordings and perceptual evaluations of 55 patients with advanced Head and Neck Cancer who were treated with concomitant chemoradiotherapy. The patients were recruited in the three main departments of Toulouse managing patients with HNC (ENT department of the University Hospital, Cancerology department of the Institut Claudius Regaud (surgery and radiotherapy), Maxillofacial surgery department of the Toulouse University Hospital). They were selected from the lists of carcinologic follow-up consultations of these 3 departments. These departments are part of the University Institute of Cancer in Toulouse (IUC-T) and associated with the unit of "Onco-réhabilitation" which is located at the IUC-T Oncopole. These patients had to meet the following inclusion criteria:

- have a T1 to T4 cancer of the oral cavity and/or pharynx;
- have been treated by surgery and/or radiotherapy and/or chemotherapy;

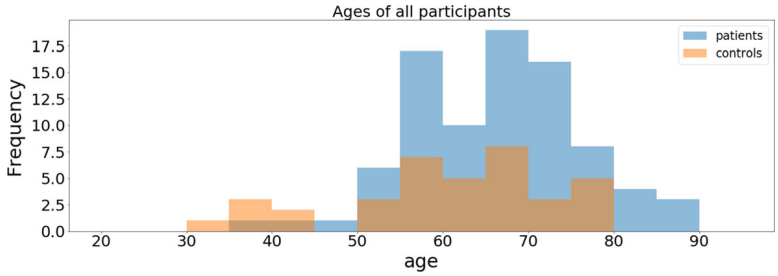


Fig. 1 Age distribution of the patients and controls groups

– be more than 6 months after the end of treatment to ensure stability of the speech disorder, whether audible or not.

Similarly, the criteria for non-inclusion were to present another source of speech disorders (eg. stuttering) or to present cognitive or visual problems that are incompatible with the assessment protocol design. These non-inclusion criteria were also used for the recruitment of the control population. Among the patients, 51 (59%) were men, and the mean age was 65.8 years old (range 36–87).

40 healthy controls (HC) were recruited. 18 control speakers (45%) were men. Figure 1 presents the age distribution of patients and controls. The control group's mean age was significantly different from the patients (56.9 years old, range 35–79, $p = 0.003$ Mann–Whitney).

4.2 Metadata

Individual metadata were collected for each speaker. They comprise civility information such as age, gender, birth and area of residence (French department), as well as clinical information including the anatomical region affected by the cancerous lesion, values of T and N criteria from UICC Tumor/Node/Metastasis (TNM) classification Brierley et al. (2016) (the internationally accepted standard for cancer staging by the UICC journals), treatment type (surgery, radiotherapy, chemotherapy), time in months since the end of treatment.

The research protocol was reviewed by the Research Ethics Committee[3] (CER) from the University Hospital Centre of Toulouse. CER analyses ethical aspects of research protocols directly or indirectly involving humans. They approved the C2SI protocol on May 17th, 2016. A processing declaration which purpose is "the recording of the speech of patients treated for ENT cancer" was registered with the Commission Nationale de l'Informatique et des Libertés (CNIL) on July 24th, 2015 under number 1876994v0.

[3] https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer.

Fig. 2 Photography of the anechoic room for audio recordings

4.3 Questionnaires

The SHI and PHI health related quality of life questionnaires presented in Sect. 2.2 are given to the patients just before the audio recordings. The SHI is composed of 30 questions shared between two dimensions (symptoms and psycho-social). The PHI is a 15 items questionnaire with 3 dimensions (symptoms, functional consequences and emotional).

4.4 Recordings

The speakers were comfortably seated in an anechoic room in front of a computer (see Fig. 2). The computer was used to visually display instructions and corpus. For some tasks, the instructions were also produced with an auditory modality (ex: pseudo-words in DAP task). The recordings were made with a Neumann TLM 102 Cardioid Condenser Microphone connected to a FOSTEX digital recorder. The sampling rate was 48 kHz, which facilitates the downsampling to 16 kHz, usually used in automatic speech processing.

87 patients and 40 control speakers were recorded in the corpus. Unfortunately, despite a similar generic recording protocol, some patients and control speakers did not carry out all the tasks. Table 2 provides detailed information related to the tasks such as the number of speakers performing each task, the mean duration per recording as well as their total duration. It can be pointed out that 5 patients were recorded twice, performing all the protocol tasks. Similarly, 2 patients performed all the tasks except the spontaneous task, and a final patient was recorded twice on the AAA, LEC and DES tasks only. Regarding now the medical information, the inclusion criteria were balanced regarding tumor localization (see Fig. 3): 39% of oral cavity cancer (Floor of mouth, Tongue, Retromolar Area and Mandibula), and

Table 2 Information about speakers and tasks: number of speakers having carried out a given task (#FC: female control speakers-#FP: female patients-#MC: male control speakers-#MP: male patients), mean duration of recordings per task, total duration per task

Task	Nb of speakers #FC-#FP-#MC-#MP	Mean duration per recording (s)	Total duration (s)
Questionnaires			
SHI	6-30- 5-50	–	–
PHI	6-29 -5-48	–	–
SF36	12-35-8-49	–	–
Speechtasks			
AAA	16-35-10-48	8.6	2978
DAP	21-36-18-44	188.6	24134
SVT	20-34-18-48	207.4	25920
LEC	21-33-18-47	33.1	3768
MOD	22-33-18-47	141.2	17645
FOC	22-33-18-46	192.7	25052
SYN	22-33-18-44	167.1	19889
DES	14-35-10-46	69.6	9397
SPO	15-34-10-43	66.0	7064

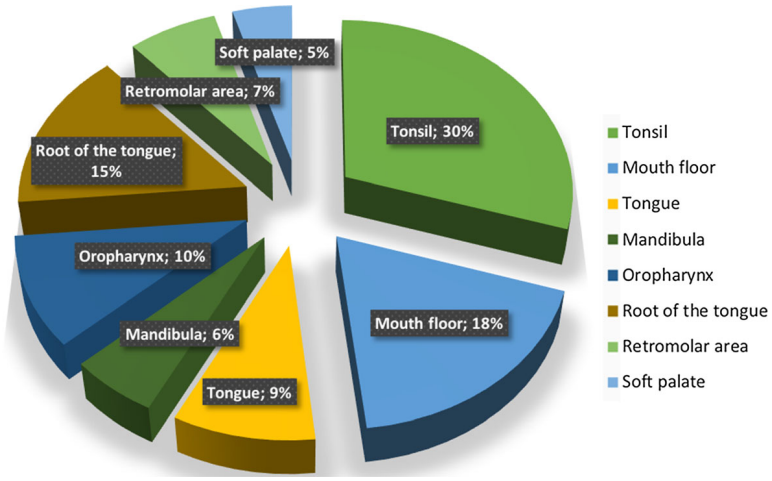


Fig. 3 Tumor localization distribution

61% of oropharyngeal cancer (Tonsil, Root of tongue, Soft Palate and when there is a larger extension "OroPharynx").

Figure 4 presents the treatment distribution of patients. The most frequent treatment related to the size of the tumors is surgery (84%). The resection of the

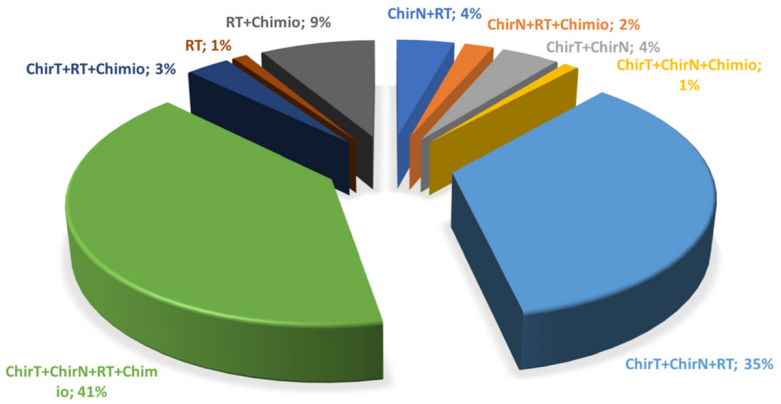


Fig. 4 Patients treatment distribution

tumor (ChirT) is associated with the node resection (ChirN) followed in 40% by a chemoradiotherapy (RT-chimio) and in 37% by radiotherapy (RT) only.

5 Enrichment data

5.1 Human perception evaluation

For the data DAP, MOD, SYN, FOC and SVT, the set of stimuli of all speakers was played back to a set of listeners via the Perceval Lancelot software[4]. We collected a large amount of perceptual data, allowing us to issue quantified indicators related to our data.

– DAP:All the 52 pseudo-words pronounced by every speaker of the database were transcribed 3 times. 40 naive listeners were involved in order to transcribe the 52 * 119 speakers = 6188 stimuli. Listeners were confronted with a task that can be considered as acoustic-phonetic decoding followed by a written transcription. The mean distance between the transcribed and expected response is considered as a score of (un)intelligibility. For the comparison operation, we used a Wagner–Fischer algorithm that integrates the phenomena of insertion, elision and substitution of units. In our case, this calculation of Levenshtein distance is not based on orthographic units but on phonemes Ghio (2018). Indeed, on the orthographic forms, in a traditional way, the distance between 2 graphemes is null if they are equal and is equal to 1 if they are different. In the case of phonemes, it is possible to introduce more subtle nuances because, for example, we can consider that a confusion between two vowels does not have the same weight as between a vowel and a voiceless consonant. We used the phonetic features theory to establish the local distance. In our preliminary

[4] http://www.lpl-aix.fr/∼lpldev/perceval/.

results, we validated this measure which is discriminant between healthy speakers vs. patients. The measure called Perceived Phonological Deviation (PPD) is strongly correlated with the clinical severity index. Moreover, experiments show that the method is not biased by a learning effect by the listeners.

In order to perceptually evaluate 3 times each recorded sentence for SVT (4816 sentences), FOC (1969 sentences, MOD (2860 sentences) and SYN (2513 sentences) tasks, the recordings were presented to 147 naive listeners:

- MOD:the recordings were presented to naive listeners, who had to recognize which modality was intended, between assertion, question and injunction Nocaudie et al. (2018).
- FOC:Each sentence previously recorded was thereafter associated with a congruous ("*Qu'as-tu vu dans le jardin, un cochon ou un canard ?*" / eg. "*What did you see in the garden, a pig or a duck?*") or incongruous ("*Où as-tu vu un canard, dans le jardin ou dans la cour ?*" eg. "*Where did you see a duck in the garden or in the yard?*") question. Listeners had to judge whether the perceived focus was congruous or incongruous in the manipulated dialogues.
- SYN:Each recorded sentence was presented to naive listeners who had to choose between two pictures representing either one or the other syntactic reading (narrow *vs.* broad scope of adjective).
- SVT:The sentences were evaluated by 3 naive listeners who had to judge whether the sentence presents a true fact or an incorrect one. This produces an indicator based of the global comprehensibility of the sentence recorded.

A perception score was thereafter calculated for each speaker*task (SYN, FOC, MOD and SVT). These scores ranging from 0 to 3 correspond to the mean of each perceptual evaluation obtained during the test. Mean listener's reaction times were also calculated. These perceptual scores were correlated to an index of severity (estimated by 5 healthcare professionals) as reported in Nocaudie et al. (2018). It appears that the mean SVT score is strongly correlated to this index of severity ($r = .81$, $p < .001$) and thus can serve as a rating of the global comprehensibility of speakers whereas prosodic tasks' mean scores are moderately correlated to the severity index (FOC: $r = .56$, $p < .001$, MOD: $r = .44$, $p < .05$; SYN: $r = .53$, $p < .001$), indicating that tested prosodic functions seem overall preserved in the patients' speech.

- LEC and DES:With the data obtained on read and spontaneous speech, an index of severity (alteration of speech signal) and subjective intelligibility was produced by a set of six experts on a scale from 0 (the strongest alteration) to 10 (perfect speech). In order to assess for inter-judge reliability, an Interclass Correlation Coefficient (ICC) was calculated. The degree of concordance between the jury ratings is therefore good ($r > 0.69$) for the set of tasks. The jury constituted by the six speech-language expert jurors is therefore homogeneous.

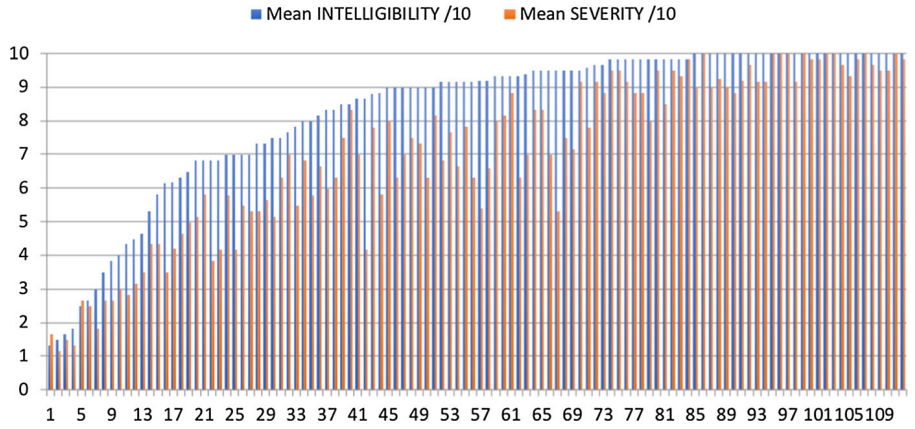


Fig. 5 Distribution of intelligibility and severity scores on DES task in ordinate and subjects (by increased scores of intelligibility) in abscissa

Although these different tasks give highly correlated results ($r > 0.8$), the intent to evaluate speech severity (alteration of the speech signal) favors a score distribution offering a better metric. The severity is on average more impacted by an oral location (5.44, sd 2.47) than oropharyngeal (6.46, sd 2.24). The semi-spontaneous speech tends to reduce the ceiling effect of the severity measure compared to the speech read (average scores at 6.06 over 10 in image description, and 6.51 over 10 in reading).

Perceptive measurement of the severity of speech disorder on semi-spontaneous speech seems to be the clinically most relevant score in the evaluation of speech disorders after cancer treatment of the oral cavity or oropharynx. More details on this perceptual task could be found in Balaguer et al. (2019). Figure 5 presents the distribution of intelligibility and severity scores across subjects on DES task. The figure is sorted by increased scores of intelligibility.

5.2 Automatic processing

In order to enrich the C2SI corpus, notably for phonetic analysis related to speech disorder analysis, audio recordings related to the LEC task was segmented automatically at the phone level thanks to a forced-alignment system. The latter takes as inputs the sequence of words pronounced in a speech utterance, and a phonetized lexicon of words coupled with different phonological variants, based on a set of 37 French phones. The forced alignment is based on a Viterbi decoding and graph-search algorithms, the core of which is the acoustic modeling of each phone, based on Hidden Markov Models (HMM). A 3-state context-independent HMM topology is used to model each phone. The HMM-based models are built thanks to the Maximum Likelihood Estimate paradigm from about 200 h of French radiophonic speech recordings Galliano et al. (2005). These models are speaker independent, however a three-iteration MAP adaptation is applied to all the HMM

parameters to get speaker-dependent models. Acoustic vectors consist of 12 Perceptual Linear Prediction coefficients plus the energy, plus their delta and delta-delta coefficients.

It is important to note that the input sequences of words come from the original text that speakers had to read for the LEC task. Indeed, no manual orthographic transcription per recording had been performed by human listeners. Therefore, alignment errors could be possible due to word repetitions, omissions, substitutions, or deletions. Nevertheless, depending on the targeted phonetic analysis, a manual phone segmentation correction could be envisaged from the automatic outputs, which may bring a large gain of time.

Thus, this corpus enrichment results in one pair of start and end boundaries per phone present in all the speech recordings associated with the LEC task, packaged into TextGrid format files.

6 Conclusions and future work

In this paper, we have presented the design and recording of a corpus of 127 speakers, which allows us to consider the automatic production of indexes with a high level of correlation. During the constitution of the corpus, we faced several issues. Considering DAP task, patients' recordings were initially achieved, using only a visual presentation of the DAP items and the pseudo-word was simultaneously read aloud by the experimenter. However, because the phonological construction of the items sometimes allows different possible pronunciations, this configuration could have modified the speaker's repetition. To cope with this drawback, we replaced the aloud reading of the experimenter with a recorded synthesized voice for each item to standardize its pronunciation and to limit the potential biases. Furthermore, some tasks were considered as particularly hard to understand and to achieve by the patients (SYN, for example): the impact of these perceived difficulties will have to be checked and studied during the analysis of the results. Perceptual evaluations are in progress in order to complete the usable metadata, and to obtain reliable intelligibility/comprehensibility scores, which will be compared to self-assessed quality of life scores. We are also working now on extracting information from the different recordings in order to analyze them and to produce automatic indexes Sicard et al. (2017), Laaridh et al. (2017, 2018) and Fredouille et al. (2019).

Our main goal is to get objective judgments, which can help speech therapists and physicians in clinical practice. Data will be available to the scientific community through the GIS Parolotheque[5]: a scientific structure ("Groupement d'Intéret Scientifique") which purpose is to facilitate access and research on pathological speech recordings (like the tumor library "thomorothèque" for access to cancer cell samples). The data come from hospital structures in a pseudo-anonymized form (waveforms cannot be totally anonymized by definition, but all metadata from the hospital are cleaned). The GIS is responsible for the storage, legal aspect and allocation of access to scientists in the context of research projects.

[5] https://www.parolotheque.fr.

Access to data and metadata is therefore facilitated and accelerated compared to the traditional approaches that are currently required to participate in this kind of research. The GIS is currently in a signing phase and operational implementation is expected to start in 2020.[6]

Acknowledgements Grant 2014-135 from Institut National pour le CAncer (INCa) in 2014, "Sciences Humaines et Sociales, épidémiologie et Santé Publique" call. Lead by Pr Virginie Woisard at University Hospital of Toulouse and Grant ANR-18-CE45-0008 from The French National Research Agency in 2018 RUGBI project "Improving the measurement of intelligibility of pathological production disorders impaired speech" lead by Jérôme Farinas at IRIT. We thank the company Voxygen[1] for providing us with their speech synthesis platform necessary for the realization of the corpus DAP.

References

Astésano, C., Bard, E. G., & Turk, A. (2007). Structural influences on initial accent placement in french. *Language and Speech*, *50*(3), 423–446.

Aura, K. (2012). Protocole d'évaluation du langage fondé sur le traitement de fonctions prosodiques : étude exploratoire de deux patients atteints de gliomes de bas grade en contexte péri-opératoire. Ph.D. thesis, Université Toulouse 2. http://www.theses.fr/2012TOU20110/document.

Balaguer, M., Boisguerin, A., Galtier, A., Gaillard, N., Puech, M., & Woisard, V. (2019). Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *Annales fran caises d'oto-rhino-laryngologie et de pathologie cervico-faciale*, *136*(5), 355–359. https://doi.org/10.1016/j.anorl.2019.05.012.

Borggreven, P. A., Aaronson, N. K., Verdonck-de Leeuw, I. M., Muller, M. J., Heiligers, M. L., & de Bree, R., et al. (2007). Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. *Oral oncology*, *43*(10), 1034–1042.

Brierley, J. D., Gospodarowicz, M. K., & Wittekind, C. (2016). *TNM classification of malignant tumours*. Hoboken: Wiley.

Cardol, M., Brandsma, J., De Groot, I., van den BOSOE, G., De Haan, R., & De Jong, B. (1999). Handicap questionnaires: what do they assess? *Disability and rehabilitation*, *21*(3), 97–105.

Dwivedi, R. C., St Rose, Rose, Roe, J. W., Chisholm, E., Elmiyeh, B., & Nutting, C. M., et al. (2011). First report on the reliability and validity of speech handicap index in native english-speaking patients with head and neck cancer. *Head & neck*, *33*(3), 341–348.

Enderby, P.M. (1983). Frenchay dysarthria assessment. Pro-ed

Enderby, P.M., & Palmer, R. (2008) FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual. Pro-ed

Fichaux-Bourin, P., Woisard, V., Grand, S., Puech, M., & Bodin, S. (2009). Validation of a self assessment for speech disorders (phonation handicap index). *Revue de laryngologie-otologie-rhinologie*, *130*(1), 45–51.

Fredouille, C., Ghio, A., Laaridh, I., Lalain, M., & Woisard, V. (2019). Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. In: Proceedings of Intl Congress of Phonetic Sciences (ICPhS'19). Melbourne, Australia

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., & Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In: Ninth European Conference on Speech Communication and Technology

Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Robert, D., Rebourg, M., Fredouille, C., Laaridh, I., & Woisard, V. (2018).Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In: XXXII éme Journées d'Etudes sur la Parole 10.21437/JEP.2018-33. https://hal.archives-ouvertes.fr/hal-01770161/file/190996.pdf.

Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., & De Looze, C., et al. (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication*, *54*(5), 664–679.

[6] http://voxygen.fr/.

Laaridh, I., Fredouille, C., Ghio, A., Lalain, M., & Woisard, V. (2018). Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers. In: Interspeech, pp. 2943–2947. ISCA, Hyderabad, India. 10.21437/interspeech.2018-1266. https://hal.archives-ouvertes.fr/hal-01962170.

Laaridh, I., Kheder, W.B., Fredouille, C., & Meunier, C. (2017). Automatic prediction of speech evaluation metrics for dysarthric speech. In: Proc. Interspeech, pp. 1834–1838

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In: Speech production and speech modelling, vol. 55, pp. 403–439. Springer, Dordrecht https://doi.org/10.1007/978-94-009-2037-8_16

Magne, C., Astésano, C., Lacheret-Dujour, A., Morel, M., Alter, K., & Besson, M. (2005). On-line processing of "pop-out" words in spoken french dialogues. *Journal of cognitive neuroscience*, *17*(5), 740–756.

Middag, C., Clapham, R., Van Son, R., & Martens, J. P. (2014). Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer speech & language*, *28*(2), 467–482.

Mlynarek, A. M., Rieger, J. M., Harris, J. R., O'Connell, D. A., Al-Qahtani, K. H., & Ansari, K., et al. (2008). Methods of functional outcomes assessment following treatment of oral and oropharyngeal cancer: Review of the literature. *Journal of otolaryngology - head and neck surgery*, *37*(1), 2–10.

Nocaudie, O., Astésano, C., Ghio, A., Lalain, M., & Woisard, V. (2018) .Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx. In: XXXIIe Journées d'Etudes sur la Parole, pp. 196–204

Pisoni, D.B., & Dedina, M.J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task: a first report. Tech. Rep. Progress report 12, Indiana University

Pisoni, D. B., Manous, L. M., & Dedina, M. J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer speech & language*, *2*(3–4), 303–320.

Porcaro, C., Evitts, P., King, N., Hood, C., Campbell, E., & White, L., et al. (2019). Effect of dysphonia and cognitive-perceptual listener strategies on speech intelligibility. *Journal of Voice in Press*,. https://doi.org/10.1016/j.jvoice.2019.03.013.

Pyo HY, S.H.S. (2007). A study of speech intelligibility affected by voice quality degradation. Communication Sciences & Disorders, 12(2), 256–278 http://www.e-csd.org/journal/view.php?number=326.

Rinkel, R. N., Leeuw, I. M. V., van Reij, E. J., Aaronson, N. K., & Leemans, C. R. (2008). Speech handicap index in patients with oral and pharyngeal cancer: Better understanding of patients' complaints. *Journal for the Sciences and Specialties of the Head and Neck*, *30*(7), 868–874.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474.

Sicard, E., Mauclair, J., & Woisard, V. (2017). Etude de paramètres acoustiques des voix de patients traités pour un cancer orl dans le cadre du projet c2si. In: 7èmes Journées de Phonétique Clinique

Thomas, L., Jones, T. M., Tandon, S., Carding, P., Lowe, D., & Rogers, S. (2009). Speech and voice outcomes in oropharyngeal cancer and evaluation of the university of washington quality of life speech domain. *Clinical Otolaryngology*, *34*(1), 34–42.

Ware, J. E, Jr., & Sherbourne, C. D. (1992). The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, *223*(6), 30–37.

Woisard, V., Espesser, R., Ghio, A., & Duez, D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique? *Revue de Laryngologie Otologie Rhinologie*, *1*(134), 27–33.

Zumbiehl, O. (2010). Evaluation perceptive des dysphonies par la sentence verification task. Master's thesis, Université Aix-Marseille . Mémoire d'Orthophonie (dir. : Cavé, C. and Ghio, Alain)

A.4 Article dans le journal IJLCD

L'article intitulé « Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer : Preliminary validation » a été publié le 31 août 2022 dans International Journal of Language and Communication disorders [Balaguer et al., 2023b]. Cet article représente une partie des résultats obtenus suite au doctorat de Mathieu Balaguer : la proposition de la réalisation d'un score montrant l'impact de la difficulté de communication suite à un cancer ORL sur les cercles sociaux.

Received: 11 February 2022 | Accepted: 8 July 2022

DOI: 10.1111/1460-6984.12766

International Journal of Language & Communication Disorders

RESEARCH REPORT

Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation

Mathieu Balaguer[1,2] | **Julien Pinquier**[1] | **Jérôme Farinas**[1] | **Virginie Woisard**[2,3]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France

[2]Hôpital Larrey, Hôpitaux de Toulouse, Toulouse, France

[3]Laboratoire de Neuro-Psycho-Linguistique LNPL, Université Toulouse II, Toulouse, France

Correspondence
Mathieu Balaguer, IRIT Institut de Recherche en Informatique de Toulouse, SAMoVA Team, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France.
Email: mathieu.balaguer@irit.fr

Funding information
This study was funded by the Hospitals of Toulouse; and by the French National Research Agency [RUGBI project, grant ANR-18-CE45-0008].

Abstract

Background: In head and neck cancer, many tools exist to measure speech impairment, but few evaluate the impact on communication abilities. Some self-administered questionnaires are available to assess general activity limitations including communication. Others are not validated in oncology. These different tools result in scores that does not provide an accurate measure of the communication limitations perceived by the patients.

Aim: To develop a holistic score measuring the functional impact of speech disorders on communication in patients treated for oral or oropharyngeal cancer, in two steps: its construction and its validation.

Methods & Procedures: Patients treated for oral/oropharyngeal cancer filled six self-questionnaires: two about communicative dynamics (ECVB and DIP), two assessing speech function (PHI and CHI) and two relating to quality of life (EORTC QLQ-C30 and EORTC QLQ-H&N35). A total of 174 items were initially collected. A dimensionality reduction methodology was then applied. Face validity analysis led to eliminate non-relevant items by surveying a panel of nine experts from communication-related disciplines (linguistics, medicine, speech pathology, computer science). Construct validity analysis led to eliminate redundant and insufficiently variable items. Finally, the holistic communication score was elaborated by principal component factor and validated using cross-validation and latent profile analysis.

Outcomes & Results: A total of 25 patients filled the questionnaires (median age = 67 years, EIQ = 12; 15 men, 10 women; oral cavity = 14, oropharynx = 10, two locations = 1). After face validity analysis, 44 items were retained ($\kappa >$ 0.80). Four additional items were excluded because of a very high correlation ($r > 0.90$) with other items presenting a better dispersion. A total of 40 items were finally included in the factor analysis. A post-analysis score prediction was performed (mean = 100; SD = 10). A total of 24 items are finally retained for the construction of the holistic communication score (HoCoS): 19 items from

questionnaires assessing communicative dynamics (13 from the ECVB and six from the DIP), four items from a perceived speech impairment questionnaire (PHI) and one from a quality-of-life questionnaire (EORTC QLQ-H&N35). The reliability is good (five-fold cross-validation: $r_s = 0.91$) and the complementary latent profile analysis shows a good validity of the HoCoS, clustering subjects by level of communication performance.

Conclusions & Implications: A global score allowing a measure of the impact of the speech disorder on communication was developed. It fills the lack of this type of score in head and neck oncology and allows the better understanding of the functional and psychosocial consequences of the pathology in the patients' follow-up.

KEYWORDS
assessment, communication, oncology, speech

What this paper adds

What is already known on the subject

- Because of their anatomical location, head and neck cancer degrades the speech abilities. Few tools currently allow the assessment of the impact of the speech disorder on communication abilities. In ENT oncology, self-administered questionnaires are available to assess activity limitations and participation restrictions (International Classification of Functioning (ICF)—WHO). Other tools from the field of neurology allow an evaluation of communication dynamics. But these different tools, constructed by items, give global additive or averaged scores. This implies an identical weighting of each item, resulting in global scores that are not very representative of the communication limitations really perceived by the patients.

What this paper adds to existing knowledge

- A new global holistic score allowing a measurement of the impact of speech impairment on communication after treatment of oral or oropharyngeal cancer has been developed. The methodology of its construction allows a better reflection of the symptomatological, pragmatic and psychosocial elements leading to a degradation of communication abilities.

What are the potential or actual clinical implications of this work?

- The developed HoCoS score fills the gap in the absence of this type of tool in head and neck oncology. It may allow a better understanding of the factors involved in the functional and psychosocial limitations of these patients, and better customize their follow-up.

INTRODUCTION

Cancer of the oral cavity and oropharynx is very common, with a high incidence: they represent more than 475,000 new cases worldwide in 2020.[1] At the same time, mortality from lip–mouth–pharynx cancer is decreasing. Thus, the increase in life expectancy following cancer means that patients are now living longer with the

sequelae of cancer and treatment (Borggreven et al., 2007).

In this context, the functional and psychosocial repercussions after oncological treatment must be considered, alongside the analytical and dynamic deficits.

For this purpose, various conceptual frameworks, based on bio-psychosocial models, have been developed. The International Classification of Functioning, Disability and Health (ICF) developed by the World Health Organization (WHO) (2001) suggests looking beyond the impairment to the functional (activity limitations) and psychosocial consequences (participation restrictions) of pathologies. In this classification, personal and environmental factors can impact functional and psychosocial levels. These models provide a better understanding of the impact of therapeutic procedures in head and neck cancer on patients' quality of life (Borggreven et al., 2007).

Complementary conceptual models have been described to specify the different levels involved in the functional and psychosocial dynamics, and to establish the causal relationships that may exist between them (Wilson & Cleary, 1995): biological and physiological factors, symptomatic status, functional status, general health perceptions and overall quality of life (Murphy et al., 2007).

Due to their location, oral and oropharyngeal cancer impact the speech abilities (Balaguer et al., 2019, 2020; Mlynarek et al., 2008), and are a common complaint of these patients. This symptomatology influences functional status, altering the patients' communication abilities (Eadie et al., 2018). Yet, while many perceptual and automatic tools currently exist in head and neck oncology to measure speech impairment (Middag et al., 2009; Woisard et al., 2021) few assess the functional impact on communication abilities (Bolt et al., 2016; Meyer et al., 2004).

Some questionnaires assess activity limitations and participation restrictions, such as the phonation handicap index (PHI) (Balaguer et al., 2020) or the 'phonation' or 'psychosocial' domains of the carcinologic handicap index (CHI) (Balaguer et al., 2021). Other items related to communication, social contact or speech are present in both modules of the European Organization for Research and Treatment of Cancer (EORTC) quality of life (QoL) questionnaires (Aaronson et al., 1993; Bjordal et al., 1999).

Other questionnaires assess communication function but are not validated in head and neck oncology, such as the ECVB (Échelle de Communication Verbale de Bordeaux) (Mazaux et al., 2006) and the dysarthria impact profile (DIP) (Letanneux et al., 2013; Walshe et al., 2009). Some others target the assessment of communication abilities but present a very short format limiting the comprehensiveness of the assessment of the communication situation, for example, the Communicative Participation Item Bank (CPIB) (Baylor et al., 2009).

Moreover, the scoring of communication impairment is a crucial matter in the development of these tools. Indeed, these questionnaires result in scores per item, grouped in global scores. Because of their construction, by addition or average, these scores and subscores relating to a functional dimension make the hypothesis that each item carries the same weight in the construction of the final communication score. However, this strong assumption is difficult to support in a clinical setting because of the specific perceptions of each patient of their communication capacity. Moreover, all these questionnaires target different aspects of communication, but none of the tools allows to obtain a global, holistic score, representative of the impact on communication of speech disorders in head and neck oncology.

The objective of the present study is to develop a holistic communication score measuring the functional impact of speech disorders on communication in patients treated for oral or oropharyngeal cancer. The development of this score includes two sub-objectives: its construction and its validation.

METHODS

Design

This is a cross-sectional observational study.

The study protocol was approved by the Committee for the Protection of Persons (CPP: Ouest IV, 19 February 2020, reference 11/20_3) within the framework of the ANR RUGBI project.[2]

Participants

Patients coming for consultation or hospitalization in an ear, nose and throat (ENT) service were recruited by the medical staff.

Inclusion criteria were: being of legal age (at least 18 years old) and having been treated for cancer of the oral cavity or oropharynx (surgical treatment and/or radiotherapy and/or chemotherapy) for at least 6 months (stable disorders). Patients with any other associated chronic disease were excluded.

All subjects who could be included during the inclusion period (October 2019–December 2020) were asked to participate in this study. The inclusion period corresponds to the inclusion period of the quality-of-life work package of the main study (RUGBI project).

Selection process of questionnaires

Communication-related questionnaires

Two questionnaires relating to communicative dynamics were retained in their entirety because of their comprehensiveness construction and their conceptual proximity to our objective of measuring the alteration of communicative abilities: the ECVB (Mazaux et al., 2006) and the DIP (Letanneux et al., 2013; Walshe et al., 2009). These questionnaires are only validated in an adult population with neurological pathologies. Nevertheless, they were retained in the constitution of the corpus of this study because they allow an ecological measurement (i.e., close to the real situations of real daily life) of communication, while taking into account the psycho-affective dimension.

The ECVB (Mazaux et al., 2006) includes 34 items divided into seven dimensions corresponding to daily communication situations: expression of intentions (three items), conversation (seven items), telephone (seven items), shopping (four items), social relations (five items), reading (four items) and writing (four items). Initially validated with stroke patients, the dimensions assessed concern all aspects of communication that may be impaired, whether oral or written. Each item is rated on the principle of a Likert scale with four levels ('never', 'rarely', 'often' and 'very often', from 0 to 3 for questions with a positive polarity, and from 3 to 0 for negative ones). The lower the scores, the greater is the discomfort. In this study, this questionnaire was filled by the patient himself, as it is usually done in the patients reported outcomes (PRO) questionnaires (Doward & McKenna, 2004).

The DIP, initially validated in English (Walshe et al., 2009), has been translated and validated in French (Letanneux et al., 2013). Intended for subjects with Parkinson's disease, it includes 48 items in four dimensions: 'the effect of dysarthria on me as a person' (12 items), 'accepting my dysarthria' (10 items), 'how I feel others react to my speech' (14 items), and 'how dysarthria affects my communication with others' (12 items). These 48 items are also presented in the form of a five-level Likert scale (strongly disagree, agree, not sure, disagree, strongly disagree: the direction of the scoring depending on the polarity of the questions). For this study, we added an item to section 'how I feel others react to my speech' to clarify an abstruse French formulation. The DIP therefore includes here 49 items.

Questionnaires assessing speech function

Some questionnaires used in the routine care include items relating to the functional or psychosocial consequences of the speech disorder in head and neck cancer: the phonation handicap index (PHI) (Balaguer et al., 2020) and the Carcinologic handicap index (CHI) (Balaguer et al., 2021).

Focusing on speech self-assessment, the PHI was retained in its entirety: five items in the physical signs (PHI-F) domain, five items in the functional impact (PHI-C) domain, five items of the psychosocial repercussions domain (PHI-E), and three complementary questions (What degree of severity do you give to your speech difficulties?; How difficult is it for you to produce understandable speech?; and How much does your speech impairment affect your daily life?').

Allowing a measure of patients' needs after cancer treatment, the CHI also includes dimensions and items related to speech: four items in the dedicated phonation domain, and four items of the psychosocial domain.

Quality of life (QoL) questionnaires

Finally, other questionnaires are designed to measure psychosocial consequences in terms of impact on QoL. The EORTC reference questionnaires were retained for further analyses: 30 items of the EORTC QLQ-C30 (Aaronson et al., 1993) and 35 of the EORTC QLQ-H&N35 (Bjordal et al., 1999) measuring global and speech-related quality of life.

Construction of the score

Characterization of the variable type for the holistic communication score (HoCoS)

The determination of the type of the targeted variable is essential because it will condition the statistical methodology used in dimensionality reduction (Carreira-Perpinán, 1997; Chadeau-Hyam et al., 2013; Cunningham, 2014). This methodology is required due to the format of the data, where a large number of items ($n = 174$) was retained.

The holistic communication score (HoCoS) was considered as a latent variable (Borsboom, 2008): it cannot be directly observed or measured, and it requires several manifest variables as indicators that are observable and measurable. In our study, the latent HoCoS influences the values of the measured manifest variables, that is, the items from the different questionnaires. Although the responses to these variables are constructed in the form of Likert scales (i.e., ordinal categorical), they were treated in this study according to a naive approach, that is, as quantitative variables. A Shapiro–Wilk test shows that less than a quarter of these manifest variables do not have a normal distribution, which supports the choice of this approach.

The HoCoS was thus elaborated as a quantitative latent variable.

Face validity

A panel of nine experts from different communication-related disciplines was surveyed to define the criteria for inclusion of items in the HoCoS. All the experts (two computer scientists, two speech and language pathologists and PhD students, two speech therapists, one phoniatrician, two researchers in linguistics) were involved for at least 5 years in several research projects related to pathological speech analysis. Moreover, all the speech therapists currently worked in ENT departments.

This selection of relevant items was performed in two steps. First, the experts were surveyed to get a consensus definition of the communication abilities in the context of oral or oropharyngeal cancer. According to the consensual definition, the items from the questionnaires that did not comply with this consensus definition were removed. The experts then participated to an individual selection of the items.

Items were finally retained if they met one of the following two criteria:

- The I-CVI (item-level content validity index) (Lynn, 1986; Polit et al., 2007) was >0.777, which corresponds to an agreement of seven out of nine experts to keep the item.
- The Kappa of agreement was ≥0.81 (Landis & Koch, 1977): 'almost perfect' agreement.

Construct validity

A statistical selection of items respecting face validity was carried out on the criteria of non-redundancy and sufficient variability.

The criterion of non-redundancy allows an analysis if the scores of the items are statistically not associated with each other. An analysis of the inter-item correlation matrix using Spearman coefficients (non-parametric) was used because of the small sample size ($n < 30$). A threshold of 0.90 was chosen: only one of the items correlated with each other at ≥0.90 could be retained for further analysis.

In this case, to obtain sufficiently variable items, and thus allowing a more specific measure of inter-individual variability, only the item with the highest coefficient of variation (if the distribution of this item is gaussian, tested by the Shapiro–Wilk test) or the highest dispersion index (if the distribution is not gaussian) was retained.

Elaboration of the holistic communication score (HoCoS)

Once relevant, non-redundant and sufficiently variable items were selected, the HoCoS was elaborated, using principal component factor (PCF) analysis (Roscoe et al., 1982). PCF analysis is commonly used in data reduction (Acock, 2018) because it attempts to explain as much as possible the variance of a set of items by a single dimension, in other words when a set of items all measure the same concept (communalities set to one, no uniqueness).

This statistical technique therefore suits the objective of this study, where a single quantitative latent holistic score (the HoCoS) is sought among the set of manifest variables (corresponding to the selected items).

Thus, a prediction of the values in PCF analysis for factor 1, corresponding to the latent variable HoCoS, was made. This prediction is derived from a regression analysis on the set of new variables created by estimation of the first factor.

Thus, per subject, the score is predicted by the sum of the standardized values of each item weighted by the regression coefficient corresponding to factor 1.

Validation of the HoCoS

The validation of the score was done in two steps. First, a five-fold cross-validation of the predicted score was led to verify the reliability of the score. A latent profile analysis (Cai, 2012) leading to the construction of a qualitative HoCoS ('HoCoS-Qual') was then compared with the (quantitative) HoCoS.

This type of analysis, based on generalized structural equation modelling (GSEM) models (Coma et al., 2013), allows one to determine which individuals are most likely to belong to a group (corresponding to a category of the latent variable) according to information carried by other variables. Two-, three- and four-class models were computed, and the model with the best Akaike information criterion (AIC) and Bayesian information criterion (BIC) criteria—i.e., the lowest parameters—was selected (Cameron & Trivedi, 2005).

Finally, the class of each subject was predicted. For each subject, the categorical latent variable HoCoS-Qual thus takes one of the values corresponding to one of the classes.

Statistical analyses

The statistical analyses were carried out using Stata 16.1 software (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX, USA: StataCorp LLC).

TABLE 1 Characteristics of the subjects

	Total (%)	
TNM Classification: T (tumour size)	25	
0	0	0.0%
1	2	10.5%
2	6	31.6%
3	3	15.8%
4	8	42.1%
Missing data	6	
Treatments	25	
Surgery alone	1	4%
Radiotherapy alone	1	4%
Surgery + radiotherapy	8	32%
Surgery + radiotherapy + chemotherapy	13	52%
Radiotherapy + chemotherapy	2	8%
Speech and language therapy	**25**	
Ongoing	12	55%
Once a week	*3*	
At least twice a week	*3*	
No information	*6*	
None	10	45%
Missing data	3	

In all analyses, a level of significance at 5% was chosen.

The normality of distribution of quantitative variables was tested using a Shapiro–Wilk test.

Correlation analyses were performed using the following thresholds (Mukaka, 2012): >0.9 (very high correlation), 0.7–0.9 (high correlation), 0.5–0.7 (moderate correlation), 0.3–0.5 (low correlation), and <0.3 (negligible correlation).

RESULTS

Participants

A total of 25 patients filled the questionnaires (median age = 67 years, IQR = 12; 15 males and 10 females; oral cavity = 14, oropharynx = 10, two locations = 1).

A total of 88% of the subjects were treated surgically, 96% by radiotherapy.

The mean time after treatment was 87.2 months (SD = 121.8; median = 40; interquartile range = 123).

Details of TNM classification and treatment strategies are given in Table 1.

Construction of the HoCoS

A set of 174 items were initially collected. The list of the 174 items and reasons for exclusion (if applicable) are given in in the additional supporting information.

The construction process is represented in Figure 1 and will be developed below.

Face validity

As a first step, the panel identified the following consensus criteria for retaining items:

- Items related to oral communication: exclusion of items related to written communication unless they allow for compensation or an increase in oral communication.
- Items relating to expression: exclusion of only comprehension-related items.
- Items relating to interaction between speaker and interlocutor, even if implicit.

This first step led to exclude 83 items: those relating to reading, writing, understanding conversation, items relating to fatigue, pain, specific speech symptoms such as speed of speech, etc.

A total of 91 items (52.3%) were thus retained for further analysis.

In a second step, an online questionnaire using the LimeSurvey tool was submitted to the nine experts. They were asked to indicate which of the remaining 91 items should be retained in the elaboration of the holistic communication score. The experts had to tick the boxes corresponding to all the items to be kept.

A total of 44 items (48.4%) met one of the two previously defined conditions (I-CVI > 0.777 or $\kappa \geq 0.81$) and were retained for further analysis (Figure 2).

Construct validity

Four items were excluded because of a too high correlation with other items (considered as redundant items). In that case, only the item with the bigger variation coefficient or dispersion index was retained (Table 2).

A total of 40 items were finally retained for the construction of the holistic communication score.

Elaboration of the holistic communication score (HoCoS)

The PCF analysis has two conditions of application: first, the first main factor must explain a 'substantial'

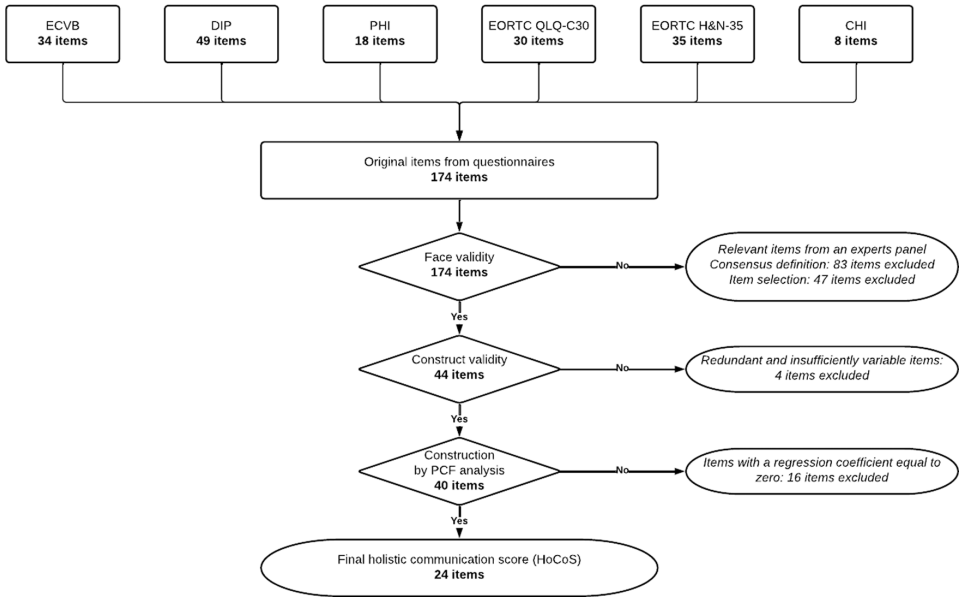


FIGURE 1 Construction process of the holistic communication score

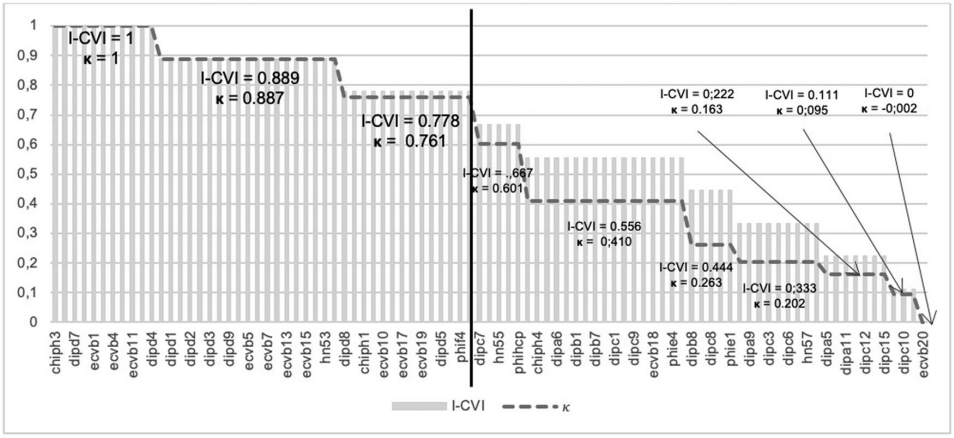


FIGURE 2 I-CVI scores and Kappa of agreement, with cut-off (black line)

part of the total variance for all the items; and second, most of the items must have a load of ≥0.4 on this factor.

Both conditions are met. The eigenvalue of factor 1 (i.e., the proportion of total variance attributable to factor 1) is 17.51 compared with 4.57 for factor 2 (Figure 3). Moreover, the proportion of variance explained solely by factor 1 is 0.44 (0.11 for factor 2). Finally, 34 out of 40 items (85%) have a loading of ≥0.4 on this factor.

Since PCF analysis is applicable, a prediction of values for factor 1, corresponding to the latent variable HoCoS, was performed.

A total of 16 items have a coefficient equal to zero, neutralizing the value taken by the item, and thus indicating that they will not be considered in the overall calculation of the HoCoS. Therefore, 24 items out of 40 (60%) have non-zero coefficients and will be retained for the HoCoS calculation.

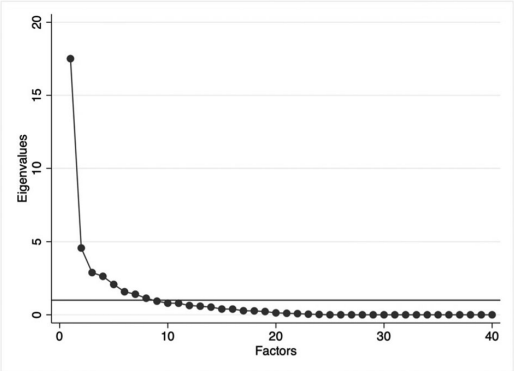


FIGURE 3 Scree plot of the eigenvalues of the factors following the PCF analysis

To facilitate readability and interpretation, the predicted score, initially centred on zero and having a standard deviation (SD) of 1, was centred on 100 with a SD of 10. This transformed score constitutes the HoCoS.

The HoCoS is therefore calculated as follows:

$$HoCoS = \left(\frac{\left[\sum \left(\frac{X_{item} - m_{item}}{s_{item}} \times \beta_{item} \right) \right] - m_{initial\_score}}{s_{initial\_score}} \times 10 \right) + 100$$

$$= \left(\frac{\left[\sum \left(\frac{X_{item} - m_{item}}{s_{item}} \times \beta_{item} \right) \right] - 0.0353408}{0.999787} \times 10 \right) + 100$$

where X_{item} represents the raw item value (score obtained on the item by the subject); m_{item} is the mean of the item obtained in the study sample, s_{item} is the SD of the item obtained in the study sample, β_{item} represents the regression coefficient of the item; $m_{\text{initial\_score}}$ is the mean of the initially predicted score; and $s_{\text{initial\_score}}$ is the SD of the initially predicted score.

The HoCoS are shown in Figure 4.

The complete formula is as follows:

$$\begin{aligned} HoCoS = {} & (((((ecvb1 - 2.36)/0.8103497^{*}0.03264) \\ & + ((ecvb4 - 2.08)/0.9539392^{*}0.01566) \\ & + ((ecvb5 - 1.8)/0.9128709^{*}0.18626) \\ & + ((ecvb6 - 1.4932)/0.8338601^{*}0.10411) \\ & + ((ecvb7 - .34)/0.8524999^{*}0.01695) \\ & + ((ecvb9 - 1.8)/0.9574271^{*} - 0.00532) \\ & + ((ecvb12 - 1.68)/1.215182^{*} - 0.01413) \\ & + ((ecvb13 - 1.56)/1.356466^{*}0.10219) \\ & + ((ecvb16 - 1.72)/1.137248^{*}0.12197) \\ & + ((ecvb17 - 2.3268)/0.8902^{*}0.06856) \\ & + ((ecvb19 - 1.4932)/1.054509^{*} - 0.01007) \\ & + ((ecvb25 - 1.82)/1.081376^{*}0.15096) \\ & + ((ecvb26 - 2.28)/0.9363048^{*} - 0.08137) \\ & + ((dipd2 - 2.88)/1.563117^{*} - 0.04681) \end{aligned}$$

TABLE 2 Non-redundancy analysis

Correlation between the two items of the pair	Item code	Item	Variation coefficient[a] or dispersion index[b]
0.95	ecvb11	Do you have difficulty calling your family?	0.67[a]
	ecvb12	**Do you have difficulty phoning your friends?**	**0.71[a]**
0.92	chiph3	Do you speak less with your family, friends, neighbours?	0.77[a]
	phie2	**My speech difficulties limit my personal and social life**	**0.84[a]**
0.91	**ecvb8**	**And with someone you don't know very well (the letter carrier or a cab driver for example), are you embarrassed to have a conversation on simple subjects? (The weather; what you did the day before; the flowers in your garden …)?**	**0.87[b]**
	ecvb10	Do you find it difficult to speak when you are with people you don't know well (at a dinner party, an outing, an evening out …)?	0.53[a]
0.90	chiph2	Do people have difficulty understanding you?	0.61[a]
	phic4	**I am asked to repeat myself because of my difficulty to speak**	**0.69[a]**

Note: Retained items are shown in bold.

TABLE 3 Akaike information criterion (AIC) and Bayesian information criterion (BIC) values of GSEM models (latent profile analysis—LPA)

Model	AIC	BIC
Latent variable with two classes	2811.97	2959.45
Latent variable with three classes	**2729.59**	**2927.04**
Latent variable with four classes	2747.96	2995.39

Note: The retained model is shown in bold.

$$
\begin{aligned}
&+((dipd3-2.2)/0.8660254^{*}0.1402)\\
&+((dipd5-3.36)/1.113553^{*}-0.03138)\\
&+((dipd6-3.48)/1.262273^{*}0.14321)\\
&+\ ((dipd7-2.64)/1.254326^{*}0.05331)\\
&+((dipd9-2.84)/1.344123^{*}-0.12441)\\
&+\ ((hn53-2.2)/1^{*}-0.071)\\
&+((phif4-2.32)/1.519868^{*}-0.00858)\\
&+((phic1-0.99)/1.251^{*}-0.24677)\\
&+\ ((phic3-1.76)/1.422439^{*}-0.08623)\\
&+((phic4-1.68)/1.180395^{*}-0.07575))\\
&-0.0353408)/0.999787^{*}10)+100
\end{aligned}
$$

Validation of the HoCoS

Five-fold cross-validation

A strong correlation of 0.91 was found between the HoCoS and the values predicted by the five-fold cross-validation (i.e., training on 20 observations and prediction on the other five observations, repeated five times in this case) (Figure 5).

Complementary validation by latent profile analysis

The same 40 items meeting the face and construct validities were retained as manifest variables of the qualitative latent score (HoCoS-Qual).

The model selection parameters AIC and BIC were calculated for models with two, three and four classes (Table 3). The model resulting in a three-class latent variable was thus retained because it has the lowest AIC and BIC criteria.

The class to which each subject belongs was then predicted. For each subject, the categorical latent variable HoCoS-Qual thus takes one of the three values corresponding to one of the three classes (1, 2 or 3).

A comparison of the values of the two latent variables, quantitative (reference HoCoS score) and categorical (HoCoS-Qual), shows that class 1 of the HoCoS-Qual score corresponds to the subjects with the lowest HoCoS scores, class 2 corresponds to the subjects with the highest HoCoS scores and finally class 3 to the subjects with intermediate HoCoS scores (Table 4).

The latent profile analysis leads to a qualitative variable corresponding to a level of impact on communication closely related to the HoCoS score. This analysis also confirms that the HoCoS does correspond to a level of impact on communication, thus validating the construction of this holistic indicator.

DISCUSSION

Psychometrics

A new index measuring communication impairment in patients treated for oral or oropharyngeal cancer was developed in this study.

Despite its construction from limited sample ($n = 25$), the HoCoS shows good performances in validity and reliability. On the one hand, the reliability in cross-validation is high ($r_S = 0.91$). On the other hand, it is also a valid score, which measures the level of impact on communication, which was confirmed by the construction of a qualitative index by a latent profile analysis.

Limitations

However, to better ensure generalizability of the results, this study needs to be completed.

First, increasing the sample size would allow a better statistical power and thus a better generalization of the results. Thus, this study concerns the first validation step of this innovative tool. This preliminary study will have to be followed by a validation of the HoCoS on a larger sample size. To date, there is no consensus on the number of subjects required for score validation. The inclusion of a new sample of about 100 subjects could thus be considered to ensure greater statistical robustness.

Then this score should be evaluated on a new sample of patients for external validation. The analysis of the performance of the HoCoS on a new sample of patients

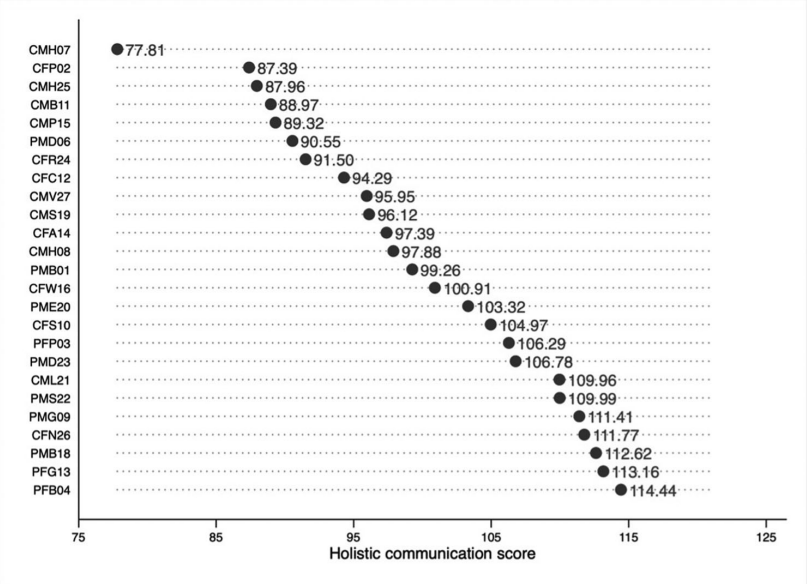


FIGURE 4 Holistic communication scores (HoCoS) by subject

TABLE 4 Comparison of latent classes predicted by latent profile analysis (LPA) analysis and the HoCoS score (subjects are ranked by increasing order of the HoCoS)

Subject	Predicted class	HoCoS	Subject	Predicted class	HoCoS	Subject	Predicted class	HoCoS
CMH07	1	77.81	CMV27	3	95.95	CFS10	2	104.97
CFP02		87.39	CMS19		96.12	PFP03		106.29
CMH25		87.96	CFA14		97.39	PMD23		106.78
CMB11		88.97	CMH08		97.88	CML21		109.96
CMP15		89.32	PMB01		99.26	PMS22		109.99
PMD06		90.55	CFW16		100.91	PMG09		111.41
CFR24		91.50	PME20		103.32	CFN26		111.77
CFC12		94.29				PMB18		112.62
						PFG13		113.16
						PFB04		114.44

would allow to ensure its reliability, and thus again its generalizability.

Finally, the temporal reliability of this score remains to be analysed. This point is closely related to the temporal reliability performance of the questionnaires from which the items used to calculate the HoCoS are taken. However, the non-preservation of all the items of the initial questionnaires modifies the global structure of these questionnaires, and the temporal reliability remains to be verified on specific items presented in a different order. The 24 items retained for the construction of the HoCoS will thus have to be presented in two stages (D0 and D7) to a new sample of patients.

Perspectives

Speaker and interlocutor in communication situation

The holistic communication score is elaborated solely from items from self-reported questionnaires. The measure

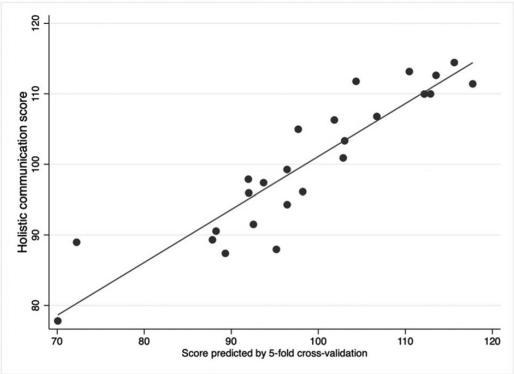


FIGURE 5 Scatterplot representing HoCoS and score predicted by cross-validation, with regression line

of communication impairment is therefore only self-reported.

However, when measuring communication abilities, the speaker is a communication partner in the same way as the interlocutor. The latter is particularly important because communication is only effective if the message is not only correctly transmitted but also correctly received.

The inclusion in the HoCoS of an external measure of communication abilities by a listener would allow to consider other dimensions linked more globally to the impact of the disorder on the comprehensibility of speech (by a listener) (Pommée et al., 2021) and by ripple effect on communication and quality of life. Different tools could be used for this purpose, such as an evaluation by the listener using a visual analogue scale, or communicative dynamics questionnaires for example. The analysis of the results of a score combining internal and external measures according to the tools used could provide new insights into the dynamics of communication in a social context.

Communication environment according to the bio-psychosocial models

Bio-psychosocial models such as Wilson's (Wilson & Cleary, 1995) represent the links between symptomatic (speech disorders) and functional (communication) status. According to these models, factors related to the characteristics of the individual or the environment can influence both the speech disorder and the communication abilities.

Taking into account these factors, such as the cognitive and anxiety–depressive state (Böhm et al., 2016; Eadie et al., 2018) the constitution of social circles around the patient (Danker et al., 2010) or the patient's self-perception of the speech impairment (Bolt et al., 2016) would allow a better understanding of the functional dynamics in patients treated for oral cavity or oropharyngeal cancer.

More globally, the association of the holistic communication score with indicators related to the individual and his environment and a measure of the speech disorder could also allow a more effective prediction of the psychosocial impact of speech disorders in these patients, and their quality of life.

Clinical implications

The HoCoS is an index that is voluntarily holistic in its construction, taking into account symptomatologic (e.g., item phif4 'I use a great deal of effort to speak'), interactional (e.g., item phic3 'I have trouble communicating with unfamiliar people'), pragmatic (e.g., ecvb25 'At the restaurant/coffee shop, do you find it difficult to place your order yourself?'), but also psycho-affective (e.g., dipd9 'I feel comfortable speaking in most situations both at home and outside') dimensions. Strategies for compensating for communication difficulties are also considered (e.g., dipd3 'I try other ways of getting my message across when people don't understand me').

In our sample, 55% of the subjects included had speech therapy. However, speech therapy follow-up can modify patients' perception of their own communication abilities (Jacobi et al., 2010). The HoCoS would have a relevant clinical applicability in the rehabilitation process of head and neck cancer patients as one of the indicators to be considered by speech therapists, by providing a valid and reliable measure of progress in follow-up.

This indicator thus fills a gap in tools for measuring the functional consequences of oral cavity and oropharyngeal cancer treatment on speech and communication in daily care. The use of the HoCoS in clinical care would allow to better target the daily problems met by the patients in their communication with peers, and thus to better adapt the therapeutic strategies to their reported needs.

It thus seems interesting to let the patient in head and neck oncology to fill the HoCoS, that is, the 24 items from the ECVB, DIP, PHI and EORTC QLQ-H&N35 questionnaires. This score structured in few items allows to quickly target the impact of the speech disorder on communication. This measurement could be systematized in consultation or evaluation and would allow the therapeutic strategy to be adjusted as closely as possible to the patient's needs.

Finally, to consider the patients' quality of life, beyond the functional dimension evaluated by our new score, the HoCoS could thus be associated with specific quality of life

questionnaires (Raquel et al., 2020) in a bio-psychosocial follow-up approach.

CONCLUSIONS

A global score allowing a measurement of the impact of speech impairment on communication after treatment of oral or oropharyngeal cancer has been developed. The methodology of its construction allows a better reflection of the symptomatological, pragmatic and psychosocial elements leading to a degradation of communication abilities. The temporal reliability of this score and its external validity remains to be explored. Nevertheless, it fills the gap in the absence of this type of tool in head and neck oncology, and may allow a better understanding of the factors involved in the functional and psychosocial limitations of these patients.

ORCID

Mathieu Balaguer https://orcid.org/0000-0003-1311-4501

Virginie Woisard https://orcid.org/0000-0003-3895-2827

NOTES

[1]Source: Globocan—Global Cancer Observatory 2020 World Health Organization, https://gco.iarc.fr/today/data/factsheets/cancers/3-Oropharynx-fact-sheet.pdf.

[2]See https://www.irit.fr/rugbi/.

ACKNOWLEDGEMENTS

The authors acknowledge Benoît Lepage for his help in methodological considerations.

CONFLICTS OF INTEREST

All authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data and the database are available from the corresponding author upon request.

REFERENCES

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J. et al. (1993) The European organization for research and treatment of cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *JNCI Journal of the National Cancer Institute*, 85(5), 365–376. Available from: http://jnci.oxfordjournals.org/content/85/5/365.short

Acock, A.C. (2018) *A gentle introduction to stata*, 6th edition, Tex: StataCorp LP, p. 570.

Balaguer, M., Boisguérin, A., Galtier, A., Gaillard, N., Puech, M. & Woisard, V. (2019) Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 136(5), 355–359. Available from: https://doi.org/10.1016/j.anorl.2019.05.012

Balaguer, M., Champenois, M., Farinas, J., Pinquier, J. & Woisard, V. (2021) The (head and neck) carcinologic handicap index: validation of a modular type questionnaire and its ability to prioritise patients' needs. *European Archives of Oto-Rhino-Laryngology*, 278(4), 1159–1169. Available from: http://link.springer.com/10.1007/s00405-020-06201-6

Balaguer, M., Farinas, J., Fichaux-Bourin, P., Puech, M., Pinquier, J. & Woisard, V. (2020) Validation of the French versions of the speech handicap index and the phonation handicap index in patients treated for cancer of the oral cavity or oropharynx. *Folia Phoniatrica et Logopaedica*, 72(6), 464–477. Available from: https://www.karger.com/Article/FullText/503448

Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V. & Speyer, R. (2020) Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: systematic review. *Head & Neck*, 42(1), 111–130. Available from: https://doi.org/10.1002/hed.25949

Baylor, C.R., Yorkston, K.M., Eadie, T.L., Miller, R.M. & Amtmann, D. (2009) Developing the communicative participation item bank: Rasch analysis results from a spasmodic dysphonia sample. *Journal of Speech, Language, and Hearing Research*, 52(5), 1302–1320. Available from: https://doi.org/10.1044/1092-4388/282009/07-0275/29

Bjordal, K., Hammerlid, E., Ahlner-Elmqvist, M., de Graeff, A., Boysen, M., Evensen, J.F. et al. (1999) Quality of life in head and neck cancer patients: validation of the European organization for research and treatment of cancer quality of life questionnaire-H&N35. *Journal of Clinical Oncology*, 17(3), 1008–1008. Available from: https://doi.org/10.1200/JCO.1999.17.3.1008

Böhm, N., Knipfer, C., Maier, A., Bocklet, T., Rohde, M., Neukam, F. et al. (2016) Sprechqualität und psychische Beeinträchtigung nach der Therapie von Mundhöhlentumoren. *Laryngo-Rhino-Otologie*, 95(09), 610–619. Available from: https://doi.org/10.1055/s-0042-102256

Bolt, S., Eadie, T., Yorkston, K., Baylor, C. & Amtmann, D. (2016) Variables associated with communicative participation after head and neck cancer. *JAMA Otolaryngology – Head & Neck*, 142(12), 1145–1151. Available from: https://doi.org/10.1001/jamaoto.2016.1198

Borggreven, P.A., Verdonck-De Leeuw, I.M., Muller, M.J., Heiligers, M., De Bree, R., Aaronson, N.K. et al. (2007) Quality of life and functional status in patients with cancer of the oral cavity and oropharynx: pretreatment values of a prospective study. *European Archives of Oto-Rhino-Laryngology*, 264(6), 651–657.

Borsboom, D. (2008) Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives*, 6(1–2), 25–53. Available from: https://doi.org/10.1080/15366360802035497

Cai, L. (2012) Latent variable modeling. *Shanghai Archives of Psychiatry*, 24(2), 118–120.

Cameron, C. & Trivedi, P. (2005) *Microeconometrics: methods and applications*. Cambridge: Cambridge, p. 1056.

Carreira-Perpinán, M. (1997) *A review of dimension reduction techniques*. Sheffield: Department of Computer Science, The University of Sheffield Techical Report CS-96-09, pp. 1–69. Available from: http://www.pca.narod.ru/DimensionReductionBrifReview.pdf

Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P. et al. (2013) Deciphering the complex:

methodological overview of statistical models to derive OMICS-based biomarkers. *Environmental and Molecular Mutagenes*, 54(7), 542–557. Available from: https://doi.org/10.1002/em.20575

Coma, E., Ferran, M., Méndez, L., Iglesias, B., Fina, F., & Medina, M. (2013) Creation of a synthetic indicator of quality of care as a clinical management standard in primary care. *Springerplus*, 2(1), 51. Available from: https://doi.org/10.1186/2193-1801-2-51

Cunningham, P. (2014) Dimension reduction. *Machine learning techniques for multimedia*. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 91–112. Available from: https://doi.org/10.1007/978-3-540-75171-7_4

Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, E. & Meyer, A. (2010) Social withdrawal after laryngectomy. *European Archives of Oto-Rhino-Laryngology*, 267(4), 593–600. Available from: https://doi.org/10.1007/s00405-009-1087-4

Doward, L.C. & McKenna, S.P. (2004) Defining patient-reported outcomes. *Value Heal*, 7(Supplement 1), S4–S8. Available from: https://doi.org/10.1111/j.1524-4733.2004.7s102.x

Eadie, T., Faust, L., Bolt, S., Kapsner-Smith, M., Pompon, R.H., Baylor, C. et al. (2018) Role of psychosocial factors on communicative participation among survivors of head and neck cancer. *Otolaryngology – Head & Neck Surgery*, 159(2), 266–273. Available from: https://doi.org/10.1177/0194599818765718

Jacobi, I., van der Molen, L., Huiskens, H., van Rossum, M.A. & Hilgers, F.J.M. (2010) Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. *European Archives of Oto-Rhino-Laryngology*, 267(10), 1495–1505. Available from: https://doi.org/10.1007/s00405-010-1316-x

Landis, J.R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159.

Letanneux, A., Walshe, M., Viallet, F. & Pinto, S. (2013) The dysarthria impact profile: a preliminary French experience with Parkinson's disease. *Parkinsons Disease*, 2013, 1–6. Available from: http://www.hindawi.com/journals/pd/2013/403680/

Lynn, M.R. (1986) Determination and quantification of content validity. *Nursing Research*, 35, 382–386. Available from: http://ijoh.tums.ac.ir/index.php/ijoh/article/view/26

Mazaux, J.-M., Daviet, J.-C., Darrigrand, B., Stuit, A., Muller, F., Dutheil, S. et al. (2006) Difficultés de communication des personnes aphasiques. *Évaluation des Troubles Neuropsychologiques en vie Quotidienne*, 73–82. Springer, Paris. http://ijoh.tums.ac.ir/index.php/ijoh/article/view/26

Meyer, T.K., Kuhn, J.C., Campbell, B.H., Marbella, A.M., Myers, K.B. & Layde, P.M. (2004) Speech intelligibility and quality of life in head and neck cancer survivors. *Laryngoscope*, 114(11 I), 1977–1981.

Middag, C., Martens, J.P., Van Nuffelen, G. & De Bodt, M. (2009) Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, 2009(5), Springer Science and Business Media LLC. https://doi.org/10.1155/2009/629030

Mlynarek, A., Rieger, J., Harris, J., O'Connell, D., Al-Qahtani, K., Ansari, K. et al. (2008) Methods of functional outcomes assessment following treatment of oral and oropharyngeal cancer: review of the literature. *Journal of Otolaryngology–Head and Neck Surgery*, 37(1), 2–10.

Mukaka, M.M. (2012) Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.

Murphy, B.A., Ridner, S., Wells, N. Dietrich, M. (2007) Quality of life research in head and neck cancer: a review of the current state of the science. *Critical Reviews in Oncology/Hematology*, 62(3), 251–267.

Polit, D.F., Beck, C.T. & Owen, S.V. (2007) Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *The Research in Nursing & Health*, 30(4), 459–467. Available from: https://doi.org/10.1002/nur.20199

Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J. & Woisard, V. (2021) Intelligibility and comprehensibility: a Delphi consensus study. *The International Journal of Language & Communication Disorders*, 57(1), 21–41. Available from: https://doi.org/10.1111/1460-6984.12672

Raquel, A.C.S., Buzaneli, E.P., Silveira, H.S.L., Simões-Zenari, M., Kulcsar, M.A.V., Kowalski, L.P. et al. (2020) Quality of life among total laryngectomized patients undergoing speech rehabilitation: correlation between several instruments. *Clinics (Sao Paulo)*, 75, e2035.

Roscoe, B.A., Hopke, P.K., Dattner, S.L. & Jenks, J.M. (1982) The use of principal component factor analysis to interpret particulate compositional data sets. *Journal of the Air Pollution Control Association*, 32(6), 637–642.

Walshe, M., Peach, R.K. & Miller, N. (2009) Dysarthria impact profile: development of a scale to measure psychosocial effects. *The International Journal of Language & Communication Disorders*, 44(5), 693–715. Available from: https://doi.org/10.1080/13682820802317536

Wilson, I.B. & Cleary, P.D. (1995) Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA The Journal of the American Medical Association*, 273(1), 59–65. Available from: https://doi.org/10.1001/jama.273.1.59

Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M. et al. (2021) Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: the Carcinologic Speech Severity Index. *Head & Neck*, 44(1), 71–88. Available from: https://doi.org/10.1002/hed.26903

World Health Organization (WHO). (2001) *International classification of functioning, disability and health: ICF*. WHO. Available from: https://apps.who.int/iris/handle/10665/42407

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Balaguer, M., Pinquier, J., Farinas, J. & Woisard, V. (2022) Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation. *International Journal of Language & Communication Disorders*, 1–13. https://doi.org/10.1111/1460-6984.12766

A.5 Article dans le journal JASA

L'article « Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech » a été publié le 29 novembre 2022 dans le Journal of Acoustical Society of America [Vaysse et al., 2022a].

Ces travaux constituent la première partie du travail de doctorat de Robin Vaysse (cf. section 6.2.5 page 101), réalisé en collaboration avec Corine Astésano dans le cadre du projet RUGBI (cf. section 5.4.15 page 83). IL en résulte des conseils sur les algorithmes d'extraction de la fréquence fondamentale à utiliser dans le cas de la parole pathologique.

JASA ARTICLE

Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech

Robin Vaysse,[1,a),b)] Corine Astésano,[2,c)] and Jérôme Farinas[1]
[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]Laboratoire de NeuroPsychoLinguistique, Université Toulouse Jean-Jaurès, France

ABSTRACT:
Reliable fundamental frequency (f_0) extraction algorithms are crucial in many fields of speech research. The current bulk of studies testing the robustness of different algorithms have focused on healthy speech and/or measurements of sustained vowels. Few studies have tested f_0 estimations in the context of pathological speech, and even fewer on continuous speech. The present study evaluated 12 available pitch detection algorithms on a corpus of read speech by 24 speakers (8 healthy speakers, 8 speakers with Parkinson's disease, and 8 with head and neck cancer). Two fusion methods' algorithms have been tested: one based on the median of algorithms and one based on the fusion between the best algorithm for voicing detection and the algorithm that generates the most accurate f_0 estimations on voiced parts. Our results show that time-domain algorithms, like REAPER, are best for voicing detection while deep neural network algorithms, like FCN-f_0, yield better accuracy for the f_0 values on voiced parts. The combination of REAPER and FCN-f_0 yields the best ratio performance/implementation complexity, since it generates less than 4% errors on voicing detection and less than 5% of gross errors in the estimation of the f_0 values for all speaker groups.
 https://doi.org/10.1121/10.0015143

(Received 2 August 2022; revised 29 September 2022; accepted 26 October 2022; published online 29 November 2022)
[Editor: B. Yegnanarayana] Pages: 3091–3101

I. INTRODUCTION

The measurement of the fundamental frequency (f_0) is an essential element of automatic speech processing, particularly in the study of prosody. It is, therefore, crucial to have a good estimate of this parameter. Many algorithms have been developed for estimating the fundamental frequency of healthy speech recorded under good conditions (without noise), which provide very good f_0 approximations (see Sec. II). In the context of pathological speech, the calculation of precise f_0 variations is necessary because most pathologies have an impact on voice quality, more specifically on speakers' inability to maintain a stable fundamental frequency (jitter, shimmer) (Jiménez-Jiménez *et al.*, 1997). In addition, the dynamics of the fundamental frequency in a sentence defines the intonation, which corresponds to the voice "melody." Intonation provides main communicative functions and is a powerful tool for the illocutionary and structural interpretation of the speaker's message (Di Cristo, 2016). Yet, some pathologies can lead to a poor control of intonation that can induce confusion as to the type of sentence the speaker is trying to produce (Le Dorze *et al.*, 1994), which affects both his/her intelligibility and comprehensibility. When we want to model intonation or stress patterns from the f_0, these types of errors can lead to distorted interpretations over large time spans.

It is, therefore, crucial to use an f_0 extraction algorithm, which is as accurate as possible and avoids gross estimation errors (such as dividing by two or doubling the real value of the fundamental frequency) or errors in the detection of voiced or unvoiced areas. When working on large corpora of pathological voice recordings, such as Cesari *et al.* (2018), this issue is even more challenging because the amount of data does not allow for precise manual annotations. The objective of the present study is, therefore, to test several different algorithms in the particular context of pathological voice, such as those resulting from head and neck cancers (H&NC) or Parkinson's disease (PD).

Several performance evaluation studies of pitch detection algorithms have been designed on non-pathological speech (de Cheveigné and Kawahara, 2001; Strömbergsson, 2016) and it seems that auto-correlation function (ACF) from Praat (Boersma and Weenink, 2020) and the YIN algorithms (de Cheveigné and Kawahara, 2002) are good methods for typical, healthy voices. Some studies also looked into the evaluation of noisy speech, which best corresponds to real recording conditions (Jouvet and Laprie, 2017; Luengo *et al.*, 2007). These results show that, while all the evaluated algorithms provide comparable results on healthy speech, an increase in background noise results in a loss of algorithm performance, specifically with regard to the detection of voicing. More specifically, the robust algorithm for pitch tracking (RAPT) and the robust epoch pitch estimator (REAPER) algorithms seem to provide good results on

[a)]Also at: Laboratoire de NeuroPsychoLinguistique, Université Toulouse Jean-Jaurès, France
[b)]Electronic mail: robin.vaysse@irit.fr
[c)]Also at: UMR 5267 Praxiling - Université Paul Valéry Montpellier, France

https://doi.org/10.1121/10.0015143 JASA

noisy data while the ACF algorithm does not provide good results on noisy speech. Indeed, according to Jouvet and Laprie (2017), ACF generates an error rate of 4.6% for voicing detection on noise-free speech, but this rate increases to 16.2% with the addition of noise with a signal to noise ratio (SNR) level of 10 db, while the REAPER error rate only increases from 5% to 8.3% with the same SNR level. Pathological speech was evaluated marginally in some studies, such as Parsa and Jamieson (1999). The authors compared seven pitch detection algorithms on sustained vowels of pathological speech where the patients showed benign vocal lesions, such as polyps, nodules, and cysts. They showed that among the compared algorithms, the ACF and average magnitude difference function (AMDF) algorithms were good fits for pathological voices. Later on, Jang *et al.* (2007) also compared seven pitch detection algorithms on sustained vowels. This last study concluded that the ACF (Boersma and Weenink, 2020) performed best on their dataset. Tsanas *et al.* (2014) compared 10 pitch detection algorithms on a sustained vowel task and showed that the sawtooth waveform inspired pitch estimator (SWIPE) and the nearly defect-free (NDF) algorithms provide the best f_0 estimates on their dataset. They also proposed a new combination algorithm based on Kalman filters that was 16% more accurate than the best algorithm tested. According to this brief review of literature, the next step is to test pitch detection algorithms in connected pathological speech, which has, to our knowledge, never been addressed. The present study tackles this issue on two different pathologies: H&NC and PD. These pathologies have quite different impacts on f_0: H&NC present a variety of f_0 alterations (e.g., hoarseness, dysfluences interrupting coherent intonation groups) while PD does not so much impact the global linguistic features of f_0 but rather the dynamics of f_0 variations. These two pathologies, thus, allow for complementary insight on f_0 detection algorithms. Furthermore, a comparison between the classical algorithms of pitch detection and the new emerging methods (Ardaillon and Roebel, 2019; Kim *et al.*, 2018) based on deep neural networks could be interesting.

Closer to our present goals, the study by Jouvet and Laprie (2017) using speech in noise was of particular interest to help us determine which algorithms to test on our pathological speech corpora. Specifically, algorithms whose performances were highest were chosen, and they have been sorted according to their differences of implementation (spectral vs time domain; post- vs pre-processing). In addition to these algorithms, three additional algorithms have been integrated: pitch estimation filter with amplitude compression (PEFAC), an algorithm that is in the spectral domain with or without pre- or post-processing (Gonzalez and Brookes, 2014), that was designed to perform on noisy speech; convolutional representation for pitch estimation (CREPE) (Kim *et al.*, 2018), which uses a pre-trained convolutional neural network; and fully convolutional networks for f_0 detection (FCN- f_0) (Ardaillon and Roebel, 2019), which is a fully-convolutional neural network that has been trained to optimize both f_0 computation and voicing detection while the other deep neural network algorithm, CREPE, has been optimized for f_0 estimation only. Section II presents the different fundamental frequency detection algorithms that have been selected for this study. Section III describes the voice recordings and the method used to extract the real values of f_0 and also the evaluation metrics used to evaluate the performances of the algorithms. Finally, Sec. IV presents our results and proposes leads to understand the differences observed between healthy and pathological speech with the various tested algorithms.

II. FUNDAMENTAL FREQUENCY DETECTION ALGORITHMS

Speech researchers need reliable fundamental frequency detection programs and have a wide variety of choice (see Sec. II A). In the case of pathological speech research, it is much trickier to decide which f_0 algorithm is best suited for degraded voice quality. This section describes our selection process leading to the choice of the f_0 algorithms that will be used in our testing section. The programs were selected primarily on the basis of their availability and ease of access, to fit the ecological situation of most speech researchers interested in f_0 detection. Our selection was also motivated by the need to cover a large variety of algorithms based on temporal and frequency representations, and those based on deep learning methods. It is commonly acknowledged that most f_0 extraction algorithms can be decomposed in three steps:

- First, a pre-processing of the signal can be applied to remove unnecessary information. For example, yet another algorithm for pitch tracking (YAAPT) (Kasi and Zahorian, 2002) applies a bandpass filter between 100 Hz and 900 Hz to the signal, while some algorithms apply a low-pass filter on the raw signal [e.g., the YIN algorithm (de Cheveigné and Kawahara, 2001) or the REAPER algorithm from Google-Open-Source (2015)].
- Then, candidates' values of f_0 are extracted using, for example, temporal or spectral representations of the signal.
- Finally, a step of post-processing takes the pitch candidates and chooses those more likely to be good f_0 estimations. For example, CREPE (Kim *et al.*, 2018) applies a Viterbi smoothing to remove isolated or incoherent values (sudden drops or increases of f_0). Some algorithms also use dynamic programming (Bellman, 1954) like REAPER, RAPT (Ghahremani *et al.*, 2014), or YAAPT (Kasi and Zahorian, 2002), allowing sudden jumps in the final f_0 curve to be minimized.

Aside from these common prerequisites and even though their global architectures are generally similar, pitch detection algorithms implementations otherwise differ in many ways. We purposely used these programs with default parameters, as researchers commonly do when first using such algorithms. Section II A will present the different approaches underlying the f_0 extraction algorithms.

JASA
https://doi.org/10.1121/10.0015143

A. Pitch detection algorithms typology

The f_0 information can be retrieved using time-domain representations of signal or frequency domain representations. Time-domain algorithms generally use the autocorrelation algorithm which consists of extracting a signal window of a few milliseconds and searching whether the detected pattern is repeated in successive signal windows. This method computes the correlation between two windowed signals. The resulting f_0 value will consist of the delay τ between windows that present the best correlation. Some of them use a slightly different method called cross correlation which computes the correlation between the signal and a modified version of itself like a downsampled version of the signal.

The following methods use time-domain algorithms: ACF (Boersma, 2000), YIN (de Cheveigne and Kawahara, 2002), AMDF (Ross *et al.*, 1974) and REAPER are based on the autocorrelation algorithm, while RAPT (Talkin and Kleijn, 1995) and enhanced RAPT (Ghahremani *et al.*, 2014) use cross correlation to extract pitch candidates.

As far as frequency domain algorithms are concerned, the f_0 values are selected by looking at the occurrence of the f_0 harmonics in the spectrum. In the present study, the following algorithms were chosen: PEFAC (Gonzalez and Brookes, 2014) and SWIPE (Camacho and Harris, 2008), which are known to be robust even for noisy speech signals (Jouvet and Laprie, 2017).

Some algorithms can also use information from time and frequency domains to refine their selection of pitch values, such as NDF (Kawahara *et al.*, 2005) and YAAPT (Kasi and Zahorian, 2002) for which the combination allows selection of the most likely pitch candidates.

Finally, new kinds of algorithms use deep neural networks like CREPE (Kim *et al.*, 2018) or FCN-f_0 (Ardaillon and Roebel, 2019). These methods rely on machine learning techniques, where the algorithm is trained to compute f_0 values from a raw signal with no explicit procedure. It is, thus, difficult to know what kind of information from the signal those algorithms use and whether it works in the time domain, the frequency domain, or both. These methods, however, provide robust results. The list of algorithms selected for the present study is given in Table I.

B. Algorithm merging

In addition to the above algorithms, we decided to integrate techniques based on combinations of different algorithms to test whether these combinations can reduce common errors. Indeed, autocorrelation-based algorithms tend to produce halving f_0 errors (estimated f_0 two times smaller than the real value) while methods based on frequency domain produce more doubling errors (estimated f_0 two times bigger than the real value). It is, thus, interesting to mix time and frequency domain algorithms to compensate for their respective common errors. An example of previous mixing f_0 algorithms techniques can be found in Espesser (1999) with the toolkit MES-SignAix (Espesser, 1996), which used a majority vote between different algorithms.

Tsanas *et al.* (2014) also used a median vote between 10 different pitch f_0 algorithms as a baseline for algorithm combinations. They found no improvement over the NDF algorithm alone on the raw accuracy of f_0 estimation on a sustained vowel task. However, we believe that using a simple median filtering on top of complementary algorithms (which produce different kinds of f_0 estimations errors) can improve the reliability of f_0 measurements. They also used a method based on a Kalman filter to merge different algorithm results. This latter method provides promising results on their sustained vowels dataset. Unfortunately, the codebase of their algorithm is not publicly available.

With this in mind, two simple methods for merging algorithms were selected. The first method is a "majority vote" obtained by using the median between the different values of several selected algorithms (Hess, 2008; Soquet, 1994; Espesser, 1999). The choice of the median allows us to eliminate gross f_0 errors. At least three different algorithms were computed at a time: because each algorithm generates different results on the same file, the median of these values was used as the f_0 estimate. A 10 ms frame was chosen to comply with the pseudo-stationary property of the

TABLE I. List of algorithms tested in the present study, with a link to the chosen implementation. The last three columns indicate whether the algorithm works on the signal's time or spectral domain, or whether it uses deep learning.

Algorithm	Implementation	Time domain	Spectral	Neural network
ACF (Boersma, 2000)	Praat	X		
AMDF (Ross et al., 1974)	Snack Sound toolkit (Kåre, 2005)	X		
REAPER (Google-Open-Source, 2015)	https://github.com/google/REAPER	X		
RAPT (Talkin and Kleijn, 1995)	Snack Sound toolkit (Kåre, 2005)	X		
Enhanced RAPT (Ghahremani *et al.*, 2014)	Kaldi (Povey *et al.*, 2011)	X		
Yin (de Cheveigne and Kawahara, 2002)	https://github.com/patrice.guyot/Yin	X		
NDF (Kawahara et al., 2005)	STRAIGHT (Kawahara, 2018)	X	X	
YAAPT (Kasi and Zahorian, 2002)	MATLAB implementation (Zahorian and Hu, 2016)	X	X	
SWIPE (Camacho and Harris, 2008)	Speech signal processing toolkit (Tokuda *et al.*, 2017)		X	
PEFAC (Gonzalez and Brookes, 2014)	VOICEBOX (Brookes, 2018)		X	
CREPE (Kim et al., 2018)	https://github.com/marl/crepe			X
FCN-f_0 (Ardaillon and Roebel, 2019)	https://github.com/ardaillon/FCN-f0			X

https://doi.org/10.1121/10.0015143 JASA

TABLE II. Illustration of the merging of algorithms by the resulting f_0 median vote: The first column indicates the start time of the 10 ms frame, the next 3 columns are the estimated values by an example of 3 different algorithms, and the last column with boldface values is the f_0 median voting.

Time (s)	f_0 Yin	f_0 ACF	f_0 *SWIPE*	f_0 *Median*
0	0	140	0	**0**
0.01	0	189	181	**181**
0.02	170	173	169	**170**
—	—	—	—	—

speech signal (Hess, 2008). An illustration of the method is described in Table II.

The second merging method consists of the actual fusion/combination of two algorithms, by taking an algorithm (Algorithm A) that is particularly efficient on the calculation of the f_0 value and another algorithm that is very accurate on the voicing detection (Algorithm B). Algorithm B is then used to select the voiced time windows and Algorithm A gives the estimated f_0 values in those windows. Table III describes the method used.

III. EXPERIMENTAL SETUP

A. Recordings description

This work is part of the ANR project RUGBI 2018-2023 (RUGBI, 2018-2023) in which the main goal is to find specific speech markers that impact speakers' intelligibility. Two pathological speech corpora from this project were used.

The first corpus is taken from the Carcinologic Speech Severity Index (C2SI) project (see Acknowledgements), in which 127 speakers were recorded, consisting of 40 control subjects and 87 patients who had been treated for cancer of the oral cavity or pharynx. Speakers were recorded in several tasks (such as sustained /a/, non-words reading, short-text reading, picture description, prosodic functions encoding) For a complete description of the corpus, see Woisard *et al.* (2021).

Our second source of pathological speech is the Aix Hospital Neuro (AHN) corpus, with 209 recordings of PD patients (112 controls) described in Ghio *et al.* (2012). A subset of this corpus is used in the RUGBI project. Speakers also recorded several tasks, such as those described in the C2SI corpus plus additional tasks specifically designed for this pathology (e.g., diadochokinesis, singing, breathing). Both corpuses used the same short-text reading (Daudet, 1870), which was used in the present study to test the f_0 algorithms on connected speech. This task has been chosen to have consistent recordings between the different speakers (the first four sentences of the text are common to both corpora), but also because the recordings are relatively long (ranging from 20 to 70 s). A subset of 24 speakers composed of 8 healthy patients, 8 patients with H&NC, and 8 patients with PD (4 men and 4 women in each group) has been selected.

The selection of patient files was based on perceptual analysis conducted by specialized clinicians to assess the quality of patients' voices on the reading task. Speakers with the most degraded voice quality were selected to test the algorithms under the most difficult conditions. In total, the corpus is composed of 120 sentences (40 per group) and represents roughly 13 min of recordings.

TABLE III. Illustration of the merging of algorithms by fusion/combination. The last column with bold values represents the final f_0 estimation based on the two previous columns. The voicing detection is based on A (if there is a 0 in A then it is reported on the resulting combination value) and the estimated f_0 values are taken from the Algorithm B.

Time (s)	Algorithm B	Algorithm A	*f0* Combination
0	0	140	**0**
0.01	173	189	**189**
0.02	170	180	**180**
0.03	0	0	**0**
—	—	—	—

B. Manual f_0 annotations

1. Manual correction of period detection errors

Since the files were not recorded using an Electro-Glotto-Graph, it was necessary to fully annotate the f_0 manually. To do so, the Praat software (Boersma and Weenink, 2020) was used to perform a first automatic annotation using an algorithm based on the autocorrelation of the signal (Boersma, 2000). This annotation was then manually corrected at the signal level by indicating the boundaries of each pattern to obtain a fundamental frequency corresponding to the real value (illustration on Fig. 1). Once corrected, the fundamental frequency curve was extracted with values every 10 ms (Fig. 2).

Our resulting manual annotated f_0 constitutes our reference (gold standard) and was then compared to the outputs of all the algorithms described in Sec. II A.

2. Simultaneous f_0 zones (diplophonia)

When analyzing pathological speech, we encounter numerous instances where the fundamental frequency seems to drop abruptly, with periodic patterns that double in length in the signal. An example is shown in Fig. 3.

The length of the patterns at the beginning of the signal corresponds to a frequency of about 126 Hz; a new periodic pattern suddenly appears, which is much longer (66 Hz) and disappears after a few milliseconds. This phenomenon leads to simultaneous frequencies, one quite low and another usually an octave higher. From a perceptual point of view, this results in a hoarse voice and an indeterminate fundamental frequency (Keating *et al.*, 2015). Further studies need to be run to better encompass this phenomenon, but they go beyond the scope of our paper. These complex areas have been annotated to observe how the algorithms behave on these segments. In these zones, the annotation of the f_0 value was performed using a linear interpolation between the previous (stable) f_0 values and the following values. This choice is justified by the fact that these segments are relatively

JASA
https://doi.org/10.1121/10.0015143

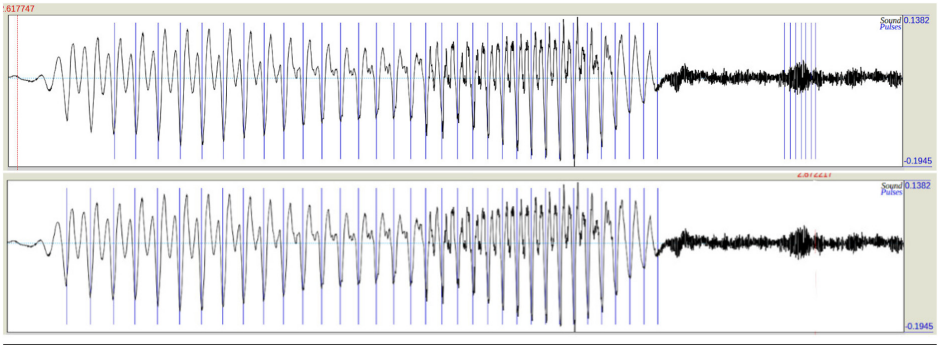


FIG. 1. (Color online) Example of annotation: the automatic marking of periodic pattern boundaries from Praat is at the top and the manually corrected annotation is at the bottom of the figure. Note that the first two periods on the left were not detected as voiced and an unvoiced segment at the end was detected as periodic.

short (less than 100 ms on average) and are, thus, hardly perceived as sudden f_0 drops. These complex periodic areas are actually quite rare. In total, in this corpus, simultaneous frequencies represent:

- 0.9% of voiced segments for healthy speech
- 4.5% of voiced segments for cancer patients (with a maximum of 13% for one speaker)
- 1.5% of voiced segments for PD patients.

C. Metrics

To have an objective evaluation of the quality of the different algorithms, three metrics and four sub-metrics classically used were computed to exhaustively evaluate the computation of f_0 (Jouvet and Laprie, 2017; Drugman and Alwan, 2011; Babacan *et al.*, 2013; Chu and Alwan, 2009). These metrics allow us to describe the different types of errors produced by the algorithms, whether they are voicing detection errors or errors in the estimation of the real value of f_0:

- Voicing detection error (VDE), which measures the 10 ms frame proportion containing errors in the detection of voicing; two sub-metrics for voicing detection were also added
- False negative rate (FNR), which computes the proportion of voiced frame detected as unvoiced by the algorithm
- False positive rate (FPR), which computes the proportion of unvoiced frame detected as voiced by the algorithm
- Gross pitch error (GPE), which measures the proportion of frames where the estimated value differs from the real value by more than 20%
- The proportion of frames where the estimated value is at least 20% higher than the reference value ($\times 2$)
- The proportion of frames where the estimated value is at least 20% lower than the reference value ($\div 2$)

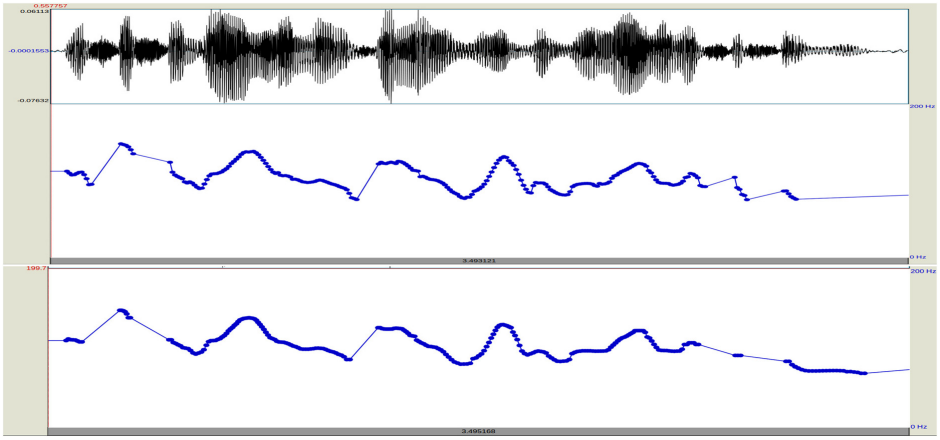


FIG. 2. (Color online) Example of file annotation from a healthy speaker on the sentence, "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres.*" Automatic f_0 from Praat is at the top and the manually corrected annotation is at the bottom. Each blue point corresponds to a f_0 value for a 10 .ms frame.

https://doi.org/10.1121/10.0015143
JASA

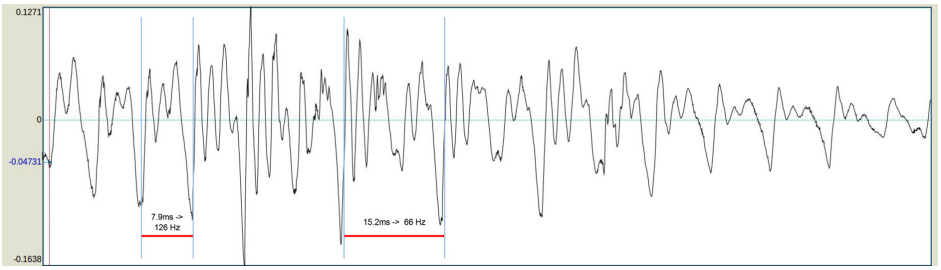


FIG. 3. (Color online) Example where the fundamental frequency seems to decrease suddenly by an octave.

- FFE (f_0 frame error), which measures the frame proportion where an error was detected, whether it is a voicing detection error or a gross pitch error.

D. Experimental protocol

Each of the 12 algorithms (see Table I) were then executed on all audio files, respecting the recommended sampling frequencies for each algorithm and using the default settings of the different implementations. Unfortunately, some algorithms are quite sensitive to parameters like pitch range, silence threshold and it is necessary to adjust them for each speaker. However, we purposely chose not to modify them, to evaluate which algorithms best adapt to the data quickly. The f_0 estimate was calculated on 10 ms windows (Hess, 2008), which corresponds to the generated annotations (see Sec. III B). Some algorithms generate a voicing probability score, which makes it possible to determine whether the analyzed window contains a fundamental frequency value. It is then possible to test different thresholds to determine whether the portion of the signal is voiced. For example, one can consider that if the probability is less than 80%, then the estimate of the f_0 is set to 0. To determine an optimal threshold, for each algorithm, the threshold that minimizes the VDE metric (see Sec. III C) was chosen.

If the algorithm used does not generate a voicing probability, then the unvoiced sections are set to 0 by the algorithm and the metrics are computed directly.

IV. RESULTS

We now present the results obtained with the different algorithms described in Table I, with the manual annotations described in Sec. III B as a reference (gold standard). In addition to these algorithms, the median vote described in 2.2 was computed on 5 methods: AMDF, Kaldi, NDF, FCN-f_0, and REAPER. The median vote was applied to these 5 methods because they give the best vote results, and they are based on various calculation methods. The results of the median vote will be referred to as "median" in Figs. 4–6.

Finally, the combination of two algorithms was also computed using REAPER as a basis for the detection of voicing, and the values estimation was given by the FCN-f_0 method. This choice is based on the fact that REAPER generates the best overall results concerning the voicing detection and FCN-f_0 provides the most accurate estimates of f_0 on our dataset. The results of the combination of those two algorithms will be referred to as ⟨⟨ Combi ⟩⟩ in the following figures.

First, the global results for the VDE were analyzed, then GPE, and the final step consists in analyzing mixed errors (FFE metric). These tests were done after excluding speech areas with simultaneous fundamental frequencies (Sec. III B 2). All the detailed results are included in the Sec. V presented in Table IV.

A. Voicing detection

Figure 4 shows the results obtained for different algorithms, on the metric VDE. The abscissa shows the different algorithms that have been tested, while the proportion (between 0 and 1) of analysis frame windows (10 ms) with a VDE is shown on the ordinate. The figure compares the results from the "Healthy group", the group with H&NC, and the group with PD.

It seems that algorithms based on the signal time domain give the best results for voicing detection. The RAPT, ACF, and AMDF algorithms give similar results with about 5% of windows containing a voicing error on the voices of speakers with H&NC and PD, and about 4% for control speakers. REAPER algorithm seems to be the best one for patients with H&NC and PD with 3.5% of errors on speech of patients with H&NC cancer, 3.3% on speech of patients with PD, and 3.6% for healthy speakers.

The majority of errors in these algorithms are found at the beginning and end of the voiced segments where detection usually starts slightly too early and ends too late. One can notice that the methods based on deep neural networks provide good results with about 8% of VDE, which is, however, not as good as the time-domain algorithms. This could be explained by the fact that the voicing decision has to be made by using a hard threshold on probabilities that can lead to isolated errors if the threshold is not strict enough. The results for the patients with PD also show the same trend with time-domain methods being more effective for voicing detection with about 4% of errors. The median

JASA
https://doi.org/10.1121/10.0015143

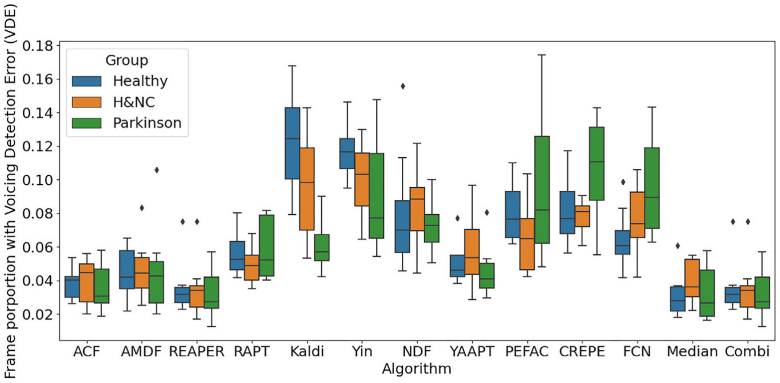


FIG. 4. (Color online) Results on voicing detection errors: Blue boxplots represent VDE for control speakers, orange boxplots are for speakers with H&NC, and green boxplots are for PD patients. Each boxplot represents an error percentage (lower percentages are optimal).

voting and YAAPT also gave similar results to time-domain algorithms.

It is noteworthy that the results obtained on the VDE metric are surprising because, for many algorithms, the performances are comparable in pathological and non-pathological speech.

B. Gross pitch errors

Figure 5 represents the results for the metric GPE, i.e., the percentage of voiced frames for which an algorithm has generated a value distanced by more than 20% from the real value of f_0.

For this metric, methods based on neural networks give good results with 1% errors for CREPE and 0.5% for FCN-f_0 on pathological voices. On the other hand, methods based on autocorrelation give good results on healthy speech, but produce more errors on pathological voice with much higher variations between speakers. A large part of these errors come from a one-octave estimation below the real value for time domain methods. Globally, GPE yields fewer good results for pathological speech compared to healthy speech, which confirms that the evaluation of the exact f_0 value on pathological speech is harder than for healthy, clean speech.

Surprisingly, the results are really good for PD patients, and are really close to the results for healthy speakers. Indeed, almost all algorithms give good results with less than 3% of mistakes. No difference in favor of one method or the other can be noticed. Overall, the performance of the algorithms on voices of speakers with H&NC show more errors on the determination of the f_0 value than for the two other groups.

C. f_0 frame errors

Figure 6 represents the results for the metric f_0 FFE, i.e., the percentage of frames where the algorithm has made a mistake, whether it is a voicing detection mistake or a gross pitch one.

Based on Fig. 6, it is difficult to choose one algorithm that outperforms the others. However, some of them provide relatively good results with little variation between speakers (small error bars). For example, YAAPT (which uses both time domain and spectral domain of the signal) provides

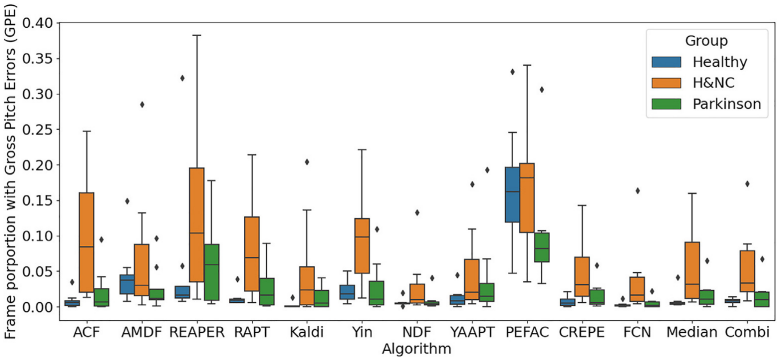


FIG. 5. (Color online) Results on gross pitch errors: Blue boxplots represent GPE for control speakers, orange boxplots are for speakers with H&NC, and green boxplots are for PD patients. Each boxplot represents an error percentage (lower percentages are better).

https://doi.org/10.1121/10.0015143

JASA

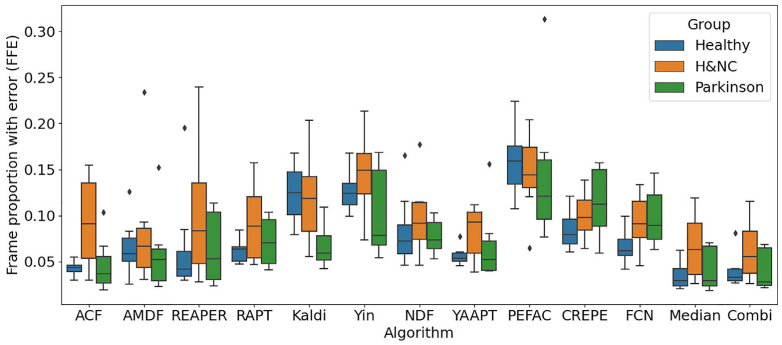


FIG. 6. (Color online) Results on f_0 frame errors for patients with H&NC and patients with PD: each boxplot represents an error percentage (lower percentages are better). Blue boxplots represent FFE for healthy speakers, orange boxplots are for speakers with H&NC, and green boxplots are for speakers with PD.

good results with only 7% of frames containing errors for pathological and healthy speech. FCN- f_0 also gives good and stable results with 7.3% of error for cancer patients and 6.2% for healthy speakers.

Finally, for PD patients, the results are again similar to those on healthy speakers. The algorithms with best detection performance are YAAPT, ACF, AMDF, and the merging methods.

D. Simultaneous f_0 zones

Concerning the simultaneous f_0 areas from the diplophonia phenomenon described in Sec. III B 2, we decided not to make a quantitative evaluation of the various pitch detection algorithms because it concerns too few segments (n = 35). Instead of computing statistics, a qualitative analysis were performed by looking more in detail at the outcomes for these particular zones. As indicated in Sec. III B 2, the f_0 values in these areas were annotated as an interpolation of the previous and next f_0 values. The reason behind this choice is that, for our future experiments, we will study stress and intonation patterns on the same corpus of pathological speech. To have a good estimation of the linguistic implementation of f_0, it is crucial to have a regular, precise f_0 curve and avoid sudden drops or increases. On the other hand, it is also valuable to have algorithmic estimations of f_0 sudden changes when searching to precisely characterize pathological speech. We we chose to compare different algorithms to our manual annotation (in black) for this reason.

After analyzing the results on the 35 overlapping f_0 intervals, a trend can be observed. The algorithms based on the time domain of the signal tend to show a sudden drop by an octave, as exhibited by the ACF curve in Fig. 7. The ACF algorithm generates a drop by an octave for 28 intervals (80%). Also, some algorithms avoid the drop by doing an interpolation between the previous and the following f_0 values. Typically, YAAPT seems to provide consistent results with 24 intervals (69%) in this particular case. Also, the neural network algorithms, like FCN- f_0, often give stable f_0 curves like YAAPT, but they also tend to tag these superposed f_0 patterns as unvoiced segments.

V. DISCUSSION AND CONCLUSION

This paper analyzed the performance from 12 algorithms based on time domain, frequency domain, or deep neural network techniques on clinical data including healthy speech, speakers with H&NC, and speakers with PD. Two methods of merging algorithms (combi and median) were also tested to determine whether the potential performance improvement induced by these merges was sufficient to alleviate its inherent technical constraints. The main objectives were to test the following:

(1) a large amount of widely used pitch detection algorithms for these two particular diseases, because they yield different f_0 detection problems
(2) these algorithms on connected speech, which is more ecological for automatic pitch detection tools but nevertheless never proposed in similar studies.

The experiments were run on a corpus composed of 24 French speakers (8 healthy, 8 for H&NC, 8 for PD) performing a reading task. The performance of the algorithms was tested through three metrics: VDE for voicing detection, GPE for estimation accuracy, and FFE for overall performance. A test were also computed based on a selection of three algorithms representative of time or frequency domain and deep neural network on specific speech areas with simultaneous f_0 typical of the diplophonia phenomenon. Indeed, although these simultaneous f_0 areas are relatively scarce, it is interesting to characterize the algorithms' behavior for future research.

Regarding the three metrics (VDE, GPE, FFE), our results indicate that algorithms based on the temporal study of the signal generate better results for voice detection.

JASA
https://doi.org/10.1121/10.0015143

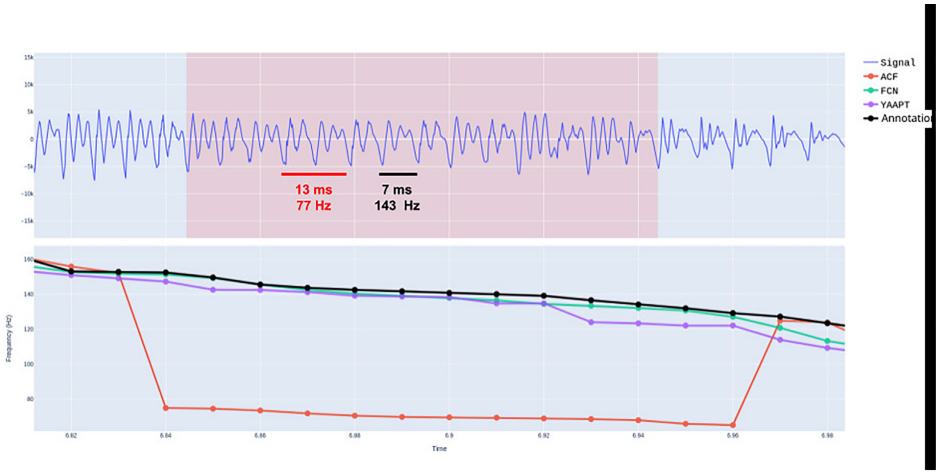


FIG. 7. (Color online) Example of results for three algorithms (ACF in red, FCN-F0 in green, and YAAPT in violet) on a particular zone with superposed f_0. The upper figure shows the raw signal with a superposed f_0 zone highlighted.

Indeed, ACF, REAPER, and AMDF algorithms offer the best VDE values with similar results for all speaker types (healthy or pathological) with an error rate below 5%. The algorithms using deep neural network however are most efficient to account for the accuracy of the estimates (GPE). Indeed, CREPE and FCN- f_0 produce less than 0.5% gross errors on healthy and Parkinsonian individuals. As a reminder, an estimation is considered as a gross error when it differs by more than 20% from the reference value. The results for speakers with H&NC generally yield worse GPE values, regardless of the algorithm chosen. Despite this, CREPE and FCN generate less than 1.5% error on average for this type of speech.

By running the three metrics on the fusion algorithm methods, our results indicate that the "median" algorithm fusion method using five algorithms (AMDF, Kaldi, NDF, FCN- f_0, and REAPER) generates overall better results than these methods taken individually. The median vote yields less than 4% of VDE for all speaker groups and less than 5% of GPE. However, the performance improvement does not seem to be significant enough to justify the technical cost (implementation and execution time) of running five algorithms. On the other hand, our results on the "combi" algorithm fusion method indicate that the simple combination of two methods (mixing REAPER for detecting voicing and FCN- f_0 for estimating the f_0 value) is efficient enough as it generates results at least similar to the "median" vote (around 5% of VDE and less than 3% of GPE), while executing only 2 algorithms.

To summarize, time-domain methods are best for voicing detection while deep neural networks generate more accurate f_0 estimations. Using a combination of just two algorithms (REAPER for good voicing detection and FCN-f_0 for accurate estimations) is a good compromise between performance and computational complexity.

From a clinical point of view, choosing the fittest f_0 detection algorithm is a crucial element depending on the studies one wishes to perform. However, this choice is rarely justified in clinical studies of pathological speech pitch. This study allowed us to highlight the strengths and weaknesses of different methods. It may, thus, help to best choose the relevant algorithm (or combination of algorithms) in future clinical studies, depending on the finality and purpose of the research. For example, if a study is particularly interested in the presence or absence of voicing, the use of an algorithm, such as ACF or AMDF, seems recommended as we did in Vaysse *et al.* (2022) where we measured the f_0 on a large corpora of Parkinsonian speech. Conversely, if one is mostly interested in retrieving the mean f_0 values of a corpus, it is more interesting to use methods, such as REAPER, FCN- f_0, or NDF, which generate robust estimates even if some voiced areas are missed. Interestingly, our results indicate that the same f_0 extraction algorithms (NDF and FCN- f_0 for accurate pitch estimation and REAPER or ACF for voicing detection) turn out to be the most reliable for speakers with PD and H&NC.

ACKNOWLEDGMENTS

This work has been carried out thanks to the French National Research Agency in 2018 as part of the RUGBI 2018 project entitled "Looking for relevant linguistic units to improve the intelligibility measurement of speech production disorders" (Grant No. ANR-18-CE45-0008). The first corpus is taken from the Carcinologic Speech Severity Index (C2SI) project (Grant No. 2014-135) from Institut National du Cancer (INCa).

https://doi.org/10.1121/10.0015143

JASA

APPENDIX: RAW RESULTS

The raw results of the tested algorithms are presented in Table IV.

TABLE IV. Raw results of the 14 tested algorithms on the different metrics and sub-metrics. C, controls (healthy) speakers; H, H&NC; P, Parkinson's disease patients; VDE, voicing detection errors; FNR, false negative rate of VDE; FPR, false positive rate of VDE; GPE, the gross pitch errors; ×2 and ÷2, rates of algorithm estimations above and below 20% from our annotations, respectively (see Sec. III B). Values in boldface are the best scores for each metric.

	VDE (%)			GPE (%)			FNR (%)			FPR (%)			×2 (%)			÷ 2 (%)		
Algorithm	C	H	P	C	H	P	C	H	P	C	H	P	C	H	P	C	H	P
ACF	3.8	4.0	3.6	0.9	10.0	2.1	**1.3**	2.4	2.4	2.5	1.7	1.1	0.1	**0.7**	0.2	0.8	9.3	1.9
AMDF	4.5	4.7	4.6	4.6	7.2	2.6	3.1	3.6	3.6	1.4	1.0	1.0	2.1	4.6	1.4	2.4	2.7	1.2
REAPER	3.6	**3.5**	**3.3**	5.8	13.7	6.3	1.5	**1.3**	**1.7**	2.0	2.2	1.6	0.2	1.4	0.2	5.6	12.3	6.2
RAPT	5.6	4.9	5.9	1.2	8.6	2.6	2.6	2.2	**1.7**	3.0	2.7	4.2	0.2	1.4	0.5	1.0	7.3	2.1
Enhanced RAPT	12.4	9.6	6.0	**0.2**	5.3	1.3	8.1	6.7	3.3	4.2	2.9	2.7	0.1	**0.9**	0.6	**0.1**	4.4	0.7
Yin	11.7	10.1	9.1	2.3	9.8	2.8	6.8	6.6	6.1	4.9	3.5	3.0	0.1	1.2	0.2	2.2	8.5	2.6
NDF	8.0	8.4	7.3	0.6	**3.0**	0.9	1.4	3.2	2.5	6.7	5.2	4.8	0.2	1.9	0.5	0.4	**1.0**	**0.3**
YAAPT	5.0	5.9	4.6	1.3	5.0	4.0	2.7	4.4	2.9	2.3	1.4	1.6	0.2	1.7	0.3	1.1	3.2	3.8
SWIPE	6.7	9.2	6.2	79.2	68.1	91.8	6.0	8.8	5.1	**0.7**	**0.4**	1.1	18.3	5.4	9.1	60.9	62.7	82.7
PEFAC	8.1	6.6	9.5	16.8	16.7	10.5	4.1	3.5	3.9	4.0	3.0	5.5	6.9	8.0	7.1	9.9	8.6	3.4
CREPE	8.2	7.8	10.6	0.8	4.8	1.6	2.9	4.4	4.7	5.3	3.4	5.9	0.4	2.1	0.7	0.4	2.7	0.9
FCN-F0	6.5	7.7	9.6	0.3	3.8	**0.5**	3.7	5.0	5.2	2.8	2.6	4.3	**0.0**	1.7	**0.1**	0.3	2.1	0.4
Median	**3.1**	3.9	**3.3**	0.9	5.7	1.7	1.9	2.8	2.1	1.2	1.1	1.2	**0.0**	1.5	0.3	0.9	4.2	1.4
Combi	3.6	**3.5**	**3.3**	0.7	5	1.6	2.6	4.0	3.6	1.4	1.2	**0.9**	0.1	1.9	0.5	0.3	**1.0**	**0.3**

Ardaillon, L., and Roebel, A. (**2019**). "Fully-convolutional network for pitch estimation of speech signals," in *Proc. Interspeech 2019*, pp. 2005–2009.

Babacan, O., Drugman, T., D'Alessandro, N., Henrich, N., and Dutoit, T. (**2013**). "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, pp. 1–5.

Bellman, R., (**1954**). "The theory of dynamic programming," Bull. Am. Math. Soc. **60**(6) 503–515.

Boersma, P. (**2000**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences 17*.

Boersma, P., and Weenink, D. (**2020**). "Praat: Doing phonetics by computer (version 6.1.16) [computer program]," http://www.praat.org (Last viewed January 20, 2022).

Brookes, M. (**2018**). {VOICEBOX: Speech Processing Toolbox for MATLAB, available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Last viewed November 12, 2022).

Camacho, A., and Harris, J. (**2008**). "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am. **124**, 1638–1652.

Cesari, U., De Pietro, G., Marciano, E., Niri, C., Sannino, G., and Verde, L. (**2018**). "A new database of healthy and pathological voices," Comput. Electr. Eng. **68**, 310–321.

Chu, W., and Alwan, A. (**2009**). "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3969–3972.

Daudet, A. (**1870**). *Lettres de mon moulin: Impressions et souvenirs (Letters from my Windmill)* (Hetzel, Paris).

de Cheveigné, A., and Kawahara, H. (**2001**). "Comparative evaluation of F0 estimation algorithms," in *Eurospeech*, NA, France, pp. 2451–2454.

de Cheveigne, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930.

Di Cristo, A. (**2016**). *Les musiques du français parlé: Essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain (The Music of Spoken French: Essays on Accentuation, Metrics, Rhythm, Prosodic Phrasing and Intonation of Contemporary French)* (de Gruyter, Berlin).

Drugman, T., and Alwan, A. (**2011**). "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1973–1976.

Espesser, R. (**1996**). "MES : un environnement de traitement du signal" ("MES: A signal processing environment"), *XXIèmes Journées d'\'Etude sur la Parole* (*XXIst Study Days on the Word*), Avignon, France, p. 447.

Espesser, R. (**1999**). "Mes signaix package," Technical Report.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (**2014**). "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2494–2498.

Gonzalez, S., and Brookes, M. (**2014**). "PEFAC - A pitch estimation algorithm robust to high levels of noise," IEEE/ACM Trans. Audio. Speech. Lang. Process. **22**(2), 518–530.

Google-Open-Source (**2015**). "Reaper: Robust epoch and pitch estimator," https://github.com/google/REAPER (Last viewed September 20, 2020).

Hess, W. J. (**2008**). *Pitch and Voicing Determination of Speech with an Extension Toward Music Signals* (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 181–212.

Jang, S.-J., Choi, S,-H., Kim, H.-M., Choi, H.-S., and Yoon, Y.-R. (**2007**). "Evaluation of performance of several established pitch detection algorithms in pathological voices," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 620–623.

Jiménez-Jiménez, F. J., Gamboa, J., Nieto, A., Guerrero, J., Orti-Pareja, M., Molina, J. A., García-Albea, E., and Cobeta, I. (**1997**). "Acoustic voice analysis in untreated patients with Parkinson's disease," Parkinsonism Relat. Disord. **3**(2), 111–116.

Jouvet, D., and Laprie, Y. (**2017**). "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1614–1618.

Kåre, S. (**2005**). The Snack Sound Toolkit (Version 2.2.10), available at https://www.speech.kth.se/snack/index.html (Last viewed November 11, 2022).

Kasi, K., and Zahorian, S. (**2002**). "Yet another algorithm for pitch tracking," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, p. 361.

Kawahara, H., Cheveigné, A., Banno, H., Takahashi, T., and Irino, T. (**2005**). "Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight," in *Ninth European Conference on Speech Communication and Technology*, pp. 537–540.

JASA
https://doi.org/10.1121/10.0015143

Kawahara, H. (**2018**). STRAIGHT, a speech analysis, modification and synthesis system, available at http://web.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html (Last viewed November 11, 2022).

Keating, P. A., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow*, Vol. 2015, pp. 2–7.

Kim, J. W., Salamon, J., Li, P., and Bello, J. P. (**2018**). "CREPE: A convolutional representation for pitch estimation," arXiv:1802.06182.

Le Dorze, G. L., Ouellet, L., and Ryalls, J. (**1994**). "Intonation and speech rate in dysarthric speech," J. Commun. Disorders **27**(1), 1–18.

Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J., and Sainz, I. (**2007**). "Evaluation of pitch detection algorithms under real conditions," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 4, pp. IV-1057– IV-1060.

Parsa, V., and Jamieson, D. G. (**1999**). "A comparison of high precision F0 extraction algorithms for sustained vowels," J. Speech. Lang. Hear. Res. **42**(1), 112–126.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (**2011**). "The Kaldi Speech Recognition Toolkit", in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (IEEE, New York).

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., and Manley, H. (**1974**). "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech, Signal Process. **22**(5), 353–362.

RUGBI (**2018–2023**). "Looking for relevant linguistic units to improve the intelligibility measurement of speech production disorders," https://www.irit.fr/rugbi (Last viewed November 10, 2022).

Soquet, A. (**1994**). "Approche coopérative de l'extraction de la fréquence fondamentale" ("A cooperative approach of f0 extraction"), in *XXèmes Journées D'Études Sur la Parole (XXth Study Days on the Word)*, Trégastel, France, pp. 229–234.

Strömbergsson, S. (**2016**). "Today's most frequently used F_0 estimation methods, and their accuracy in estimating male and female pitch in clean speech," in *Proc. Interspeech 2016*, pp. 525–529.

Talkin, D., and Kleijn, W. B. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier Science B. V., Amsterdam), pp. 495–518.

Tsanas, A., Zañartu, M., Little, M. A., Fox, C., Ramig, L. O., and Clifford, G. D. (**2014**). "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," J. Acoust. Soc. Am. **135**(5), 2885–2901.

Tokuda, K., Oura, K., Yoshimura, T., Tamamori, A., Sako, S., Zen, H., Nose, T., Takahashi, T., Yamagishi, J., and Nankaku, Y. (**2017**). Speech Signal Processing Toolkit (Version 3.11), available at https://sp-tk.sourceforge.net/ (Last viewed Novembver 12, 2022).

Vaysse, R., Ghio, A., Astésano, C., Farinas, J., and Viallet, F. (**2022**). "Analyse macroscopique des variations et modulations de F0 en lecture dans la maladie de Parkinson: Données sur 320 locuteurs "Macroscopic analysis of f0 variations and modulations in read speech for Parkinson disease patients: Data from 320 speakers"), in *34e Journées D'Études Sur la Parole (JEP2022)*, [34th 740 Speech Study Days (JEP2022)] (Association Française de la Communication Parlée, Noirmoutier, France, to be published).

Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., and Roger, V. (**2021**). "C2SI corpus: A database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers," Lang. Resour. Eval. **55**(1), 173–190.

Zahorian, S. A., and Hu, H. (**2016**). YAAPT Pitch Tracking MATLAB Function, available at http://www.ws.binghamton.edu/zahorian/yaapt.htm (Last viewed November 11, 2022).

Annexe B

Diplôme de doctorat

REPUBLIQUE FRANCAISE
Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche
ACADEMIE DE TOULOUSE

UNIVERSITE PAUL SABATIER
TOULOUSE III

ATTESTATION
DOCTORAT DE L'UNIVERSITE PAUL SABATIER

Arrêté du 25 Avril 2002

Le Secrétaire Général de l'Université Paul Sabatier de Toulouse, soussigné, certifie que :

Monsieur Jérôme FARINAS

Né(e) le : 16/08/1974

A présenté en soutenance le : 15/11/2002

Devant ladite Université, une Thèse

Intitulé : UNE MODELISATION AUTOMATIQUE DU RYTHME POUR L'IDENTIFICATION DES LANGUES

Après délibération , le jury a prononcé l'admission de :

Monsieur Jérôme FARINAS

Au titre de : DOCTEUR DE L'UNIVERSITE PAUL SABATIER

Mention : TRES HONORABLE

Fait à Toulouse, le 03 décembre 2002

Pour le Secrétaire Général
Et par délégation
L'Attachée Principale d'Administration
Scolaire et Universitaire,
Chef de la Division
De la Scolarité

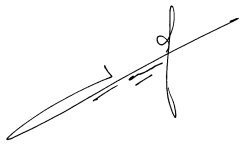
Annick PAYTAU

Il ne peut être délivré de duplicata de la présente Attestation. Il appartient à l'intéressé d'en établir des Copies qu'il fait certifier conformes à l'original

Annexe C

Attestation sur l'honneur de non double inscription

Je sous-signé, Jérôme Farinas, atteste m'inscrire à l'Université Paul Sabatier, et déclare sur mon honneur n'avoir pas réalisé d'inscription dans un autre établissement pour l'obtention du diplôme d'Habilitation à Diriger les Recherches.



Annexe D

Programmation dynamique

Sakoe et Chiba [Sakoe and Chiba, 1978] ont formalisé l'application de la méthode de programmation dynamique, introduite par Bellman [Bellman et al., 1954], aux signaux de parole. Sakoe a également généralisé cette cet algorithme à la prise en compte de mots connectés [Sakoe, 1979].

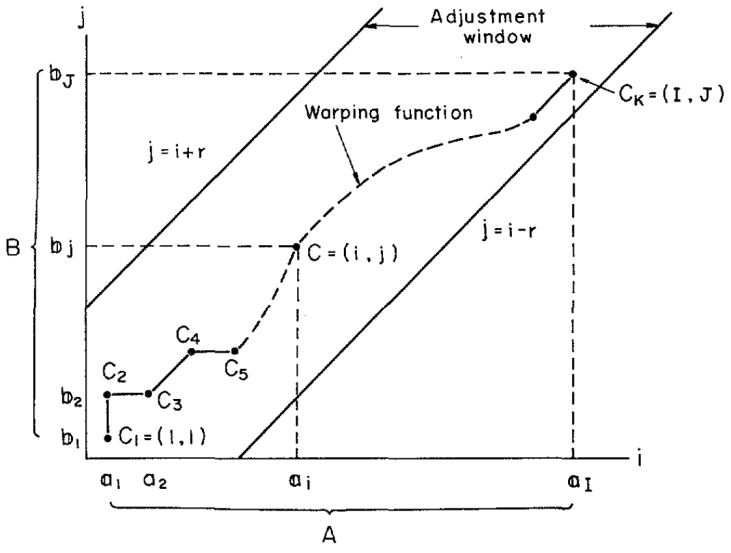


FIGURE D.1 – Fonction d'alignement entre deux séquences acoustiques (extrait de [Sakoe and Chiba, 1978])

Le principe repose sur le besoin d'alignement et de comparaison entre deux séquences temporelles de longueur différente. Deux enregistrements de signaux audio ne sont pas en général identiques, même si l'on contrôle les conditions d'enregistrement, que l'on utilise le même locuteur et que le même mot est prononcé. Les phénomènes d'articulation et co-articulation vont immanquablement produire des distorsions entre les deux signaux. En effet on peut retrouver de la variance au niveau de la réalisation des fréquences mais également au niveau temporel : allongements et rétrécissements temporels peuvent faire varier le signal. Si l'on représente ces enregistrements par une séquence de vecteur représentatifs des fréquences (par exemple une séquence de vecteurs MFCC [Davis and Mermelstein, 1980]), chaque élément représente donc une petite fenêtre temporelle prise sur le signal. Donc deux séquences représentant des réalisations différentes peuvent donc être représentés par des séquences de longueur

différente, en fonction de l'élasticité des contraintes articulatoires. La méthode de programmation dynamique permet de mesurer les dissimilarités entre ces deux séquences, tout en permettant des allongements ou des rétrécissements dans les séquences à comparer, et en recherchant un chemin optimal d'appariement. La figure D.1 illustre cette recherche d'optimum entre deux séquences de longueurs I et J.

L'algorithme consiste à calculer un tableau de dissemblance g en respectant certaines contraintes (monotonie, continuité, coïncidence des extrémités). La construction est faite de manière itérative, en calculant par accumulation la dissemblance des séquences. Dans le domaine discret, il n'est pas possible d'introduire des contraintes de continuité : on introduit des notions de contraintes locales et globales.

Les contraintes globales (cf. figure D.1), limitent le nombre de chemins possible « autour » de la diagonale. Elles permettent de ne pas considérer les cas extrêmes correspondant à un très grand nombre d'insertions et de suppression, qui ont peut de chance de contenir les chemins optimaux d'alignement. Cela a une très grande incidence sur le nombre de calculs à réaliser et influe donc grandement sur la rapidité de l'algorithme.

Les contraintes locales entre en jeu lors du calcul de la dissemblance cumulée minimale entre deux éléments à comparer des séquences. Le calcul est réalisé en cumulant les dissemblances antérieures à la comparaison des éléments courants. On peut pondérer les cheminements, et privilégier certains passages.

Exemple de contrainte simple (avec une pondération de $(1, 2, 1)$) :

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2 * d(i,j \\ g(i,j-1) + d(i,j) \end{cases} \tag{D.1}$$

La fonction de distance $d(i,j)$ permet de comparer les éléments en position i de la première séquence à celui en position j de la deuxième. En général une distance euclidienne est utilisée. La figure D.2 reprend cette formulation et montre graphiquement les cheminements possibles en rouge.

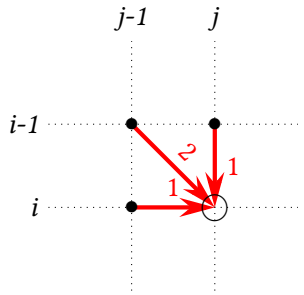


FIGURE D.2 – Contrainte locale simple

Les contraintes locales peuvent être plus étendues sur les valeurs antérieures à considérer. Il est également possible d'envisager des contraintes non symétriques.

Exemple de contrainte plus complexe :

$$g(i,j) = \min \begin{cases} g\left(i-2,j-1\right) + 2 * d\left(i-1,j\right) + d\left(i,j\right) \\ g\left(i-1,j-1\right) + 2 * d\left(i,j\right) \\ g\left(i-1,j-2\right) + 2 * d\left(i,j-1\right) + d\left(i,j\right) \end{cases} \tag{D.2}$$

La figure D.3 en donne une représentation schématique.

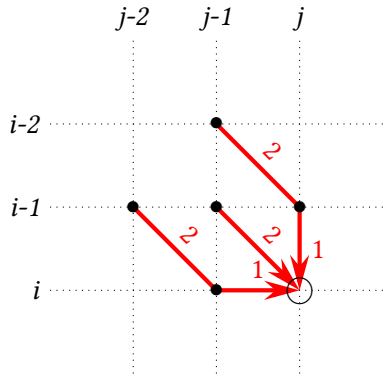


FIGURE D.3 – Contrainte locale plus complexe et symétrique

L'algorithme 1 est un exemple d'implémentation qui ne prend pas en compte les contraintes globales, et qui utilise une contrainte locale simple (cf. figure D.3).

Algorithme 1 : Exemple d'algorithme de programmation dynamique avec une contrainte locale simple

Données : Deux séquences (1..I) et (1..J) et une distance d pour comparer chaque élément
Résultat : Distorsion cumulée entre les deux séquences

début
  $w_0 \leftarrow 1$;
  $w_1 \leftarrow 2$;
  $w_2 \leftarrow 1$;
  $g(0,0) \leftarrow 0$;
  pour $j \leftarrow 1$ **à** J **faire**
    $g(0,j) \leftarrow +\infty$;
  fin
  pour $i \leftarrow 1$ **à** I **faire**
    $g(i,0) \leftarrow +\infty$;
    pour $j \leftarrow 1$ **à** J **faire**
      /* recherche du chemin minimal */
      $g(i,j) \leftarrow min(g(i-1,j) + w_0 * d(i,j), g(i-1,j-1) + w_1 * d(i,j),$
       $g(i,j-1) + w_2 * d(i,j))$;
    fin
  fin
  $D = g(I,J)/(I+J)$;
fin

L'algorithme permet d'obtenir la mesure de dissemblance optimale entre deux mots. Si l'on souhaite réaliser un système de reconnaissance de la parole, il convient de se constituer une base de mots de référence pour se constituer un dictionnaire. Ensuite lorsque l'on souhaite reconnaître un mot, il faut calculer les dissemblances entre tous les mots du dictionnaire. La minimisation de la dissemblance permet de rapprocher prendre une décision par rapport au mot le plus proche. Il est également possible de mettre en place un seuil sur la dissemblance (qui est normalisée par rapport aux longueurs des mots) afin de pouvoir traiter les cas de mots n'appartenant pas au dictionnaire. Mais la reconnaissance ne

peut donc se faire que si le dictionnaire d'enregistrements est suffisamment proche de l'enregistrement à reconnaître. Donc le genre et le locuteur sont des critères importants pour l'appariement des enregistrements.

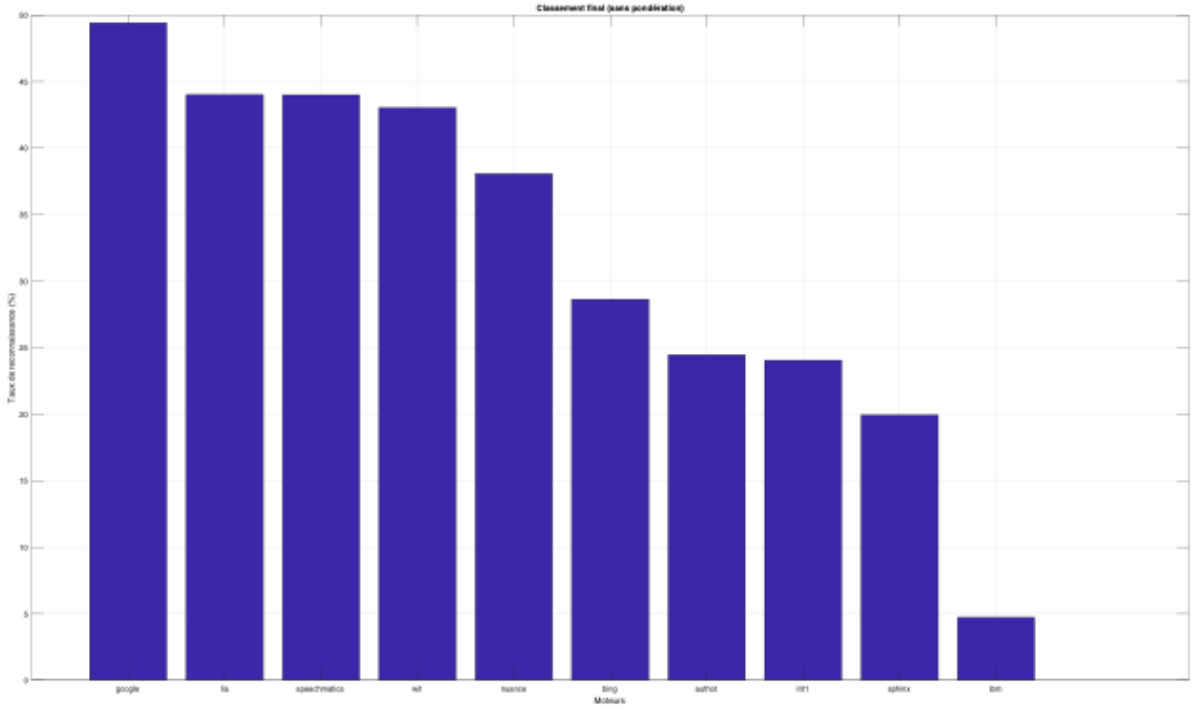
Annexe E

Résultats prestation TTT ASR

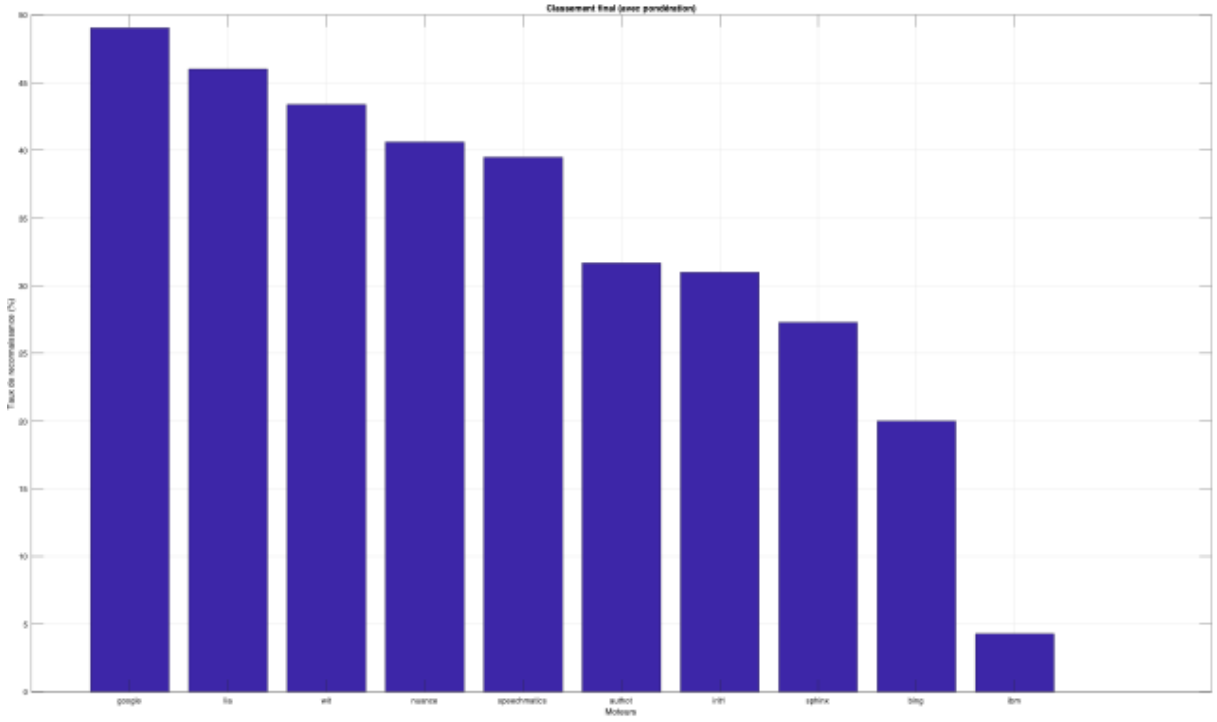
Voici les résultats bruts issus de la prestation TTT ASR (cf. section 5.6.2). Le cahier des charges, le corpus et les systèmes utilisés ont été décrits en section 2.2.

E.1 Résultats généraux

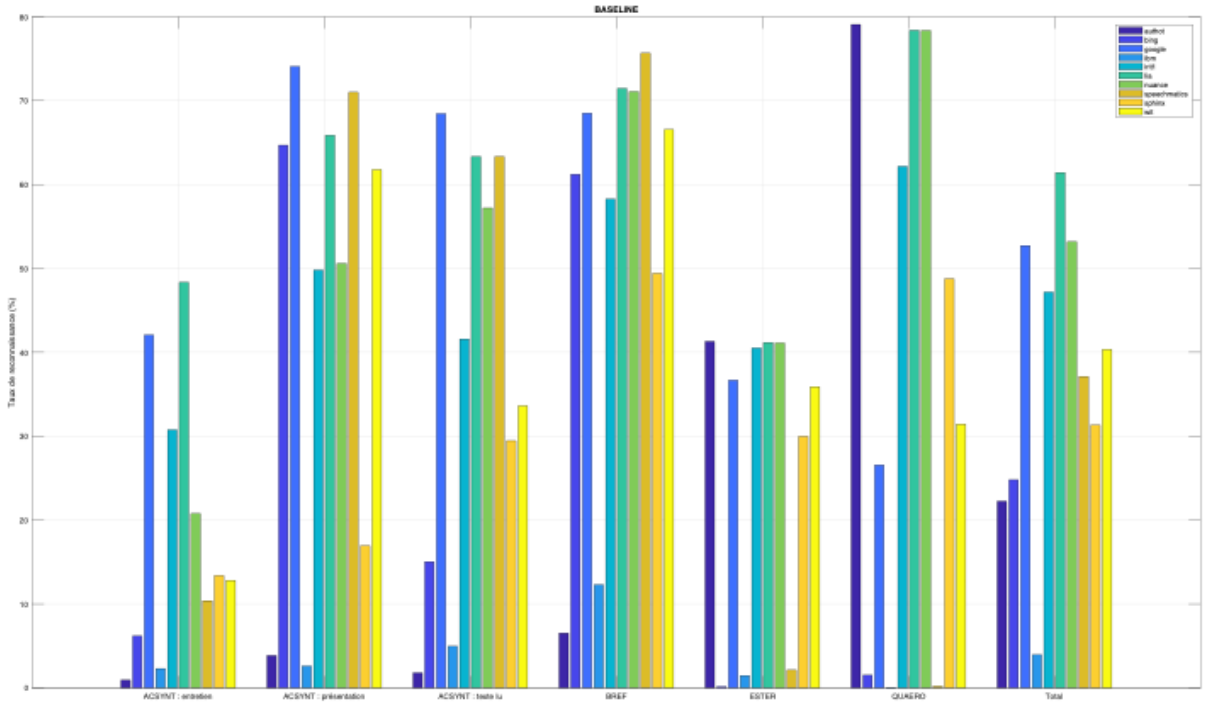
Résultats généraux (sans pondération) :



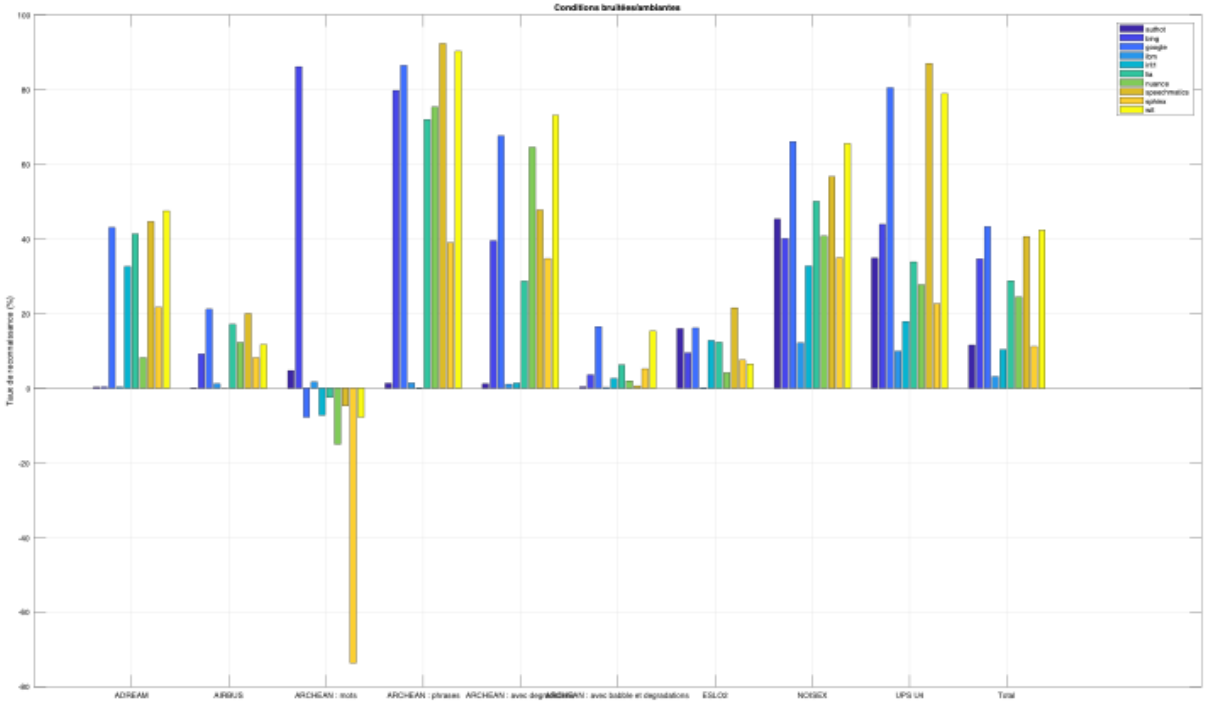
Résultats généraux (avec pondération) :



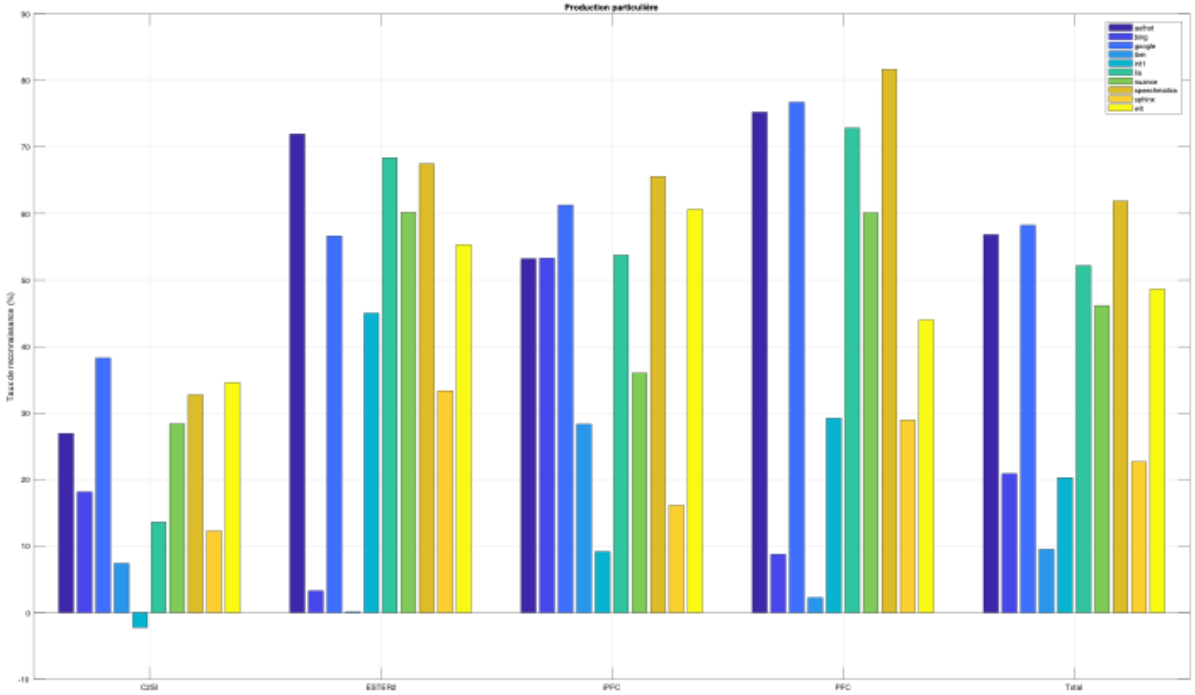
E.2 Résultats parole préparée



E.3 Résultats parole en condition bruitée

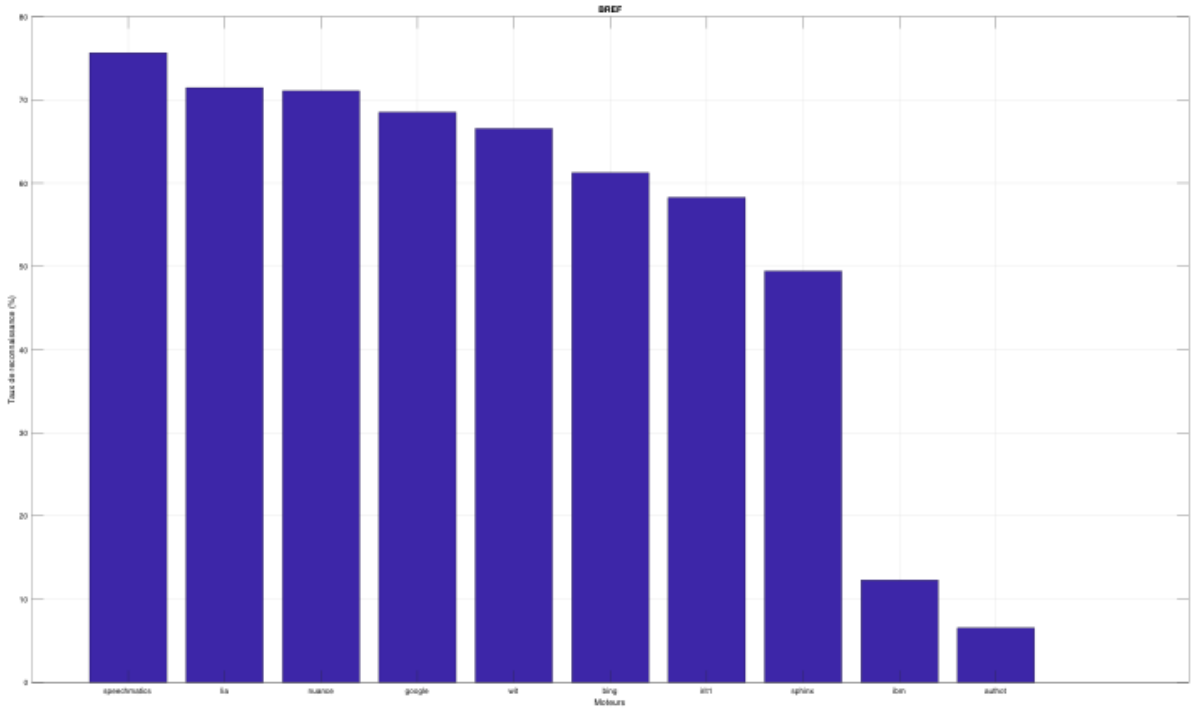


E.4 Résultats parole atypique

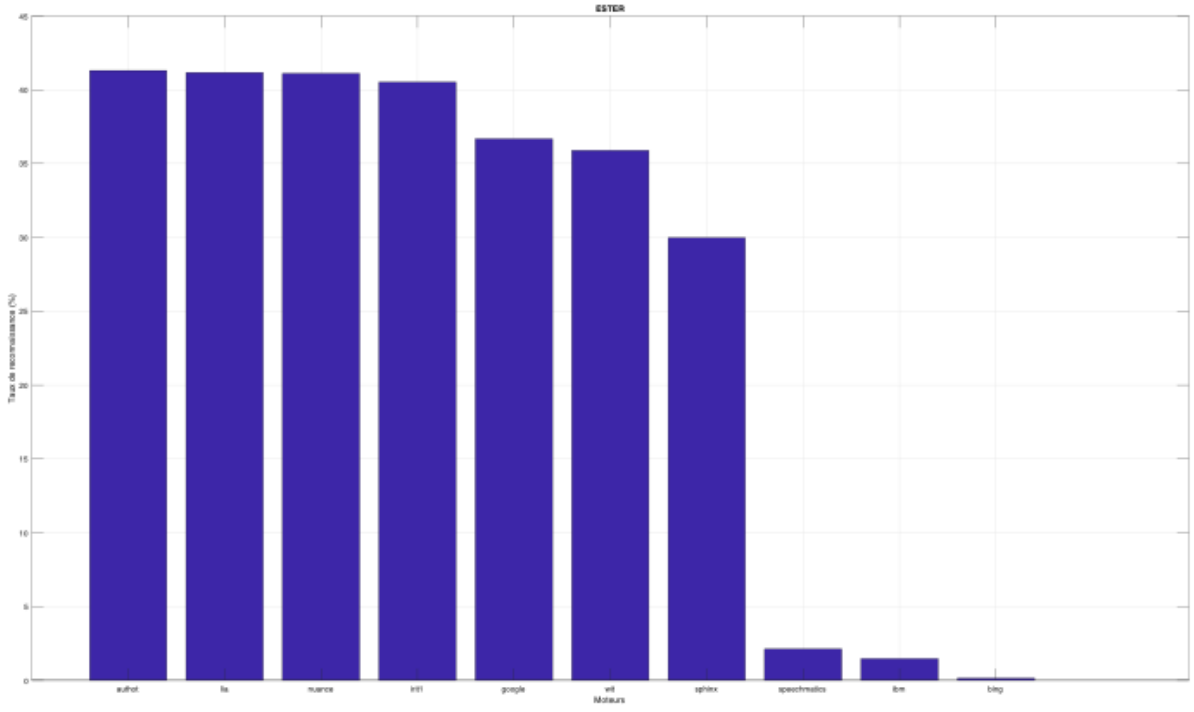


E.5 Résultats par corpus

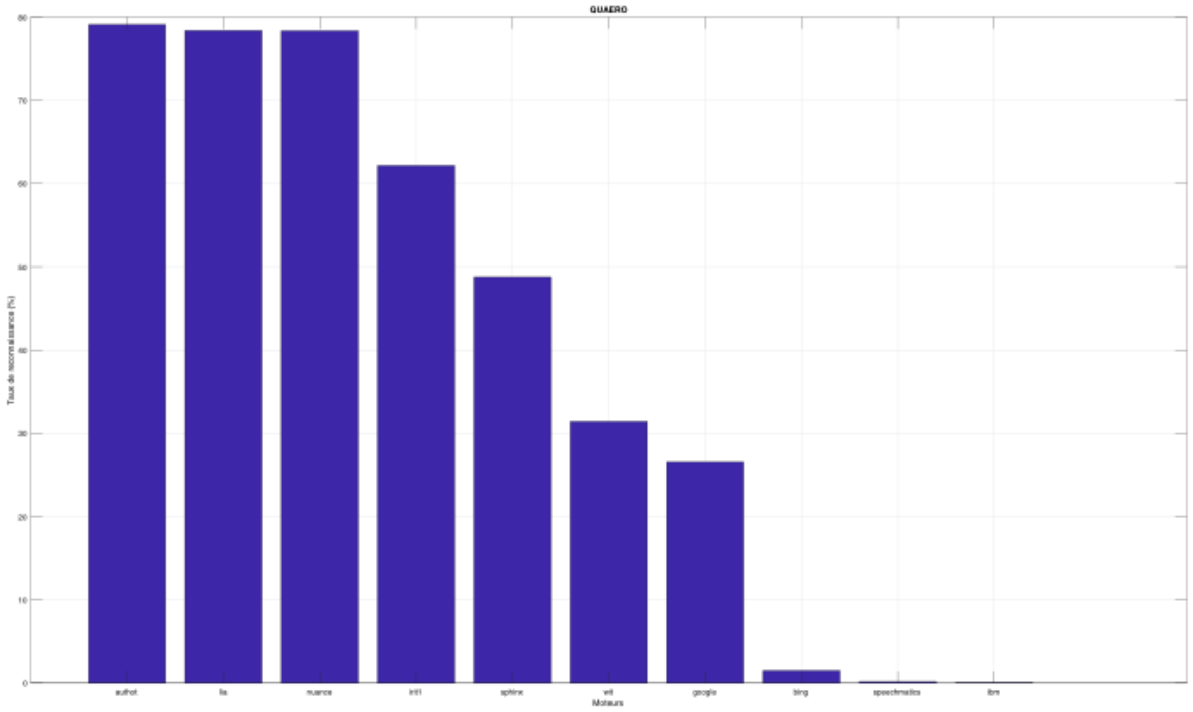
E.5.1 BREF80



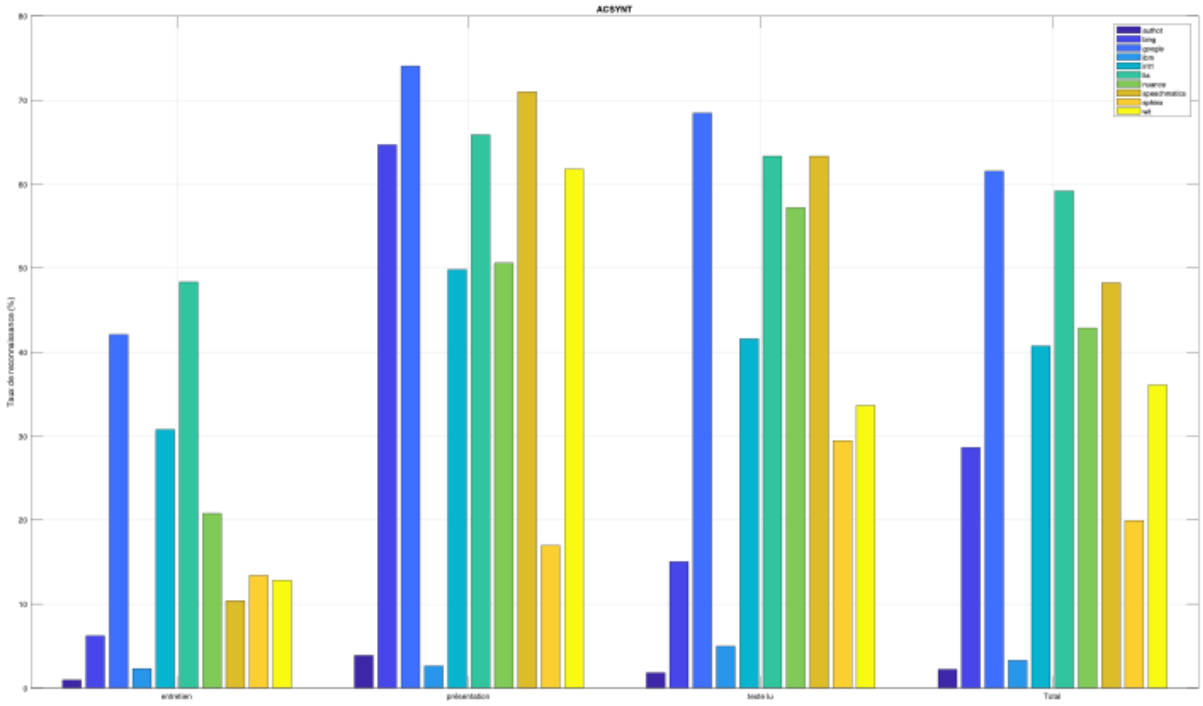
E.5.2 ESTER

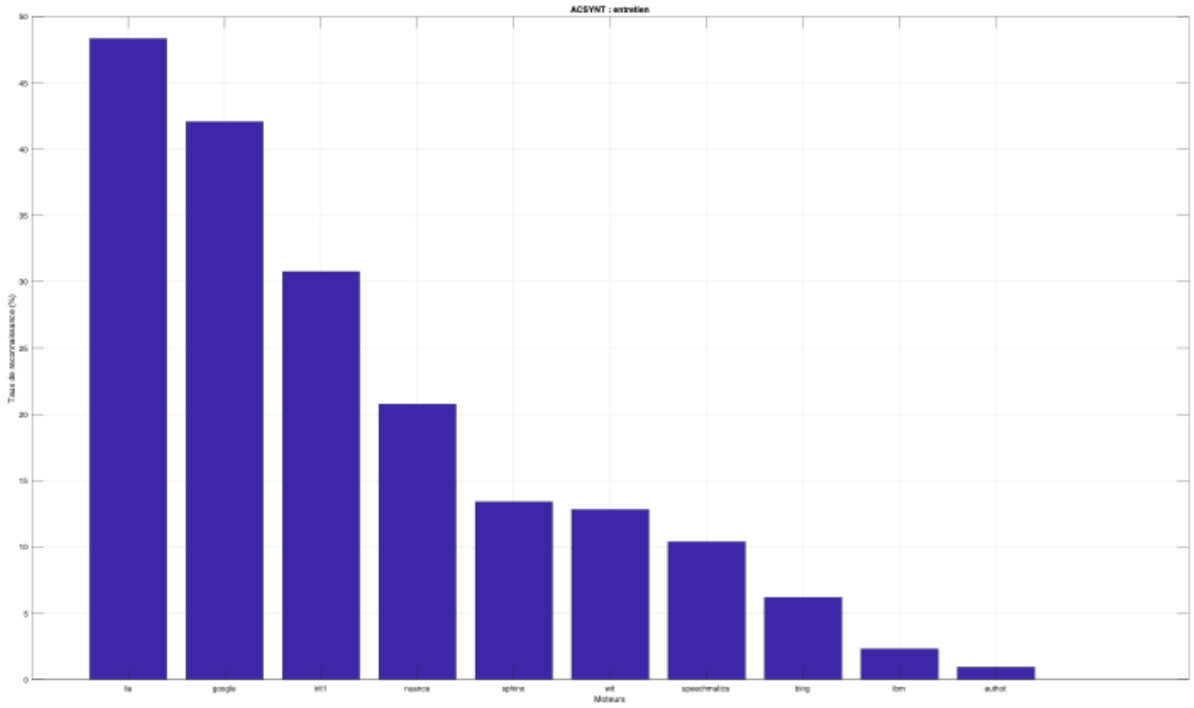


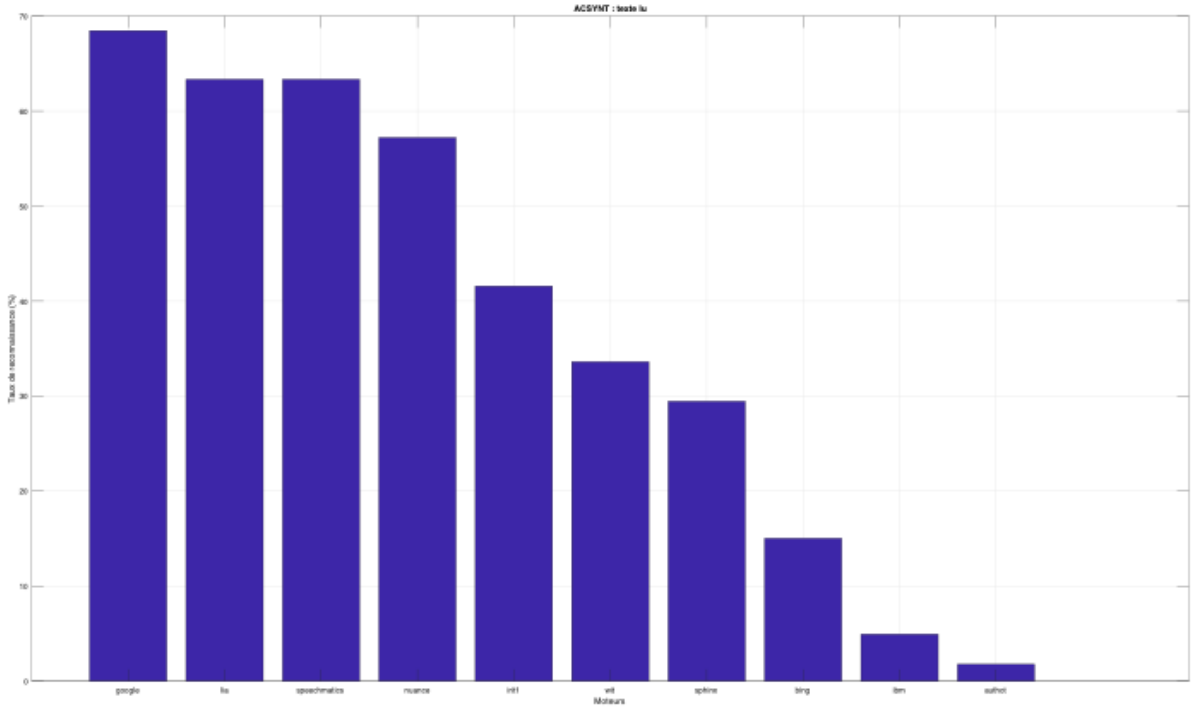
E.5.3 QUÆRO

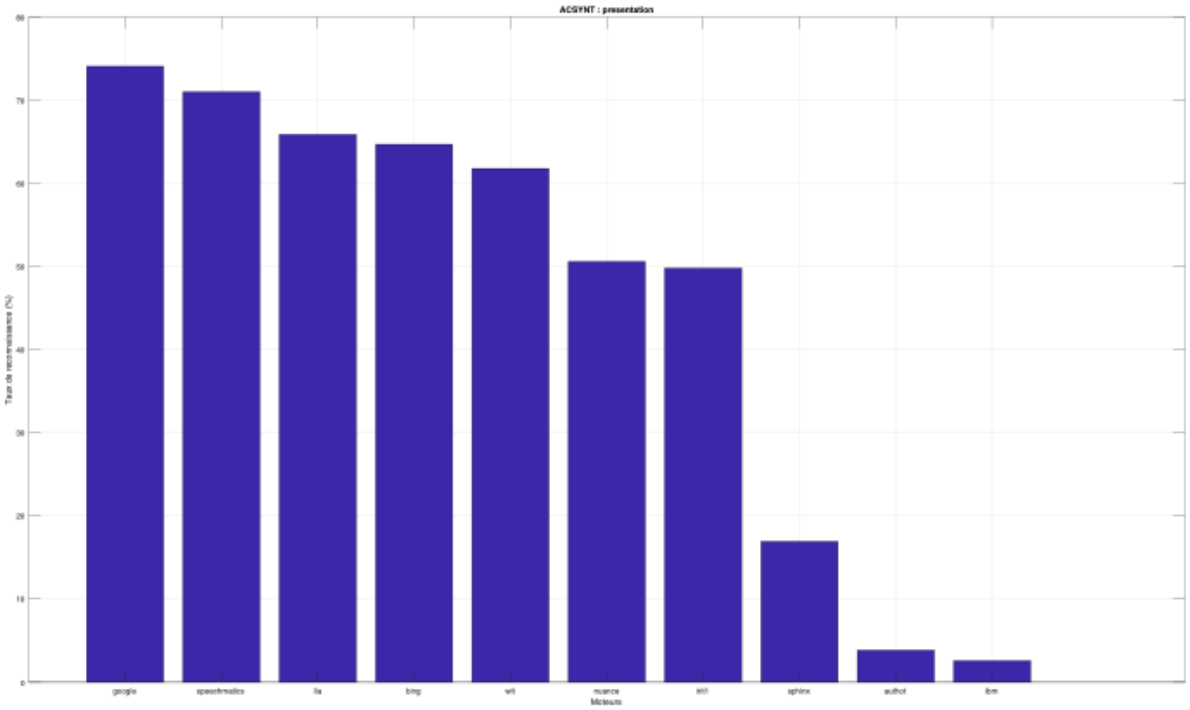


E.5.4 ACSYNT

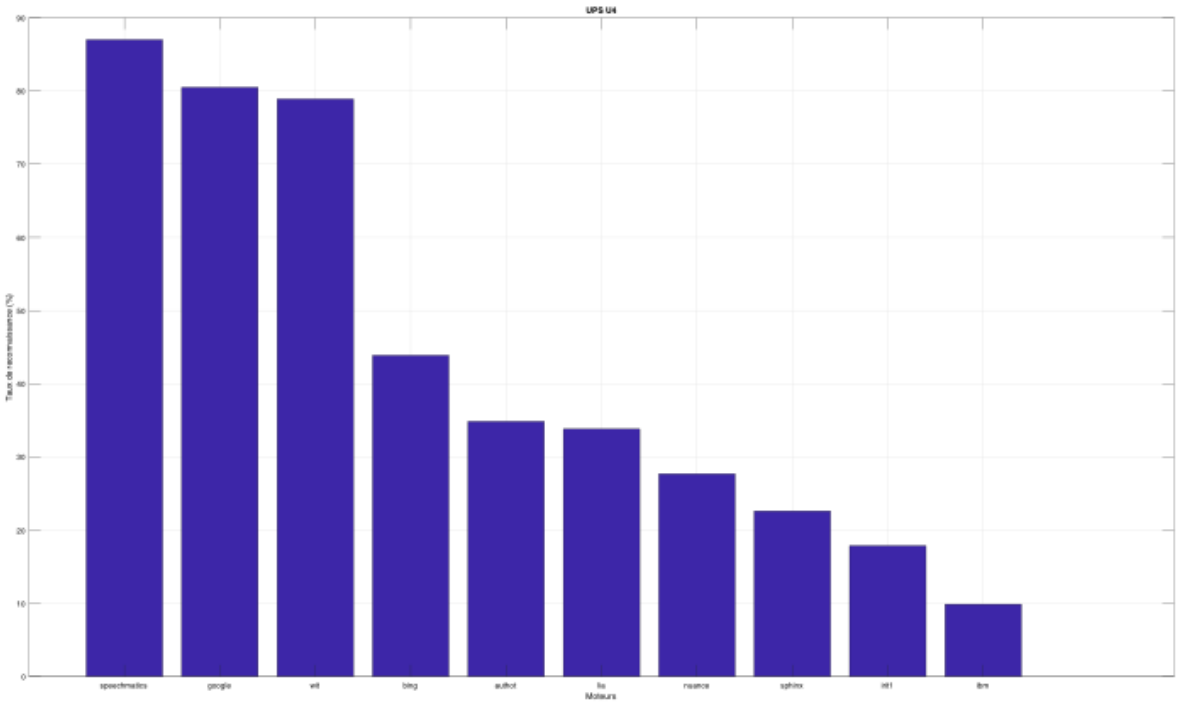



ACSYNT : entretien
Taux de reconnaissance (%)
0
5
10
15
20
25
30
35
40
45
50
lia
google
intl
nuance
sphinx
wit
speechmatics
bing
ibm
authot
Moteurs

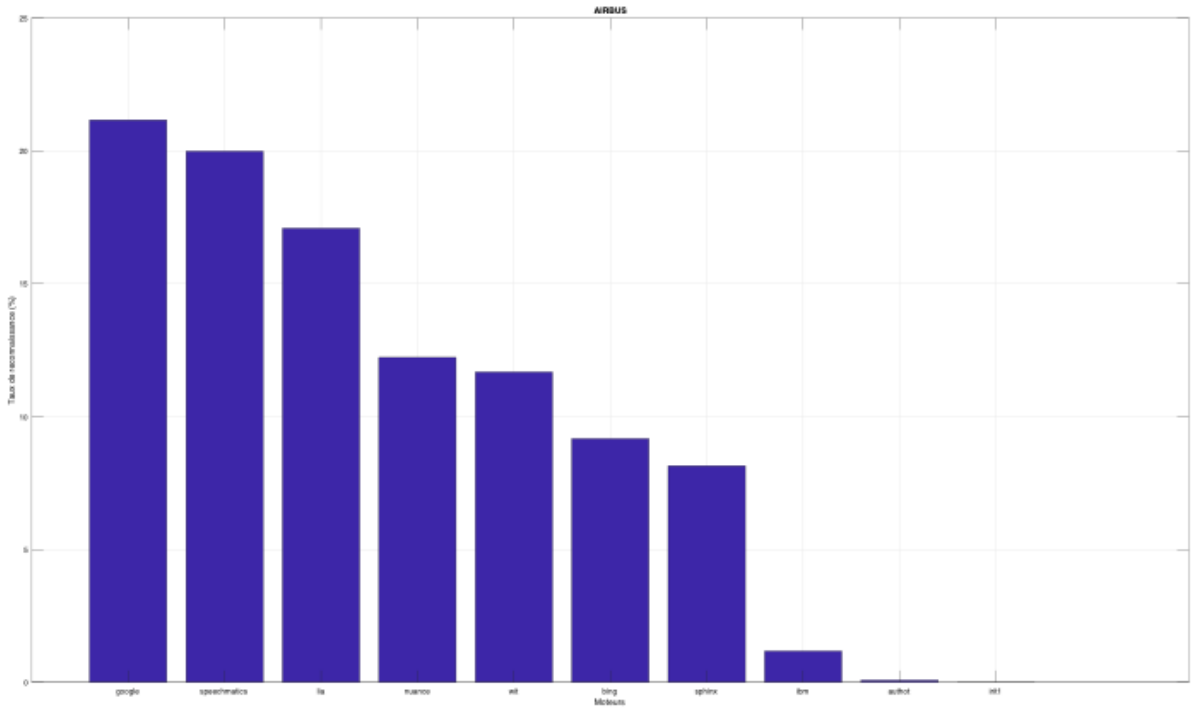

ACSYNT : texte lu
Taux de reconnaissance (%)
0
10
20
30
40
50
60
70
google
lia
speechmatics
nuance
intl
wit
sphinx
bing
ibm
authot
Moteurs


ACSYNT : presentation
Taux de reconnaissance (%)
Moteurs

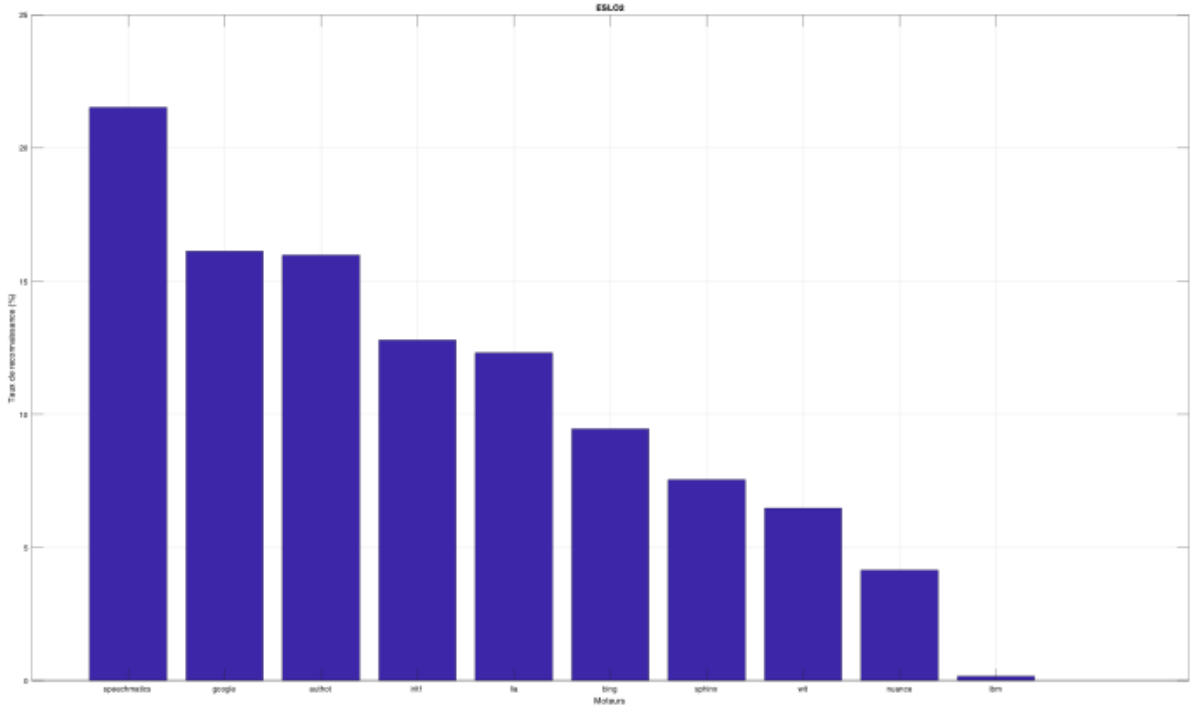
E.5.5 UPS-U4


UPS U4
Taux de reconnaissance (%)
Moteurs

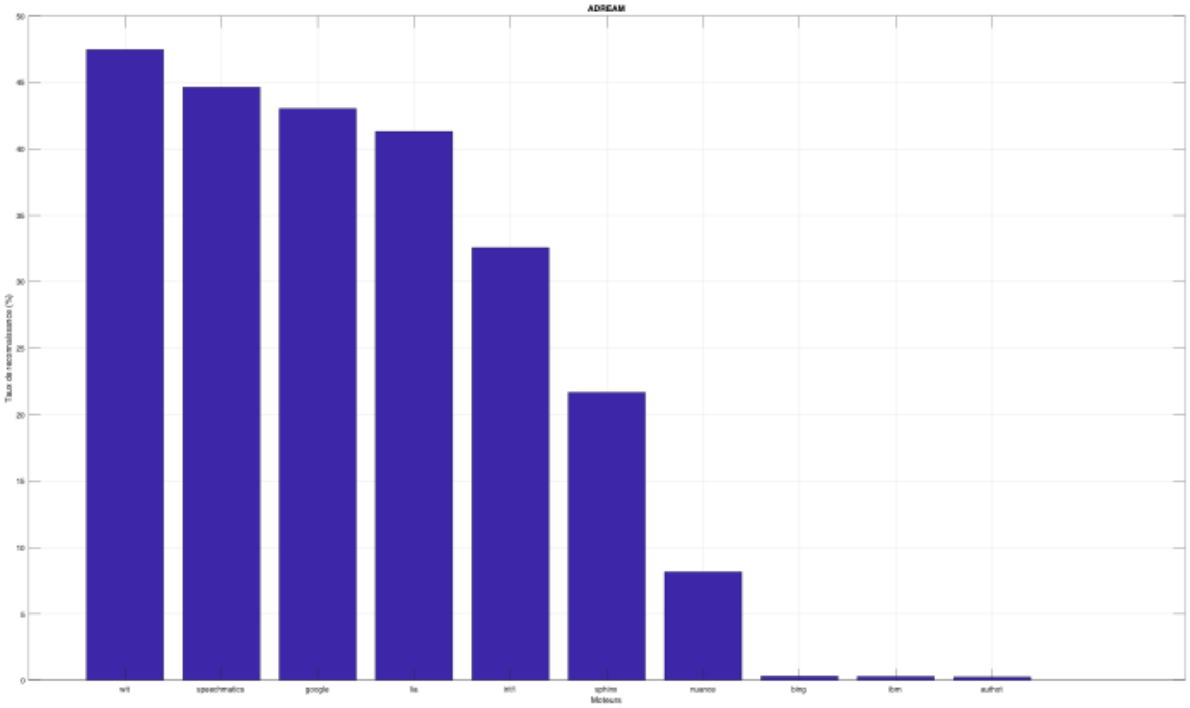
E.5.6 AIRBUS



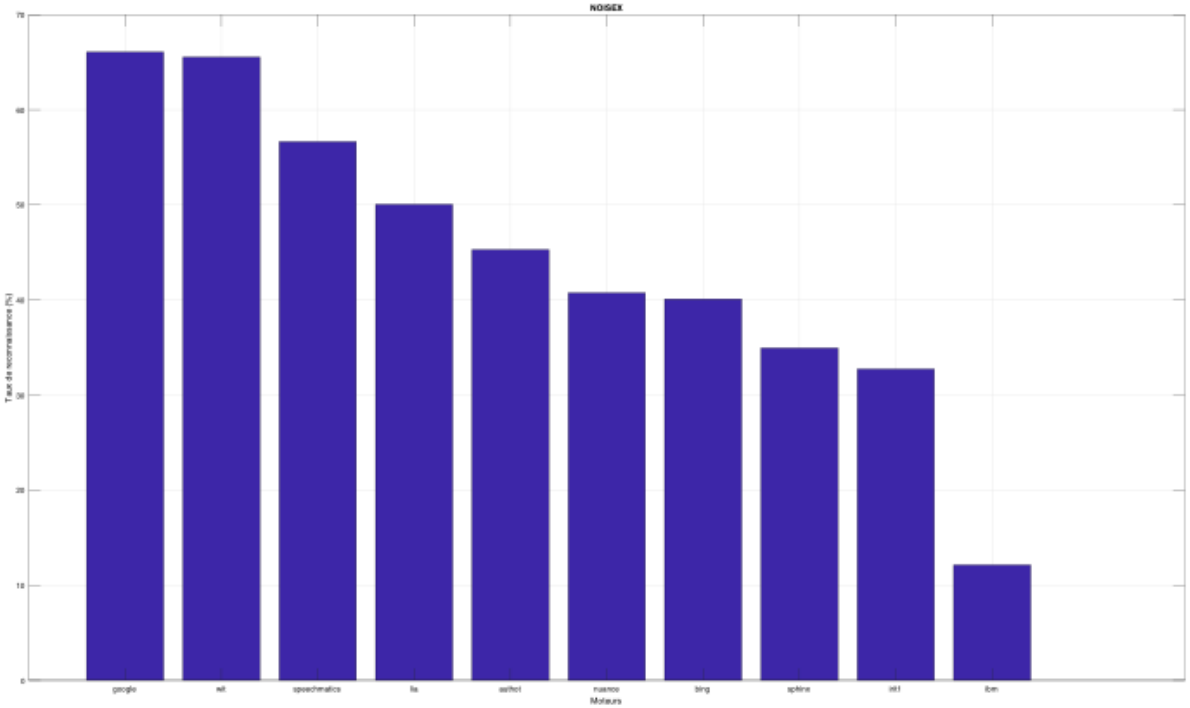
E.5.7 ESLO 2



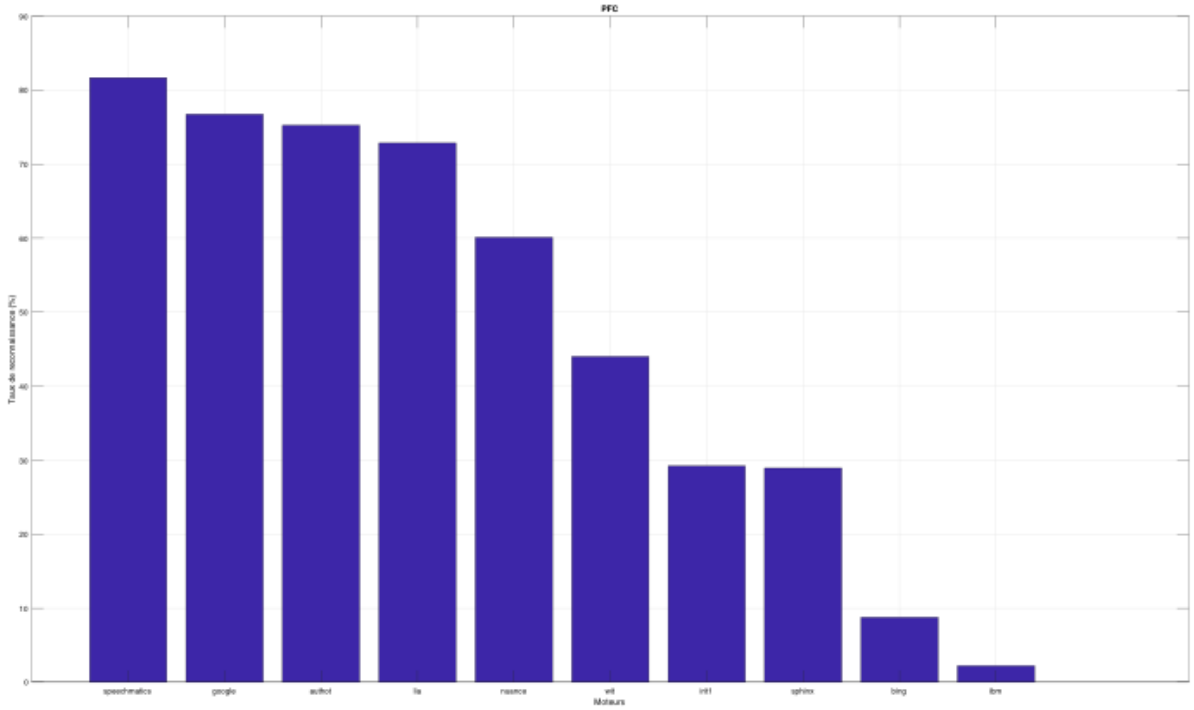
E.5.8 ADREAM



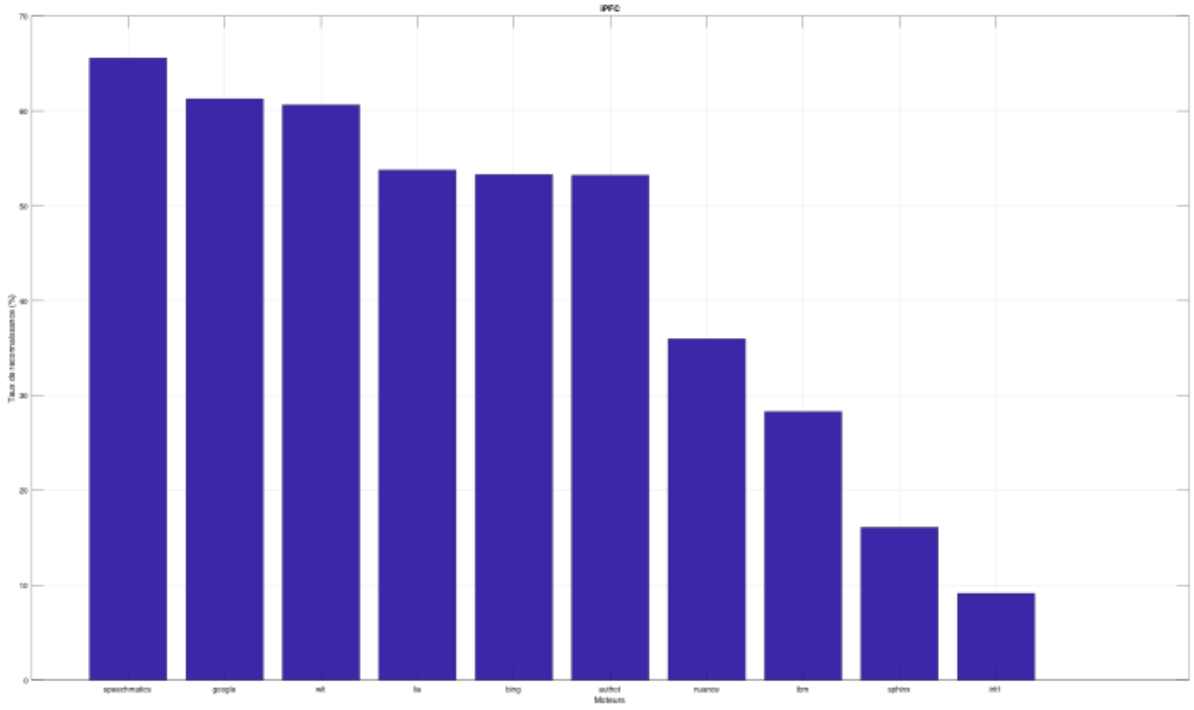
E.5.9 NOISEX



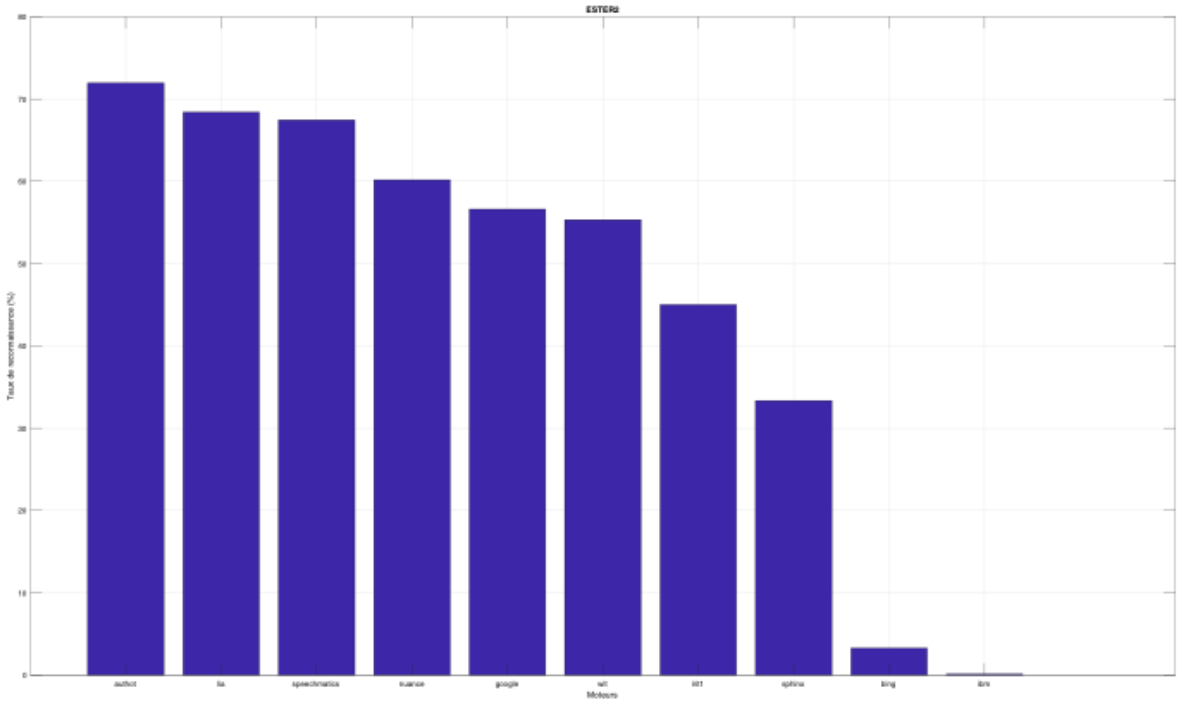
E.5.10 PFC



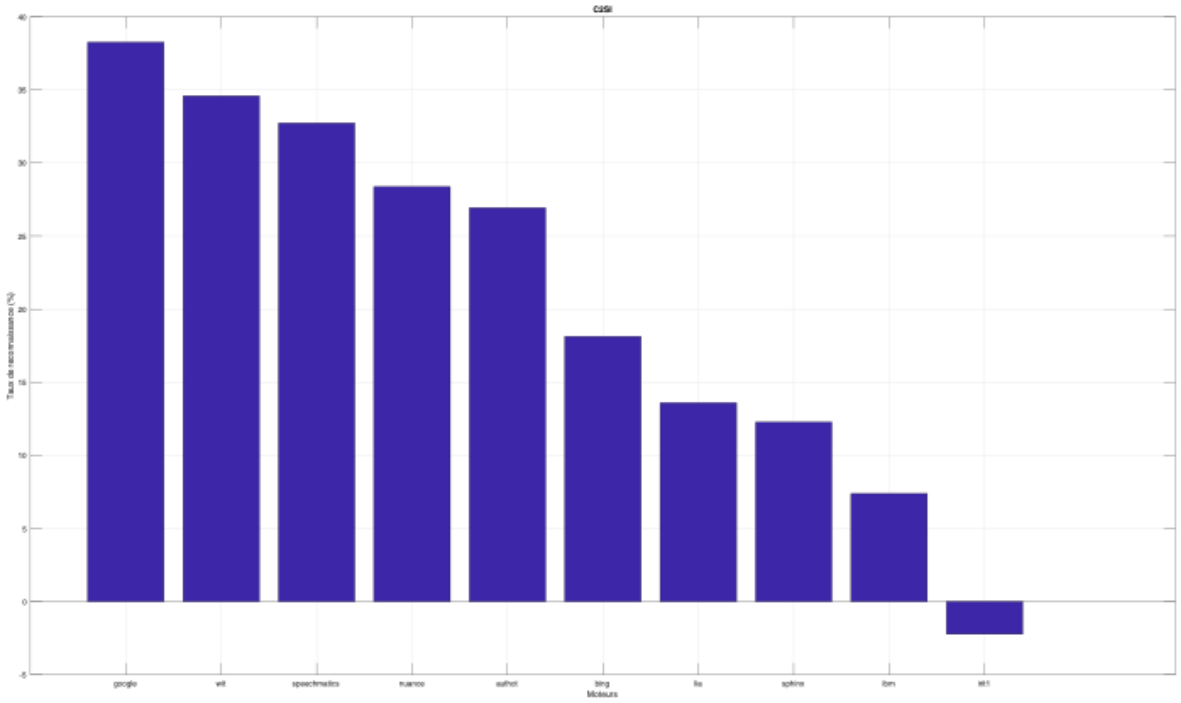
E.5.11 iPFC



E.5.12 ESTER 2



E.5.13 C2SI



Bibliographie

[Arias, 2004] Arias, J. A. (2004). Méthodes à vecteurs de support et indexation sonore. Rapport de stage DEA IIL, Université Paul Sabatier, Toulouse, France.

[Arias, 2008] Arias, J. A. (2008). *Méthodes spectrales pour le traitement automatique de documents audio.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

[Arias et al., 2008a] Arias, J. A., André-Obrecht, R., and Farinas, J. (2008a). Automatic low-dimensional analysis of audio databasis. In *International Workshop on Content-Based Multimedia Indexing (CBMI 2008), London, UK, 18-20/06/08*, pages 556–559. IEEE.

[Arias et al., 2008b] Arias, J. A., André-Obrecht, R., and Farinas, J. (2008b). Représentations de séquences de parole en espaces de faible dimensionalité. In *XXVIIèmes Journées d'Etudes sur la Parole (JEP 2008)*, pages 373–376, Avignon, France. Association Francophone de la Communication Parlée (AFCP).

[Arias et al., 2008c] Arias, J. A., André-Obrecht, R., and Farinas, J. (2008c). Unsupervised signal segmentation based on temporal spectral clustering. In *European Signal and Image Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, 25–29/08/08*, page (online). EURASIP.

[Arias et al., 2006] Arias, J. A., André-Obrecht, R., Farinas, J., and Pinquier, J. (2006). Etude de la réduction non linéaire de la dimension du signal de parole en vue de modélisations discriminatives par SVM. In *Journées d'Etudes sur la Parole*, pages 77–80, Dinard, France. Association Francophone de la Communication Parlée (AFCP).

[Astésano, 2017] Astésano, C. (2017). Le statut de l'accent initial dans la phonologie prosodique du français : enjeux descriptifs et psycholinguistiques. *Habilitation à diriger des Recherches, UT2J.*

[Astesano et al., 2018] Astesano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pont, O., Pouchoulin, G., Puech, M., Robert, D., Sicard, E., and Woisard, V. (2018). Carcinologic Speech Severity Index Project : A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer (regular paper). In Calzolari, N., editor, *Language Resources and Evaluation Conference (LREC 2018), Miyazak, Japon, 07/05/2018-12/05/2018*, pages 4265–4271. European Language Resources Association (ELRA).

[Balaguer, 2021] Balaguer, M. (2021). *Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé.* Theses, Université Paul Sabatier - Toulouse III.

[Balaguer et al., 2019a] Balaguer, M., Boisguerin, A., Galtier, A., Puech, M., Farinas, J., Pinquier, J., and Woisard, V. (2019a). Facteurs influençant l'intelligibilité et la sévérité du trouble chronique de la parole des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. In *Journées de Phonétique Clinique (JPC 2019), Université de Mons, 14–16/05/2019*, pages 23–24. Centre international de Phonétique Appliquée (CIPA).

[Balaguer et al., 2021a] Balaguer, M., Champenois, M., Farinas, J., Pinquier, J., and Woisard, V. (2021a). The (head and neck) carcinologic handicap index : validation of a modular type questionnaire and its ability to prioritise patients' needs. *European Archives of Oto-Rhino-Laryngology*, 278 :1159–1169.

[Balaguer et al., 2019b] Balaguer, M., Farinas, J., Pinquier, J., and Woisard, V. (2019b). Construction of an automatic Carcinologic Speech Severity Index (C2SI) score (regular paper). In *World Congress of the International Association of Logopedics and Phoniatrics (IALP 2019), Taipei, Taiwan, 18–22/08/2019*, page to appear. IALP : International Association of Logopedics and Phoniatrics.

[Balaguer et al., 2021b] Balaguer, M., Gelin, L., Woisard, V., Farinas, J., and Pinquier, J. (2021b). Measurement of speech intelligibility after oral or oropharyngeal cancer by an automatic speech recognition system. In *12th International Workshop MAVEBA (Models and analysis of vocal emissions for biomedical applications)*, Firenze, Italy. Università degli Studi Firenze.

[Balaguer et al., 2021c] Balaguer, M., Gelin, L., Woisard, V., Farinas, J., and Pinquier, J. (2021c). Mesure de l'intelligibilité après cancer oral ou oropharyngé par un système de reconnaissance automatique de la parole. In *1ère Journée Scientifique d'Orthophonie*, Congrès en ligne, France. SURO Société Universitaire de Recherche en Orthophonie.

[Balaguer et al., 2021d] Balaguer, M., Pinquier, J., Farinas, J., Lepage, B., and Woisard, V. (2021d). Construction d'un index holistique d'impact sur la communication des troubles de la parole chez des patients traités pour un cancer oral ou oropharyngé. 15ème conférence francophone d'EPIdémiologie CLINique (EPICLIN 2021) et les 28èmes Journées des Statisticiens des Centres de Lutte Contre le Cancer.

[Balaguer et al., 2021e] Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2021e). Construction d'un index holistique d'impact sur la communication des troubles de la parole chez des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021.

[Balaguer et al., 2023a] Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023a). Analyse des performances de systèmes de reconnaissance automatique de la parole spontanée après cancer oral ou oropharyngé. 9èmes Journées de Phonétique Clinique (JPC 2023). Poster - ISBN : 978-2-917490-35-8.

[Balaguer et al., 2023b] Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023b). Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer : Preliminary validation. *International Journal of Language and Communication Disorders*, 58(1) :39–51.

[Balaguer et al., 2023c] Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023c). Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer : preliminary validation. In *32nd World Congress of the IALP*, Auckland (Nouvelle Zelande), New Zealand. IALP International Association of Communication Sciences and Disorders.

[Balaguer et al., 2023d] Balaguer, M., Pinquier, J., Farinas, J., and Woisard, V. (2023d). Measurement of communication impairment after treatment for oral and oropharyngeal cancer by automatic analyses of spontaneous speech associated with biopsychosocial factors. In *32nd World Congress of the IALP*, Auckland (Nouvelle Zelande), New Zealand. IALP International Association of Communication Sciences and Disorders.

[Balaguer et al., 2020a] Balaguer, M., Pommée, T., Farinas, J., Fichaux-Bourin, P., Puech, M., Pinquier, J., and Woisard, V. (2020a). Validation of the French versions of the Speech Handicap Index and the Phonation Handicap Index in patients treated for cancer of the oral cavity or oropharynx. *International Journal of Phoniatrics, Speech Therapy and Communication Pathology : Folia Phoniatrica et Logopaedica*, 72(6) :464–477.

[Balaguer et al., 2021f] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., and Woisard, V. (2021f). Paramètres perceptifs expliquant la sévérité du trouble de parole mesurée automatiquement en cancérologie ORL. *Rééducation orthophonique*, 286 :1–13. "Rééducation Orthophonique" est une revue scientifique trimestrielle, réalisée par la Fédération Nationale des Orthophonistes. Chaque numéro est thématique.

[Balaguer et al., 2020b] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V., and Speyer, R. (2020b). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Journal of the sciences and specialities of the Head and Neck*, 42(1) :111–130.

[Balaguer et al., 2022] Balaguer, M., Pommée, T., Pinquier, J., Farinas, J., Woisard, V., and Sordes, F. (2022). Development and preliminary validation of the questionnaire 'Evaluation of the constitution of social circles (ECSC)' in patients treated for cancer of the upper aerodigestive tract. *Folia Phoniatrica et Logopaedica*.

[Balaguer et al., 2020c] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., and Woisard, V. (2020c). Functional impact of speech disorders in patients treated for oral or oropharyngeal cancer, assessed by perceptual and automatic measurements (education paper). In *Motor Speech Conference, Hyatt Centric Santa Barbara, California, USA, 20/02/2020-23/02/2020*.

[Balaguer et al., 2019c] Balaguer, M., Woisard, V., Farinas, J., and Pinquier, J. (2019c). Impact du trouble de la production de la parole sur les actes communicationnels de la vie quotidienne dans les cancers de la cavité buccale et de l'oropharynx.

[Baude and Dugua, 2011] Baude, O. and Dugua, C. (2011). (re) faire le corpus d'orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, 1(10) :99–118.

[Beckman, 1996] Beckman, M. (1996). The parsing of prosody. *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, 11 :17–68.

[Bellman et al., 1954] Bellman, R. et al. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6) :503–515.

[Bloch and Joffrin, 2017] Bloch, L. and Joffrin, L., editors (2017). *Comment l'intelligence artificielle va changer nos vies : Voyage au cœur de l'IA*. Libération, France Inter.

[Bloch and Joffrin, 2018] Bloch, L. and Joffrin, L., editors (2018). *Comment l'intelligence artificielle va changer nos vies : L'IA au cœur de l'humain*. Libération, France Inter.

[Boersma and Weenink, 1992] Boersma, P. and Weenink, D. (1992). Praat : Doing phonetics by computer [computer program]. Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

[Bredin et al., 2010] Bredin, H., Koenig, L., and Farinas, J. (2010). IRIT TRECVid 2010 : Hidden Markov Models for Context-aware Late Fusion of Multiple Audio Classifiers. TREC Video Retrieval Evaluation, Gaithersburg, MD, USA, 15–17/11/2010.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42 :335–359.

[Caron, 1989] Caron, J. (1989). *Précis de psycholinguistique*. Presses Universitaires de France.

[CLLE, 2013] CLLE (2013). Acsynt. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[Dard, 1998] Dard, F. (1998). *Du sable dans la vaseline*. Fleuve Noir, Paris.

[Davis and Mermelstein, 1980] Davis, S. and 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.

[De Boissezon et al., 2018] De Boissezon, X., Danet, L., Fabre, C., Farinas, J., Gaume, B., Hathout, N., Ho-Dac, L.-M., Jucla, M., Peran, P., Pierrejean, B., Pinquier, J., and Tanguy, L. (2018). Le projet EvoLex : Aller plus loin dans l'étude de la fluence et de l'accès au lexique. Qui-Quoi-Où de la recherche sur langage, culture et société à Toulouse, CHU Purpan, Pavillon Baudot, 14/05/2018.

[de Calmès et al., 2005] de Calmès, M., Farinas, J., Ferrané, I., and Pinquier, J. (2005). Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire. Atelier ESTER, Avignon, 30–31/03/2005.

[Decroix, 2017] Decroix, F.-X. (2017). *Apprentissage en ligne de signatures audiovisuelles pour la reconnaissance et le suivi de personnes au sein d'un réseau de capteurs ambiants*. Theses, Université Paul Sabatier - Toulouse III.

[Delattre, 1966] Delattre, P. (1966). Les dix intonations de base du français. *French review*, pages 1–14.

[Deléglise et al., 2005] Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2005). The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2005), Lisbonne, Portugal*.

[Destiné, 2011] Destiné, A. G. (2011). Recherche des expressions clés caractéristiques de l'interaction entre locuteurs dans les documents audiovisuels. Rapport de master, Université Paul Sabatier, Toulouse, France.

[Detey and Kawaguchi, 2008] Detey, S. and Kawaguchi, Y. (2008). Interphonologie du français contemporain (ipfc) : récolte automatisée des données et apprenants japonais. *Journées PFC : Phonologie du français contemporain : variation, interfaces, cognition*, pages 11–13.

[Deviren and Daoudi, 2002] Deviren, M. and Daoudi, K. (2002). Apprentissage de structures de réseaux bayésiens dynamiques pour la reconnaissance de la parole. *XXIVemes Journées d'étude sur la parole*.

[Di Cristo, 2000] Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français (seconde partie). *Journal of French language studies*, 10(1) :27–44.

[Di Cristo, 2011] Di Cristo, A. (2011). Une approche intégrative des relations de l'accentuation au phrasé prosodique du français. *Journal of French Language Studies*, 21(1) :73––95.

[Estève et al., 2011] Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., and Farinas, J. (2011). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news (regular paper). In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., and Odijk, J., editors, *Language Resources and Evaluation Conference (LREC 2010), Valletta, Malte, 19/05/10-21/05/10*, pages 1686–1689. European Language Resources Association (ELRA).

[ETSI, 2007] ETSI, E. (2007). Speech Processing, Transmission and Quality Aspects (STQ) ; Distributed speech recognition ; Advanced front-end feature extraction algorithm ; Compression algorithms. Technican report 202 050 v1.1.5, IRIT.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile : the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

[Farinas, 1999] Farinas, J. (1999). Comment modéliser la prosodie dans un but d'identification des langues ? Rencontres Jeunes Chercheurs en Parole (RJC Parole'99), Avignon, 18–19/11/1999.

[Farinas, 2001] Farinas, J. (2001). A differentiated approach and prosody improvements in automatic language identification. Student Forum, IEEE Signal Processing Society, International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2001), Salt Lake City, Etats-Unis, 7–11/05/2001.

[Farinas, 2002] Farinas, J. (2002). *Une modélisation automatique du rythme pour l'identification des langues* . Thèse de doctorat, Université Paul Sabatier, Toulouse, France. (Soutenance le 15/11/2002).

[Farinas, 2006] Farinas, J. (2006). Identification et Classification Automatique de Langues. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 4-6 sept. 2006.

[Farinas, 2007] Farinas, J. (2007). Reconnaissance Automatique de Langues. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 4-6 septembre 2007.

[Farinas, 2008a] Farinas, J. (2008a). Reconnaissance Automatique de Langues. Tutoriel. Ecole Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 24-26/09/2008.

[Farinas, 2008b] Farinas, J. (2008b). Structuration de documents Audio Vidéo. Tutoriel. École Recherche Multimodale d'Information - Techniques et Sciences (ERMITES), Giens, 24-26/09/2008.

[Farinas, 2016] Farinas, J. (2016). IRIT, un acteur de la recherche en Santé et Autonomie, Axe stratégique "Systèmes Informatiques pour la Santé et l'Autonomie", composante gestion de données (Université d'été de la e-santé, Castres, France, 05/07/2016). (Conférencier invité).

[Farinas, 2017] Farinas, J. (2017). GIS Parolothèque : une plateforme de recherche sur des données médicales de parole (Infrastructure pour l'expérimentation en ligne, la science participative/collaborative ou la recherche inter-disciplinaire par les données (InfraMed), Toulouse, France, 04/12/2017). (Conférencier invité).

[Farinas, 2018] Farinas, J. (2018). Voice and Speech Perception : Human vs Automatic : How can Computer Speech Recognition be used ? Tutoriel. 29th Congress of Union of The European Phoniatricians, Helsinki, Finland.

[Farinas, 2019] Farinas, J. (2019). Interests of using Automatic Speech recognition for Speech-Language Therapists (regular paper). In *World Congress of the International Association of Logopedics and Phoniatrics (IALP 2019), Taipei, Taiwan, 18–22/08/2019*. IALP : International Association of Logopedics and Phoniatrics.

[Farinas, 2021] Farinas, J. (2021). Evaluation automatique de l'intelligibilité pour des patients présentant une atteinte de la voix. In *Premier webinaire de Start in Lab Santé 2021*, Toulouse, France. Digital 113.

[Farinas, 2022] Farinas, J. (2022). Jérôme FARINAS, Maître de Conférence UT3 au département SI – Équipe SAMoVA, explique son travail de recherche sur le traitement automatique de la parole. Université Toulouse 3 https://www.youtube.com/watch?v=7Aj8HUVAzgA.

[Farinas, 2023] Farinas, J. (2023). Évaluation objective de la compréhensibilité d'un patient ayant des difficultés à s'exprimer. Institut Carnot Cognition https://www.youtube.com/watch?v=_XOJDHBiWEM.

[Farinas and André-Obrecht, 2000] Farinas, J. and André-Obrecht, R. (2000). Identification Automatique des Langues : variations sur les multigrammes. In *XXIIIème Journées d'Etude sur la Parole (JEP)*, pages 373–376, Aussois, France. Groupe Francophone de la Communication Parlée (GFCP).

[Farinas and André-Obrecht, 2001a] Farinas, J. and André-Obrecht, R. (2001a). Modélisation phonotactique de grandes classes phonétiques en vue d'une approche différenciée en identification automatique des langues. In *XVIIIème colloque GRETSI sur le traitement du signal et des images*, Toulouse, France. Télécommunications Spatiales et aéronautiques (TéSA).

[Farinas and André-Obrecht, 2001b] Farinas, J. and André-Obrecht, R. (2001b). Phonotactic modeling of broad acoustic classes for a differentiated approach of Automatic Language Identification. In *17th International Congress on Acoustics, ICA'2001, Rome, Italie, 02/09/01-07/09/01*, Italie. ASA.

[Farinas et al., 2023] Farinas, J., Astesano, C., and Vaysse, R. (2023). Caractérisation automatique du rythme de la parole. In *Journée scientifique de Toulouse Mind and Brain Institute (TMBI 2023)*, Toulouse, France. Toulouse Mind and Brain Institute.

[Farinas and Fredouille, 2017] Farinas, J. and Fredouille, C. (2017). La place du traitement automatique de la parole (Journée d'Etudes "Intelligibilité de la parole", Maison de la Recherche, D29, 24/03/2017). (Conférencier invité).

[Farinas and Pellegrini, 2018] Farinas, J. and Pellegrini, T. (2018). Analysis of Data Provided for the 2018 AIRBUS Air Traffic Control Challenge (2018 AIRBUS Air Traffic Control Challenge Workshop, Toulouse, France, 04/10/2018). https ://www.irit.fr/recherches/SAMOVA/pagechallenge-airbus-atc-workshop.html.

[Farinas et al., 2019] Farinas, J., Pellegrini, T., and Pinquier, J. (2019). Comparaison de systèmes automatiques de reconnaissance grand vocabulaire appliqué à de la parole pathologique (regular paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14/05/2019-16/05/2019*, pages 53–54. Centre international de Phonétique Appliquée (CIPA).

[Farinas and Pellegrino, 2001] Farinas, J. and Pellegrino, F. (2001). Comparison of two approaches to Language Identification. In *7th International Conference on Speech Communication and Technology, Eurospeech 2001, Aalborg, Danemark, 3–7/09/2001*, volume I, pages 399–402. International Speech Communication Association (ISCA).

[Farinas et al., 2000] Farinas, J., Pellegrino, F., and André-Obrecht, R. (2000). Automatic Language identification : from a phonetic differentiated model to a complete system. In *Workshop on Friendly Exchanging through the Net, COST 254'2000, Bordeaux, 23–24/03/2000*, pages 97–102, ENITA/ENSERB, Bordeaux. C. Germain, E. Grivel and O. Lavialle.

[Farinas et al., 2001] Farinas, J., Pellegrino, F., and André-Obrecht, R. (2001). Automatic Language Identification : From a Phonetic Differentiated Model to a Complete System. 5th workshop on Electronics, Control, Modelling, Measurement and Signals, Toulouse, 30/05–1/06/2001.

[Farinas et al., 2002] Farinas, J., Pellegrino, F., Rouas, J.-L., and André-Obrecht, R. (2002). Merging segmental and rhythmic features for Automatic Language Identification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002), Orlando, USA, 13–17/05/2002*, volume I, pages 753–756. IEEE.

[Farinas et al., 2022] Farinas, J., Pinquier, J., Ghio, A., and Petiot, J. (2022). Plateforme traitement de Parole Atypique. Evènement Kick-off de l'Institut Carnot Cognition 2022. Poster.

[Farinas et al., 2005] Farinas, J., Rouas, J.-L., Pellegrino, F., and André-Obrecht, R. (2005). Extraction automatique de paramètres prosodiques pour l'identification automatique des langues. *Traitement du Signal*, 22(2) :81–97.

[Ferreira, 2021] Ferreira, S. (2021). *"Prédiction a priori de la qualité de la transcription automatique de la parole par l'analyse de l'environnement sonore"*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

[Ferreira et al., 2020a] Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020a). A new set of superimposed speech features to predict a priori the performance of automatic speech recognition systems. Speech In Noise Workshop, Toulouse, France, 10–11/01/2020.

[Ferreira et al., 2020b] Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020b). Analyse de l'effet de la réverbération sur la reconnaissance automatique de la parole. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, pages 235–243, Nancy, France. ATALA.

[Ferreira et al., 2020c] Ferreira, S., Farinas, J., Pinquier, J., Mauclair, J., and Rabant, S. (2020c). Une nouvelle mesure de la réverbération pour prédire les performances a priori de la transcription de la parole. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).*, pages 226–234, Nancy, France. ATALA.

[Ferreira et al., 2018] Ferreira, S., Farinas, J., Pinquier, J., and Rabant, S. (2018). Prédiction a priori de la qualité de la transcription automatique de la parole bruitée (regular paper). In *XXXIIe Journées d'Etudes sur la Parole (JEP 2018), Aix-En-Provence, France, 4–8/06/2018*, pages 249–257. Association Francophone de la Communication Parlée (AFCP).

[Ferreira et al., 2019a] Ferreira, S., Farinas, J., Pinquier, J., and Rabant, S. (2019a). Analyse du bruit pour la prédiction de la qualité de la transcription automatique de la parole (regular paper). In *Colloque sur le Traitement du Signal et des Images (GRETSI 2019), Lille, France, 26/08/19-29/08/19.*

[Ferreira et al., 2019b] Ferreira, S., Pinquier, J., Farinas, J., Mauclair, J., and Rabant, S. (2019b). A new measure to predict the a priori performance of automatic transcription systems on reverberated speech. Speech In Noise Workshop, Ghent, Belgique, 10-11/01/2019.

[Fontan, 2012] Fontan, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication.* Theses, Université Toulouse le Mirail - Toulouse II.

[Fontan et al., 2015a] Fontan, L., Farinas, J., Ferrané, I., Pinquier, J., and Aumont, X. (2015a). Automatic intelligibility measures applied to speech signals simulating age-related hearing loss (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany, 6–10/09/15*, pages 663–667. International Speech Communication Association (ISCA).

[Fontan et al., 2020] Fontan, L., Farinas, J., Segura, B., Stone, M., and Füllgrabe, C. (2020). Using automatic speech recognition to predict aided speech-in-noise intelligibility. Speech In Noise Workshop, Toulouse, France, 10-11/01/2020.

[Fontan et al., 2016] Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., and Aumont, X. (2016). Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA, 8–12/09/16*, pages 650–654. International Speech Communication Association (ISCA).

[Fontan et al., 2017] Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., and Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language and Hearing Research*, 60 :2394–2405.

[Fontan et al., 2013] Fontan, L., Gaillard, P., and Woisard, V. (2013). Comprendre et agir : Les tests pragmatiques de compréhension de la parole et elokanz. *La voix et la parole perturbées*, pages 131–144.

[Fontan et al., 2019] Fontan, L., Laaridh, I., Farinas, J., Pinquier, J., Le Coz, M., and Füllgrabe, C. (2019). Using automatic speech recognition for the prediction of impaired speech identification. Speech In Noise Workshop.

[Fontan et al., 2015b] Fontan, L., Magnen, C., Tardieu, J., Ferrané, I., Pinquier, J., Farinas, J., Gaillard, P., and Aumont, X. (2015b). Comparaison de mesures perceptives et automatiques de l'intelligibilité :

application à de la parole simulant la presbyacousie. *Traitement Automatique des Langues*, 55(2) :151–174.

[Fontan et al., 2014] Fontan, L., Magnen, C., Tardieu, J., and Gaillard, P. (2014). Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit. In *30e édition des Journées d'étude sur la parole (JEP 2014)*, Le Mans.

[Fontan et al., 2015c] Fontan, L., Pellegrini, T., Farinas, J., Mauclair, J., Laborde, V., Sahraoui, H., Aumont, X., Olcoz, J., and Abad, A. (2015c). Vers des outils automatiques pour l'évaluation de locuteurs atypiques. Colloque Interphonologie du Français Contemporain Evaluation de la parole non native et corpus oraux, Paris, 08/12/2015.

[Fournier, 1951] Fournier, J.-E. (1951). *Audiométrie vocale : les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités.* Maloine.

[Galibert et al., 2014] Galibert, O., Leixa, J., Adda, G., Choukri, K., and Gravier, G. (2014). The etape speech processing evaluation. In *Proceedings of 9th International Conference on Language Ressources and Evaluation (LREC 2014)*, pages 3995–3999, Reykjavik, Islande.

[Galliano et al., 2005] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proc. 6th Conference on Speech Communication and Technology (Interspeech 2005)*, pages 1149–1152, Lisbonne, Portugal.

[Galliano et al., 2009] Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Proceedings of 10th Conference on Speech Communication and Technology (Interspeech 2009)*, Brighton, Angleterre.

[Gallois et al., 2021] Gallois, Y., Farinas, J., Paugam, M.-W., Nicolini, L., and Woisard, V. (2021). L'identification automatique des différents bruits de gorge chez le sujet sain : une étude pilote. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021.

[Garofolo et al., 1993] Garofolo, J., Graff, D., Paul, D., and Pallett, D. (1993). Csr-i (wsj0) complete ldc93s6a. *Web Download. Philadelphia : Linguistic Data Consortium*, 83.

[Gaume et al., 2018] Gaume, B., Tanguy, L., Fabre, C., Ho-Dac, L.-M., Pierrejean, B., Hathout, N., Farinas, J., Pinquier, J., Danet, L., Peran, P., De Boissezon, X., and Jucla, M. (2018). Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures (regular paper). In *Natural Language Processing and Cognitive Science, Krakow, Poland, 11–12/09/18*, pages 19–26. Jagiellonian Library.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set : An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

[Ghio and Farinas, 2021] Ghio, A. and Farinas, J. (2021). La laryngophoniatrie du futur. In *Congrès National de la Société Française d'ORL*, Paris, France.

[Ghio et al., 2012] Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., De Looze, C., Robert, D., Viallet, F., and Giovanni, A. (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers ? *Speech Communication*, 54(5) :664–679.

[Ghio et al., 2006] Ghio, A., Teston, B., Viallet, F., Jankowski, L., Purson, A., Duez, D., Locco, J., Legou, T., Pinto, S., Marchal, A., Giovanni, A., Robert, D., Révis, J., Fredouille, C., Bonastre, J.-F., and Pouchoulin, G. (2006). Corpus de parole pathologique, état d'avancement et enjeux méthodologiques. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, 25 :109–126. Autorisation No.3015 : TIPA est la revue du Laboratoire Parole et Langage.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard : Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

[Goldsmith, 1976] Goldsmith, J. (1976). *Autosegmental phonology*. PhD thesis, MIT Press London.

[Google, 2017] Google (2017). Audioset. http://research.google.com/audioset/, accédé le 1er septembre 2023, sous licence Creative Commons Attribution 4.0 International (CC BY 4.0).

[Gravellier et al., 2023a] Gravellier, L., Coz, M. L., Farinas, J., and Pinquier, J. (2023a). Détection automatique de la déglutition dans les signaux d'auscultation cervicale à haute résolution. In *29ième Colloque sur le traitement du signal et des images*, pages 789–792, Grenoble. GRETSI - Groupe de Recherche en Traitement du Signal et des Images. https://gretsi.fr/data/colloque/pdf/2023_gravellier1269.pdf du 6 août au 9 Sept 2023.

[Gravellier et al., 2023b] Gravellier, L., Le Coz, M., Farinas, J., Neveu, F., and Pinquier, J. (2023b). Étude des signaux vibroacoustiques de déglutition chez les sujets sains. 9èmes Journées de Phonétique Clinique (JPC 2023). Poster - ISBN : 978-2-917490-35-8.

[Graves et al., 2006] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

[Graves et al., 2013] Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

[Gravier et al., 2004] Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K., and Choukri, K. (2004). The ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of 4th International Conference on Language Ressources and Evaluation (LREC 2004)*.

[Gutierrez et al., 2004] Gutierrez, J., Rouas, J.-L., Farinas, J., and André-Obrecht, R. (2004). Stratégies de fusion des décisions multiexpert en identification automatique des langues. In *Identification des langues et des variétés dialectales par les humains et par les machines - Modélisation pour l'identification des langues (MIDL)*, pages 71–76, Paris, France. ENST - Télécom Paris.

[Hanusse, 2010] Hanusse, P. (2010). Théorie de l'anharmonicité des phénomenes périodiques non-linéaires. *Résumés des exposés de la 15e Rencontre du Non-Linéaire Paris 2012*, page 127.

[Hanusse, 2011] Hanusse, P. (2011). A novel approach to anharmonicity for a wealth of applications in nonlinear science technologies. In *AIP Conference Proceedings*, volume 1339, pages 303–308. American Institute of Physics.

[Hanusse and Gomez-Gesteria, 1996] Hanusse, P. and Gomez-Gesteria, M. (1996). Towards a normal form for spiral waves. *Physica Scripta*, 1996(T67) :117.

[Hanusse et al., 1994] Hanusse, P., Perez-Muñuzuri, V., and Gomez-Gesteira, M. (1994). Relaxation behavior and pattern formation in reaction-diffusion systems. *International Journal of Bifurcation and Chaos*, 4(05) :1183–1191.

[Heba, 2021] Heba, A. (2021). *Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End*. Theses, Université Paul Sabatier - Toulouse III.

[Hershey et al., 2017] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal processing magazine*, 29(6) :82–97.

[Hirst, 2007] Hirst, D. J. (2007). A PRAAT plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In *16th International Congress of Phonetic Sciences ICPhS XVI*, Saarbrücken, Germany.

[Hixon et al., 2008] Hixon, T. J., Weismer, G., and Hoit, J. D. (2008). *Preclinical speech science : Anatomy, physiology, acoustics, and perception.* Plural Publishing.

[Holz, 2022] Holz, D. (2022). Midjourney. [En ligne ; accédé le 2 mars 2023].

[Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets : Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861. http://arxiv.org/abs/1704.04861.

[Ide and Véronis, 1994] Ide, N. and Véronis, J. (1994). Multext : Multilingual text tools and corpora. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 588–592. Association for Computational Linguistics.

[Jones and Hunter, 1995] Jones, J. and Hunter, D. (1995). Qualitative research : consensus methods for medical and health services research. *Bmj*, 311(7001) :376–380.

[Jun and Fougeron, 2000] Jun, S.-A. and Fougeron, C. (2000). A phonological model of french intonation. In *Intonation*, pages 209–242. Springer.

[Khanagha et al., 2014] Khanagha, V., Daoudi, K., Pont, O., Yahia, H., and Turiel, A. (2014). Non-linear speech representation based on local predictability exponents. *Neurocomputing*, 132 :136–141.

[Laaridh et al., 2018a] Laaridh, I., Tardieu, J., Magnen, C., Gaillard, P., Farinas, J., and Pinquier, J. (2018a). Perceptual and Automatic Evaluations of the Intelligibility of Speech Degraded by Noise Induced Hearing Loss Simulation (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India, 2–6/09/18*, pages 2953–2957. International Speech Communication Association (ISCA).

[Laaridh et al., 2018b] Laaridh, I., Tardieu, J., Magnen, C., Gaillard, P., Farinas, J., and Pinquier, J. (2018b). Évaluations perceptive et automatique de l'intelligibilité de la parole dégradée par simulation de la surdité professionnelle. In *Journées d'Etudes sur la Parole (JEP 2018), Aix-en-Provence, 4-8/06/2018*, pages 392–400. Association Francophone de la Communication Parlée (AFCP).

[Laborde et al., 2016] Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H., and Farinas, J. (2016). Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA, 8–12/09/16*, pages 2686–2690. International Speech Communication Association (ISCA).

[Lamel et al., 2011] Lamel, L., Courcinous, S., Despres, J., Gauvain, J.-L., Josse, Y., Kilgour, K., Kraft, F., Le, V.-B., Ney, H., Nußbaum-Thom, M., et al. (2011). Speech recognition for machine translation in quaero. In *Proceedings of the 8th International Workshop on Spoken Language Translation : Evaluation Campaign*.

[Lamel et al., 1991] Lamel, L. F., Gauvain, J.-L., Eskénazi, M., et al. (1991). Bref, a large vocabulary spoken corpus for french1. *training*, 22(28) :50.

[Lausson, 2023] Lausson, J. (2023). Microsoft intègre chatgpt partout et pourrait bouleverser le web. [En ligne ; accédé le 7 février 2023].

[Le Coz, 2014] Le Coz, M. (2014). *Spectre de rythme et sources multiples : au coeur des contenus ethnomusicologiques et sonores*. Thèse de doctorat, Université des Sciences Sociales, Toulouse, France. (Soutenance le 10/07/2014).

[Lee et al., 2001] Lee, A., Kawahara, T., Shikano, K., et al. (2001). Julius-an open source real-time large vocabulary recognition engine. In *INTERSPEECH*, pages 1691–1694.

[Lefloch, 2005] Lefloch, L. (2005). Etude constractive de décodeurs acoustico-phonétique. Rapport de stage Master Recherche 2IH, Université Paul Sabatier, Toulouse, France.

[Levenshtein et al., 1966] Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8) :707–710.

[Liberman and Prince, 1977] Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2) :249–336.

[Liberman, 1975] Liberman, M. Y. (1975). *The intonational system of English*. PhD thesis, Massachusetts Institute of Technology.

[Lotfian and Busso, 2017] Lotfian, R. and Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4) :471–483.

[Martin and Garofolo, 2007] Martin, A. F. and Garofolo, J. S. (2007). Nist speech processing evaluations : Lvcsr, speaker recognition, language recognition. In *2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, pages 1–7. IEEE.

[Martin and Przybocki, 2003] Martin, A. F. and Przybocki, M. A. (2003). Nist 2003 language recognition evaluation. In *Eighth European Conference on Speech Communication and Technology*.

[Meignier et al., 2008] Meignier, S., Merlin, T., Lévy, C., Larcher, A., Charton, E., Bonastre, J.-F., Besacier, L., Farinas, J., and Ravera, B. (2008). Mistral : plate-forme open source d'authentification biométrique. In *Journées d'Etudes sur la Parole (JEP 2008), Avignon, France, 9–13/06/2008*, pages 81–84. Association Francophone de la Communication Parlée (AFCP).

[Meyer et al., 2010] Meyer, B. T., Jürgens, T., Wesker, T., Brand, T., and Kollmeier, B. (2010). Human phoneme recognition depending on speech-intrinsic variability. *The Journal of the Acoustical Society of America*, 128(5) :3126–3141.

[Michelas and D'Imperio, 2010] Michelas, A. and D'Imperio, M. (2010). Durational cues and prosodic phrasing in french : evidence for the intermediate phrase. In *Speech Prosody*, page 4.

[Middag et al., 2014] Middag, C., Clapham, R., Van Son, R., and Martens, J.-P. (2014). Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer speech & language*, 28(2) :467–482.

[Minier, 2019] Minier, B. (2019). *M, au bord de l'abîme*. XO editions.

[Moisan et al., 2018] Moisan, F., Kab, S., Moutengou, E., Boussac-Zerebska, M., Carcaillon-Bentata, L., and Elbaz, A. (2018). *Fréquence de la maladie de Parkinson en France : données nationales et régionales 2010–2015*. Santé Publique France, Saint Maurice. https://sante.gouv.fr/IMG/pdf/rapport-frequence-maladie-parkinson-france.pdf, accédé le 1er septembre 2023.

[Moore, 2007] Moore, B. C. (2007). *Cochlear hearing loss : physiological, psychological and technical issues*. John Wiley & Sons.

[Morinière et al., 2008] Morinière, S., Boiron, M., Alison, D., Makris, P., and Beutter, P. (2008). Origin of the sound components during pharyngeal swallowing in normal subjects. *Dysphagia*, 23(3) :267–273.

[Mozilla, 2018] Mozilla (2018). Common Voice. https://voice.mozilla.org.

[Muthusamy et al., 1992] Muthusamy, Y. K., Cole, R. A., and Oshika, B. T. (1992). The ogi multi-language telephone speech corpus. In *Second International Conference on Spoken Language Processing*.

[Nilsson et al., 1994] Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2) :1085–1099.

[OpenAI, 2022] OpenAI (2022). Introducing chatgpt. [En ligne ; accédé le 2 mars 2023].

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech : an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

[Parlangeau et al., 2003] Parlangeau, N., Farinas, J., Fohr, D., Illina, I., Magrin-Chagnolleau, I., Mella, O., Pellegrino, F., Pinquier, J., Senac, C., and Smaili, K. (2003). Audio Indexing On The Web : A Preliminary Study Of Some Audio Descriptors. In *7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, Florida, USA, 27–30/07/03*. International Institute of Informatics and Systemics.

[Pellegrini et al., 2019] Pellegrini, T., Farinas, J., Delpech, E., and Lancelot, F. (2019). The Airbus Air Traffic Control speech recognition 2018 challenge : towards ATC automatic transcription and call sign detection. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2019), Graz, Autriche, 15-19/09/2019*, pages 2993–2997, Graz, Autriche. International Speech Communication Association (ISCA).

[Pellegrini et al., 2015] Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Alazard-Guiu, C., Robert, M., and Gatignol, P. (2015). Automatic Assessment of Speech Capability Loss in Disordered Speech. *ACM Transactions on Accessible Computing (TACCESS)*, 6(3) :8 :1–8 :14.

[Pellegrini et al., 2014] Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., and Robert, M. (2014). The Goodness of Pronunciation algorithm applied to disordered speech (regular paper). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH 2014), Singapour, 14–18/09/14*, pages 1463–1467. International Speech Communication Association (ISCA).

[Pellegrino et al., 2002] Pellegrino, F., Chauchat, J.-H., Rakotomalala, R., and Farinas, J. (2002). Can automatically extracted rhythmic units discriminate among languages ? In *International Conference of Speech Prosody 2002, Aix-en-provence, France, 11–13/04/02*, pages 563–566. International Speech Communication Association (ISCA).

[Pellegrino et al., 1999a] Pellegrino, F., Farinas, J., and André-Obrecht, R. (1999a). Comparison of Two Phonetic Aproaches to Language Identification. In Olaszy, G., Németh, G., and Erdohegyi, K., editors, *6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hongrie, 5–9/09/1999*, volume I, pages 399–402. International Speech Communication Association (ISCA).

[Pellegrino et al., 1999b] Pellegrino, F., Farinas, J., and André-Obrecht, R. (1999b). Vowel System Modeling : A Complement to Phonetic Modeling in Langage Identification. In *Multi-Lingual Interoperability in Speech Technology11, (RTO MP-28), Leusden, The Netherlands, 13–14/09/1999*, pages 119–124. RTO/NATO 2000.

[Pellegrino et al., 2000] Pellegrino, F., Farinas, J., and André-Obrecht, R. (2000). Identification Automatique des Langues par une modélisation différenciée des systèmes vocaliques et consonantiques. In *XXIIème congrès francophone AFRIF-AFIA de la Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 131–139, Paris, France. AFRIF-AFIA.

[Pellegrino et al., 2004] Pellegrino, F., Farinas, J., and Rouas, J.-L. (2004). Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech. In Bel, B. and Marlien, I., editors, *International Conference on Speech Prosody 2004, Nara, Japon, 23/03/04-26/03/04*, pages 517–520, ISBN 2-9518233-1-2. ISCA Special Interest Group on Speech Prosody (SproSIG).

[Petiot et al., 2021] Petiot, J., Gravellier, L., Jucla, M., Monnier, N., Quillion-Dupre, L., Péran, P., Danet, L., De Boissezon, X., Farinas, J., and Pinquier, J. (2021). EVOLEX : la reconnaissance vocale au service du diagnostic des dysfonctionnements langagiers. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021.

[Pierrehumbert, 1980] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. PhD thesis, Cambridge : Massachusetts Institute of Technology.

[Pinquier et al., 2019] Pinquier, J., Farinas, J., De Boissezon, X., Peran, P., Danet, L., and Jucla, M. (2019). EVOLEX : apport de la reconnaissance vocale pour le diagnostic des dysfonctionnements cognitifs légers (poster). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14–16/05/2019*, pages 105–106. Centre international de Phonétique Appliquée (CIPA).

[Plakal and Ellis, 2020] Plakal, M. and Ellis, D. (2020). Yamnet. GitHub repository https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, accédé le 1er septembre 2023, sous licence Apache Version 2.0.

[Pommée et al., 2022] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. (2022). Intelligibility and comprehensibility : A Delphi consensus study. *International Journal of Language and Communication Disorders*, 57(1) :21 – 41.

[Pommée et al., 2019] Pommée, T., Mauclair, J., Woisard, V., Farinas, J., and Pinquier, J. (2019). Génération de la « banane de la parole » en vue d'une évaluation objective de l'intelligibilité (regular paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Université de Mons, 14–16/05/2019*, pages 107–108. Centre international de Phonétique Appliquée (CIPA).

[Pons et al., 2018] Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. (2018). End-to-end learning for music audio tagging at scale. In *19th International Society for Music Information Retrieval Conference (ISMIR2018)*.

[Povey et al., 2011a] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011a). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[Povey et al., 2011b] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011b). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

[Rabiner and Juang, 1986] Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP magazine*, 3(1) :4–16.

[Racine et al., 2012] Racine, I., Detey, S., Zay, F., and Kawaguchi, Y. (2012). Des atouts d'un corpus multitâches pour l'étude de la phonologie en l2 : l'exemple du projet «interphonologie du français contemporain»(ipfc). *Recherches récentes en FLE*, pages 1–19.

[Ramesh et al., 2022] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv :2204.06125*.

[Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.

[Reby et al., 2006] Reby, D., André-Obrecht, R., Galinier, A., Farinas, J., and Cargnelutti, B. (2006). Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (Cervus elaphus) stags. *Journal of the Acoustical Society of America (JASA)*, 120(6) :4080–4089.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

[Roger, 2022] Roger, V. (2022). *Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond : d'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique*. Theses, Université Paul Sabatier - Toulouse III. Thèse de doctorat dirigée par Julien Pinquier, Jérôme Farinas et Virginie Woisard.

[Roger et al., 2022a] Roger, V., Farinas, J., and Pinquier, J. (2022a). Deep neural networks for automatic speech processing : a survey from large corpora to limited data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(19) :15. Soumis : 19 novembre 2021, accepté : 15 juillet 2022, publié le 17 août 2022.

[Roger et al., 2021] Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2021). Une méthode automatique non supervisée pour évaluer le score de sévérité de la parole chez les patients traités pour un cancer ORL. Journée AFCP de Phonétique Clinique (JPC 2021), France, Toulouse, 25/05/2021.

[Roger et al., 2022b] Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2022b). Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures. In *34èmes Journées d'Études sur la Parole (JEP 2022)*, pages 117–125, Ile de Noirmoutier, France. Association Française de la Communication Parlée.

[Roger et al., 2022c] Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2022c). Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures. In *Actes XXXIVe Journées d'Études sur la Parole (JEP2022)*, pages 117–125, Noirmoutier, France. Association Française de la Communication Parlée.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

[Rouas and Farinas, 2004] Rouas, J.-L. and Farinas, J. (2004). Comparaison de méthodes de caractérisation du rythme des langues. In *Identification des langues et des variétés dialectales par les humains et par les machines - Modélisation pour l'identification des langues*, pages 45–50, Paris, France. ENST - Télécom Paris.

[Rouas et al., 2002a] Rouas, J.-L., Farinas, J., and Pellegrino, F. (2002a). Merging segmental, rhythmic and fundamental frequency features for automatic language identification. In *Eusipco 2002, Toulouse, 03–06/09/02*, volume III, pages 591–594. EURASIP.

[Rouas et al., 2003a] Rouas, J.-L., Farinas, J., and Pellegrino, F. (2003a). Automatic Modelling of Rhythm and Intonation for Language Identification. In *15th International Congress of Phonetic Sciences (15th ICPhS), Barcelona, Spain, 3-9/08/2003*, pages 567–570.

[Rouas et al., 2004] Rouas, J.-L., Farinas, J., and Pellegrino, F. (2004). Evaluation automatique du débit de la parole sur des données multilingues spontanées. In *XXVe Journées d'Etude sur la Parole (JEP'2004), Fès, Maroc, 19–21/04/2004*, pages 437–440. Association Francophone de la Communication Parlée (AFCP).

[Rouas et al., 2002b] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2002b). Fusion de paramètres rythmiques et segmentaux pour l'identification automatique des langues. In *XXVIéme Journées d'Etude sur la Parole (JEP'2002), Nancy, 24–27/06/2002*, pages 105–108. Groupe Francophone de la Communication Parlée (GFCP).

[Rouas et al., 2003b] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2003b). Modeling Prosody for Language Identification on Read and Spontaneous Speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003), Hong Kong, China, 06–10/04/2003*, volume I, pages 40–43. IEEE.

[Rouas et al., 2005] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R. (2005). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, 47(4) :436–456.

[Sakoe, 1979] Sakoe, H. (1979). Two-level dp-matching–a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(6) :588–595.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1) :43–49.

[Salingre et al., 2017] Salingre, M., Farinas, J., and Rabant, S. (2017). Automatic identification of French regional accent. Rapport de Master IRIT/RR–2017–13–FR, IRIT, Université Paul Sabatier, Toulouse.

[Sanchez-Soto and Farinas, 2007] Sanchez-Soto, E. and Farinas, J. (2007). Irit system description for Nist 2007 Language Recognition Evaluation. NIST evaluation workshop, Orlando, USA, 11–12/12/2007.

[Saon and Chien, 2012] Saon, G. and Chien, J.-T. (2012). Large-vocabulary continuous speech recognition systems : A look at some recent advances. *IEEE signal processing magazine*, 29(6) :18–33.

[Saon et al., 2017] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (2017). English Conversational Telephone Speech Recognition by Humans and Machines. In *Proc. Interspeech 2017*, pages 132–136.

[Schuller et al., 2014] Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., and Zhang, Y. (2014). The INTERSPEECH 2014 computational paralinguistics challenge : cognitive & physical load. In *Proc. Interspeech 2014*, pages 427–431.

[Shigeyoshi et al., 2004] Shigeyoshi, K., Shinya, K., Toshihiko, I., and Campbell, N. (2004). Japanese multext : a prosodic corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal*, pages 2167–2170.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Stojnic et al., 2022] Stojnic, R., Taylor, R., Kardas, M., Saravia, E., Cucurull, G., Poulton, A., and Scialom, T. (2022). Sota papers with code on speech. https://paperswithcode.com/area/speech. Accédé le 18 novembre 2022.

[Synnaeve, 2022] Synnaeve, G. (2022). Wer are we ? https://github.com/syhw/wer_are_we. Accédé le 18 novembre 2022.

[Szymański et al., 2020] Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J., and Carmiel, Y. (2020). Wer we are and wer we think we are.

[Teixeira et al., 2013] Teixeira, J. P., Oliveira, C., and Lopes, C. (2013). Vocal acoustic analysis – jitter, shimmer and hnr parameters. *Procedia Technology*, 9 :1112–1122. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

[Vaillancourt et al., 2005] Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S. D., and Giguère, C. (2005). Adaptation of the hint (hearing in noise test) for adult canadian francophone populations : Adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos. *International Journal of Audiology*, 44(6) :358–361.

[Varga and Steeneken, 1993] Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition : Ii. noisex-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3) :247–251.

[Vaysse, 2023] Vaysse, R. (2023). *Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

[Vaysse et al., 2021a] Vaysse, R., Astesano, C., and Farinas, J. (2021a). Analyse des performances des algorithmes d'estimation de la fréquence fondamentale dans le cadre de la voix pathologique. Séminaire AFCP – Phonétique Clinique. Poster.

[Vaysse et al., 2022a] Vaysse, R., Astesano, C., and Farinas, J. (2022a). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *Journal of the Acoustical Society of America (JASA)*, 152(5) :3091–3101.

[Vaysse et al., 2023] Vaysse, R., Astésano, C., and Farinas, J. (2023). Représentation automatique du rythme de la parole pathologique via le spectre de modulations d'amplitude. In de Recherche en Informatique de Toulouse, I., editor, *9èmes Journées de Phonétique Clinique (JPC 2023)*, pages 59–61, Toulouse, France. Université de Toulouse. ISBN : 978-2-917490-35-8.

[Vaysse et al., 2021b] Vaysse, R., Farinas, J., Astésano, C., and André-Obrecht, R. (2021b). Automatic extraction of speech rhythm descriptors for speech intelligibility assessment in the context of Head and Neck Cancers. In *INTERSPEECH*, Proceeding of Interspeech 2021, Brno, Czech Republic. ISCA.

[Vaysse et al., 2022b] Vaysse, R., Ghio, A., Astésano, C., Farinas, J., and Viallet, F. (2022b). Analyse macroscopique des variations et modulations de F0 en lecture dans la maladie de Parkinson : données sur 320 locuteurs. In *Actes XXXIVe Journées d'Études sur la Parole (JEP2022)*, pages 307–315, Noirmoutier, France. Association Française de la Communication Parlée.

[Wang et al., 2020] Wang, W., Wang, G., Bhatnagar, A., Zhou, Y., Xiong, C., and Socher, R. (2020). An Investigation of Phone-Based Subword Units for End-to-End Speech Recognition. In *Proc. Interspeech 2020*, pages 1778–1782.

[Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1) :36–45.

[Wilson and Cleary, 1995] Wilson, I. B. and Cleary, P. D. (1995). Linking clinical variables with health-related quality of life : a conceptual model of patient outcomes. *Jama*, 273(1) :59–65.

[Witt et al., 1999] Witt, S. M. et al. (1999). *Use of speech recognition in computer-assisted language learning*. PhD thesis, University of Cambridge Cambridge, United Kingdom.

[Witt and Young, 2000] Witt, S. M. and Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3) :95–108.

[Woisard et al., 2018] Woisard, V., Astesano, C., and Farinas, J. (2018). Carcinologic Speech Severity Index Project : A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer. Qui-Quoi-Où de la recherche sur langage, culture et société à Toulouse, CHU Purpan, pavillon Baudot, 14/05/2018.

[Woisard et al., 2021] Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., and Roger, V. (2021). C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55 :173–190.

[Woisard et al., 2022] Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astesano, C., Pinquier, J., and Lepage, B. (2022). Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx : The carcinologic speech severity index. *Head & Neck*, 44(1) :71–88. publié en ligne 2/11/2021.

[Woisard et al., 2019a] Woisard, V., Farinas, J., Ahmed, B., and Wren, Y. (2019a). Place of Automatic Speech Recognition for Assessing Speech Disorders. Présentation orale. Educational Committee for Phoniatrics Committee at International Association of Logopedics and Phoniatrics (IALP).

[Woisard et al., 2019b] Woisard, V., Farinas, J., and Astesano, C. (2019b). Intelligibilité de la parole et qualité de vie. Réflexions à partir des résultats de l'étude « carcinologic speech severity index » (short paper). In DELVAUX, V., HUET, K., PICCALUGA, M., and HARMEGNES, B., editors, *Journées de Phonétique Clinique (JPC 2019), Mons, Belgique, 14-16/05/2019*, pages 15–16. Centre international de Phonétique Appliquée (CIPA). (Conférencier invité).

[Xiong et al., 2016] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv :1610.05256*.

[Young et al., 2002] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2002). The htk book. *Cambridge University Engineering Department*, 3.

[Young and Young, 1993] Young, S. J. and Young, S. (1993). *The HTK hidden Markov model toolkit : Design and philosophy*. University of Cambridge, Department of Engineering Cambridge, England.

Résumé

Ce document présente un panorama de mes travaux en vue de passer l'Habilitation à Diriger les Recherches. La première partie synthétise les travaux en commençant par le fait que la parole, est un signal temporel, mais dont les modélisations ont beaucoup évolué au cours de ces vingt dernières années. Le passage des modélisations stochastiques, qui séparaient la modélisation acoustique et la modélisation du langage à des modélisations qui optimisent en une seule modélisation ces deux parties, en utilisant des représentations neurales profondes a constitué un changement de paradigme profond dans la communauté, qui a permis de rendre accessible les traitements audio au grand public. Mais ce signal est très variable, très perturbé, mais il reste prévisible ! Je ferai le point sur les performances des différents systèmes de transcription et de nos travaux qui cherchent à anticiper les performances des systèmes de reconnaissance. Puis je détaillerai la problématique de la mesure de l'intelligibilité, et cela sous différents aspects : de la collecte de données à la modélisation, et du point de vue de la perception de la parole. J'ai, en effet, abordé ce domaine au départ en cherchant à mesurer l'intelligibilité perçue, dans le but de pouvoir améliorer le réglage de prothèses auditives de personnes atteintes de presbyacousie. Puis je me suis plus particulièrement intéressé à la mesure d'intelligibilité de personnes atteintes de cancers de la tête et du cou, afin de pouvoir produire une mesure objective pour guider les traitements et le suivi médical. La seconde partie détaille les deux axes de recherche sur lesquels je souhaite orienter mes travaux : la modélisation automatique de la prosodie, afin de continuer les travaux sur la représentation du rythme, et d'autre part la modélisation automatique de la déglutition. Cette modélisation permettra une mesure de l'efficacité pharyngo-laryngée, et permettra de prédire les risques de complications chez des patients atteints de dysphagie. La similitude des signaux capturés avec la parole permet d'envisager des approches translationnelles et complémentaires.

Mots-clés: traitement automatique de la parole, modélisation automatique de la prosodie, modélisation de la déglutition avec un dispositif non invasif.

Abstract

This document presents an overview of my work for the Habilitation à Diriger les Recherches. The first part summarizes the work, starting with the fact that speech is a temporal signal whose modeling has evolved considerably over the last twenty years. The shift from stochastic modeling, which separated acoustic and language modeling, to modeling that optimizes both parts in a single model, using deep neural representations, has been a profound paradigm shift in the community, making audio processing accessible to the general public. But this signal is highly variable, highly perturbed, but still predictable ! I'll review the performance of various transcription systems and our work on anticipating the performance of recognition systems. Then I'll look at the problem of intelligibility measurement from various angles : from data collection to modelling, and from the point of view of speech perception. Initially, I approached this field by seeking to measure perceived intelligibility, with a view to improving the fitting of hearing aids for people suffering from presbycusis. I then turned my attention to measuring the intelligibility of people suffering from head and neck cancer, in order to produce an objective measure to guide treatment and medical follow-up. The second part details the two lines of research on which I wish to focus my work : the automatic modeling of prosody, in order to continue the work on rhythm representation, and secondly the automatic modeling of swallowing. This modeling will make it possible to measure pharyngo-laryngeal efficiency, and to predict the risk of complications in patients with dysphagia. The similarity of the signals captured with speech opens up the possibility of complementary, translational approaches.

Keywords: Automatic Speech Processing, Automatic Prosody Modeling, Swallowing Modeling with non-invasive Device