Actes de JEP-TALN-RECITAL 2024. Actes des 35èmes Journées d'Études sur la Parole

Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, José G. Moreno, Julien Pinquier

35èmes Journées d'Études sur la Parole (JEP 2024)
31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024)
26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)

https://jep-taln2024.sciencesconf.org

Actes des 35èmes Journées d'Études sur la Parole

Mathieu BALAGUER, Nihed BENDAHMAN, Lydia-Mai HO-DAC, Julie MAUCLAIR, Jose G MORENO, Julien PINQUIER (Éds.)

Toulouse, France, 8 au 12 juillet 2024

Avec le soutien de

Actes JEP-TALN-RECITAL Toulouse 2024 – ISBN : 978-2-917490-37-2
Actes des 35èmes Journées d'Études sur la Parole – ISBN : 978-2-917490-38-9

Préface

Organisée conjointement par les équipes de recherche IRIS, MELODI et SAMoVA de l'Institut de Recherche en Informatique de Toulouse (IRIT UMR 5505), l'équipe PLC du laboratoire Cognition, Langues, Langage, Ergonomie (CLLE UMR 5263) et l'axe neurocognition langagière, linguistique et phonétique cliniques du laboratoire de NeuroPsychoLinguistique (LNPL URI EA 4156), sous l'égide de l'Association Francophone de la Communication Parlée (AFCP) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence JEP-TALN-RECITAL 2024 regroupe :

— les 35ème Journées d'Études sur la Parole (JEP),
— la 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).

Les conférences TALN et JEP sont un rendez-vous qui offre le plus important forum d'échange francophone aux acteurs universitaires et industriels des technologies de la langue et la parole. Pour cette édition, nous avons plus de 200 inscrits dont une grande partie des étudiants qui construisent le futur de la recherche francophone et assurent le relais de son développement.

En tant que conférenciers invités, nous aurons Véronique HOSTE de l'Université de Ghent, Laurent BESACIER de Naver Labs Europe et Catia CUCCHIARINI de l'Université de Radboud. Ces trois conférenciers qui représentent un large spectre de thématiques entre le texte et la parole vont aborder les dernières avancées de leurs domaines d'expertise.

Cette édition permet aussi de célébrer les 30 ans de TALN. À cette occasion, nous avons dédié une session spéciale dans le programme. La session a comme objectif de rappeler l'historique de la conférence avec l'intervention des participants qui ont participé à sa pérennité afin de mieux transmettre les enjeux de ce rassemblement à la communauté scientifique du traitement automatique des langues naturelles.

En termes des soumissions, pour TALN, 66 articles pour la conférence principale ont été soumis, dont respectivement 18 ont été acceptés pour une présentation orale et 30 pour une présentation sous forme de posters. Également, nous avons reçu 13 résumés des articles publiés lors de conférences internationales qui ont été acceptés pour une présentation en format poster. En ce qui concerne RECITAL, 11 articles ont été soumis dont 7 ont été acceptés. L'ensemble des soumissions acceptées seront présentées sous forme de posters et 3 d'entre elles donneront lieu à une présentation orale. Pour les JEP, 64 articles ont été soumis et 62 ont été acceptés (17 sous forme de présentation orale et 45 sous format poster). L'alternance de sessions communes entre TALN, JEP et RECITAL et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux. En complément de la conférence principale, se tiennent les ateliers "Parole Spontanée", "Défi Fouille de Texte" (DEFT), "Jurisprudence Prédictive" (JP'24), "Evaluation des modèles génératifs" (EvalLLM) et l'activité HackaTAL 2024. Ces événements illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Il convient d'exprimer une profonde reconnaissance envers toutes les personnes qui ont participé à faire vivre la conférence, d'un côté les auteurs de toutes les soumissions et de l'autre les membres de différents comités scientifiques de la conférence. Un remerciement très chaleureux aux relecteurs qui ont accepté une charge importante et qui ont fait des relectures d'urgence afin de faciliter le bon déroulement de la conférence. La bienveillance et l'expertise des comités de programme ont permis la constitution d'un programme riche en thématiques et d'un niveau scientifique correspondant aux attentes de la communauté. Il est également essentiel d'exprimer notre gratitude envers les sponsors et les organisations qui ont subventionné la conférence. Leur soutien financier a permis à cet événement scientifique de se réaliser dans les meilleures conditions, rappelant l'importance des aspects financiers dans la réussite de telles

initiatives. Finalement, un grand merci aux différentes équipes présentes pour le bon fonctionnement, notamment des équipes de l'ATALA, l'AFCP et le CPRS qui nous ont accompagnés dans les différentes étapes de l'organisation.

Jose G Moreno	Lydia-Mai Ho-Dac Nihed Bendahman	Julie Mauclair
Président de TALN	Présidentes de RECITAL	Présidente de JEP

Comités

Comité de relecture

— Adda-Decker Martine, Laboratoire de Phonétique et Phonologie
— Alazard-Guiu Charlotte, Université de Toulouse Jean Jaurès
— Astesano Corine, Université de Toulouse Jean Jaurès
— Audibert Nicolas, Laboratoire de Phonétique et Phonologie
— Balaguer Mathieu, IRIT
— Beautemps Denis, GIPSA-Lab
— Bonastre Jean-François, Université d'Avignon et des Pays de Vaucluse
— Bougarès Fethi, LIUM
— Boula de Mareüil Philippe, LIMSI-CNRS
— Bredin Hervé, IRIT
— Colotte Vincent, Université de Lorraine - LORIA
— Crouzet Olivier, Université de Nantes
— Delais-Roussarie Elisabeth, UMR6310 - LLING & Université de Nantes
— Delvaux Véronique, Institut de Recherche en Sciences et Technologies du Langage, Université de Mons
— Didirkova Ivana, UR1569 TransCrit, Université Paris 8
— Dodane Christelle, Université Paul Valéry, Laboratoire Praxiling UMR5267
— Matrouf Driss, LIA
— Dufour Richard, LS2N - Nantes University
— Elie Benjamin, University of Edinburgh
— Evrard Marc, LISN - Université Paris Saclay
— Farinas Jérôme, IRIT
— Fauth Camille, Université de Strasbourg
— Ferrané Isabelle, IRIT
— Fontan Lionel, Archean Labs
— Fougeron Cécile, Laboratoire de Phonetique et Phonologie
— Fredouille Corinne, CERI/LIA - University of Avignon
— Gao Jiayin, Laboratoire de Phonétique et Phonologie
— Garnier Maëva, GIPSA-Lab
— Gelin Lucile, Lalilo, Paris, France
— Ghio Alain, LPL
— Guinaudeau Camille, University Paris Sud / Japanese French Laboratory for Informatics, CNRS
— Guitard-Ivent Fanny, Praxiling UMR 5267 CNRS
— Harmegnies Bernard, Institut de Recherche en Sciences et Technologies du Langage, Université de Mons
— Henrich-Bernardoni Nathalie, CNRS
— Hueber Thomas, Laboratoire de Phonétique et Phonologie
— Jabaian Bassam, LIA - Université d'Avignon
— Laprie Yves, Loria
— Le Blouch Olivier, Orange
— Lecorvé Gwénolé, Orange
— Leonardo Lancia, Laboratoire Parole et Langage/ (CNRS AMU)
— Lolive Damien, IRISA/ Université Rennes 1
— Marczyk Anna, Université de Toulouse Jean Jaurès
— Mauclair Julie, IRIT
— Meynadier Yohann, LPL

— Michelas Amandine, Laboratoire Parole et Langage/ (CNRS AMU)
— Moreno Jose, IRIT/UPS
— Ouni Slim, LORIA - Université de Lorraine
— Pellegrini Thomas, IRIT
— Perrotin Olivier, Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab
— Pillot-Loiseau Claire, Université Paris 3
— Pinquier Julien, IRIT
— Rilliard Albert, LIMSI-CNRS
— Rouas Jean-Luc, LaBRi, CNRS
— Sahraoui Halima, Université de Toulouse Jean Jaurès
— Savariaux Christophe, GIPSA-lab
— Tahon Marie, LIUM
— Vallée Nathalie, GIPSA-lab UMR CNRS 5216
— Vasilescu Ioana, LIMSI-CNRS
— Wottawa Jane, LIUM
— Yamaguchi Naomi, Laboratoire de Phonétique et Phonologie

Table des matières

Première partie

Articles présentés oralement

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Autisme et compliance phonique

Eva Goeseels[1], Kathy Huet[1], Myriam Piccaluga[1], Virginie Roland[1], Véronique Delvaux[1,2]
(1) Institut de Recherche en Sciences et Technologies du Langage, Service de Métrologie et Sciences du Langage, Université de Mons, Belgique
(2) Fond National de la Recherche Scientifique, Belgique
eva.goeseels@umons.ac.be, veronique.delvaux@umons.ac.be

RÉSUMÉ

Cet article traite de la flexibilité phonétique, définie comme la capacité d'adapter son comportement de parole aux contraintes internes/externes au locuteur et qui pèsent sur la situation de communication. Les individus avec Trouble du Spectre de l'Autisme (TSA), de par leurs caractéristiques, constituent une population pertinente pour étudier la flexibilité phonétique. Cette étude compare la flexibilité phonétique chez des sujets neurotypiques (NT) et avec TSA francophones via un protocole de compliance phonique, nécessitant de devoir répéter des voyelles synthétiques non présentes dans le répertoire vocalique du français. Trois indices ont été calculés pour caractériser la compliance phonique et les résultats montrent des stratégies différentes dans les groupes NT vs. TSA. L'étude met en lumière la préservation possible des compétences sensori-motrices nécessaires à la flexibilité phonétique chez les personnes atteintes de TSA.

ABSTRACT

Autism and phonetic compliance.
This article deals with phonetic flexibility, defined as the ability to adapt speech behavior to internal/external constraints to the speaker that can influence the communication situation. Individuals with Autism Spectrum Disorder (ASD), due to their specific characteristics, constitute a relevant population to study through phonetic flexibility. This study therefore compares phonetic flexibility in neurotypical (NT) and ASD French-speaking participants, using a protocol of phonic compliance that requires the repetition of synthetic vowels not present in the French vowel repertoire. Three indices were calculated to characterize phonetic compliance, and the results show different strategies in the NT vs. ASD groups. The study highlights the possible preservation of the sensory-motor skills necessary for phonetic flexibility in people with ASD.

MOTS-CLES : Flexibilité phonétique, compliance phonique, autisme, sensori-moteur.
KEYWORDS : Phonetic flexibility, phonetic compliance, autism, sensori-motor.

Introduction

Le phénomène de flexibilité phonétique relève de l'aptitude à adapter son comportement de parole aux contraintes internes/externes au sujet qui pèsent sur la situation de communication, et peut se manifester de diverses façons. Une bonne compétence en flexibilité peut par exemple permettre à un locuteur de maîtriser rapidement des patrons prosodiques non familiers en apprenant une langue étrangère ou de compenser efficacement une perturbation des conditions de communication (e.g. : par un *lip tube*, Ménard et al., 2016). Elle peut également se manifester par une bonne adaptation communicative à la situation ou à l'interlocuteur comme adapter sa production à une personne âgée, à un enfant (*infant directed speech* ; Kitamura et al., 2002) ou à un non natif par exemple (*non native directed speech* ; Piazza et al., 2023). Enfin, être flexible phonétiquement peut se traduire par de la

convergence phonétique en communication parlée, ou encore à avoir de bonnes aptitudes à imiter la voix/parole d'autrui et à déguiser sa propre voix (Delvaux et al., 2017).

La flexibilité phonétique est associée à plusieurs autres concepts, dont certains méritent d'être définis dans cet article. Tout d'abord, la convergence phonétique est définie comme le processus par lequel le locuteur-auditeur parvient à rendre ses patrons de production phonétiques et acoustiques plus similaires à ceux de son interlocuteur en communication parlée (Yu et al., 2013). La convergence existe dans sa forme inverse, la divergence phonétique. Aussi, on désigne par « alignement phonétique » le phénomène par lequel les sujets en interaction ne rapprochent pas leurs caractéristiques phonétiques mais évoluent ensemble parallèlement dans la même direction (Lelong, 2012). Ensuite, la compliance phonique, introduite par Delvaux et al., (2014), désigne la capacité intrinsèque à l'individu à produire des sons de parole inhabituels dans sa langue maternelle et contribue en partie à la capacité à acquérir la phonétique/phonologie d'une seconde langue. Enfin, il existe l'imitation phonétique, décrivant le processus se déroulant lorsqu'un individu imite les attributs phonétiques de la parole d'un interlocuteur, le plus souvent consciemment après qu'on lui ait explicitement demandé d'imiter.

Déjà dans les années 1980, la littérature sociolinguistique s'est emparée du phénomène global de convergence (Communication Accommodation Theory (CAT) (Giles et al., 1987, 1991) stipulant qu'il permettrait de minimiser ou de maximiser les distances sociales entre les individus, et de renforcer leurs identités sociales. La convergence entre les personnes ne concerne pas que la parole, elle peut concerner les postures, les gestes ou encore les expressions faciales (Chartrand & Bargh, 1999). Dans la lignée de la CAT, la convergence phonétique est un processus envisagé via l'aspect social des interactions langagières, soit un processus principalement contrôlé, impliquant des mécanismes de « haut niveau » accessibles à la conscience et motivés socialement. D'autres auteurs (Nielsen, 2011 ; Goldinger, 1998) considèrent la convergence phonétique comme un processus avant tout inconscient, d'ordre sensori-moteur, c'est-à-dire largement automatique et de « bas niveau ». L'implication de ces deux types de processus est aujourd'hui largement reconnue mais leur part et rôle respectifs restent à préciser.

Dans la littérature phonétique, la convergence chez les adultes a été étudiée à travers divers tâches et paradigmes. Elle a été étudiée notamment via des paradigmes avec ou sans interaction directe entre locuteurs. Elle a été mesurée au niveau acoustique sur diverses mesures (segmentales (e.g. : formants : Babel, 2009, 2010 ; VOT : Nielsen, 2008 ; Sanchez et al., 2010) et suprasegmentales (rythme, débit, intonation, pauses dans les tours de parole : e.g. : Babel & Bulatov, 2011 ; Kim et al., 2011 ; Pardo et al., 2010, 2013)). La convergence peut également être évaluée via des mesures perceptives, dans des études où des auditeurs externes à la situation de communication sont tenus de juger de la possible similarité entre les productions des locuteurs (e.g. : Goldinger, 1998 ; Pardo et al., 2006). Certains facteurs modulant les effets de convergence phonétique ont pu être mis en évidence tels que le genre (e.g. : Namy et al., 2002), la personnalité, le rôle dans l'interaction, ou encore des facteurs psycholinguistiques comme la fréquence lexicale (e.g. : Babel, 2010).
Pour les individus NT, la tendance à la convergence phonétique voire plus généralement l'aptitude à la flexibilité phonétique est basée sur trois grands domaines de compétences :

1. Les compétences cognitives : les fonctions exécutives, et plus particulièrement, les aptitudes attentionnelles et de flexibilité mentale.
2. Les compétences sociales et communicatives : leur rôle est mis en évidence par l'influence des facteurs sociaux sur l'ampleur de la convergence phonétique comme les rôles sociaux (leader/mené), le genre, la race, la distance linguistique, l'attractivité du locuteur, etc. Elles fondent la compétence pragmatique des locuteurs.

3. Les compétences sensori-motrices : elles constituent le fondement du contrôle moteur de la parole (*feedback*, *feedforward*) et sont associées à des représentations mentales phonétiques riches (détails phonétiques fins) et multimodales.

Les personnes atteintes de Trouble du Spectre de l'Autisme (TSA) sont réputées être déficitaires dans certains de ces domaines mais préservées ou même pouvant être très performantes dans d'autres. Au regard de ces fondements, étudier ce phénomène chez une population atteinte de TSA permettrait donc à la fois de mieux comprendre les mécanismes qui sous-tendent la flexibilité phonétique mais également de mieux comprendre les particularités des personnes avec TSA. En effet, un large consensus dans la littérature (e.g. Demetriou et al., 2018 ; Hill, 2004 ; Lai et al., 2016) indique un déficit des fonctions exécutives dans l'autisme. La flexibilité mentale, notion complexe définie comme la capacité d'adapter ses pensées et ses actions selon les exigences de la situation (Geurts et al., 2008 ; Hill, 2004 ; Rinehart et al., 2001 cités par Conill et al., 2014) étant notamment l'un des déficits exécutifs les plus facilement observables chez les individus TSA : grande rigidité dans les comportements, manque de flexibilité devant une nouvelle tâche, difficulté à gérer plusieurs sources d'informations et à changer de stratégie pendant les activités de la vie quotidienne, difficultés à adapter leur perspective de pensée durant les interactions sociales (théorie de l'esprit : Baron-Cohen et al., 1985). Au niveau de la perception de la parole, les enfants avec TSA pourraient avoir un focus attentionnel atypique à certaines dimensions du signal de parole, préférant les indices prosodiques aux indices lexicaux, alors que le constat inverse est mis en exergue chez les enfants NT (Ploog., 2009, 2010, cité par Hu et al., 2023).

De plus, les compétences sociales et communicatives sont également largement déficitaires dans l'autisme. Il s'agit d'ailleurs du premier critère de diagnostic du TSA selon le DSM-5-TR. De surcroit, ces compétences sont évaluées dans la plupart des tests de diagnostic de l'autisme (e.g. : ADT, ESCP). Mentionnons enfin un déficit de la pragmatique du langage dans le TSA, ces individus ayant de grandes difficultés d'adaptation de leur communication à la situation et au contexte communicatif, ou à adopter le point de vue de leur interlocuteur.

Le dernier fondement (contrôle moteur) chez les personnes avec TSA n'a que peu été étudié dans la littérature. En ce qui concerne la production de la parole, au niveau suprasegmental, la prosodie des personnes TSA peut souvent être qualifiée d'atypique (e.g. Peppé et al., 2007), et leur parole peut être perçue comme monotone, mécanique, bizarre ou exagérée (Baltaxe et Simmons., 1985 ; Lord et al., 1994 cités par Kissine et al., 2021). Certaines études font référence à un déficit phonologique chez les TSA (e.g., Gepner et al., 2002), qui peut être influencé par un traitement perceptif atypique entraînant une sur-catégorisation ou un sur-fonctionnement du phénomène de perception catégorielle (Gepner et al., 2002). D'autres auteurs en revanche proposent que les personnes avec TSA privilégient un traitement "local" de la parole, axé sur les détails phonétiques fins (Shah & Frith, 1983 ; Frith, 1989, cités par Gepner et al., 2002).

Ensuite, à propos d'imitation phonétique, il est important de mentionner la présence d'écholalies chez certains individus avec TSA. Ces répétitions en écho d'énoncés se caractérisent la plupart du temps par une reproduction très proche au niveau prosodique et temporel (débit) de ce qui a été précédemment entendu. Et ce, parfois plusieurs heures (écholalies différées) après avoir entendu l'énoncé répété. Il n'est d'ailleurs pas rare que la qualité phonétique de leurs écholalies soit nettement meilleure que ce qu'ils sont capables de produire de manière spontanée (sans répétition).

Deux études récentes ont étudié précisément le lien entre flexibilité phonétique et autisme. Menées par le même groupe de chercheurs, elles aboutissent pourtant à des constats presque opposés.

Premièrement, Kissine et Geelhand (2019) ont pu mettre en évidence dans leur étude une variabilité moindre dans la production de voyelles de différents types chez des sujets adultes TSA en comparaison à leurs pairs neurotypiques (sur F1, F2, F3, en parole spontanée). Ils ont donc conclu à une production articulatoire « inflexible » chez les TSA, et proposé que celle-ci puisse être en partie responsable de l'impression subjective d'un ton de voix monotone des TSA.

Deuxièmement, Kissine et al. (2021) ont utilisé le paradigme de compliance phonique (Delvaux et al., 2014; voir détails ci-dessous) pour étudier la capacité d'adultes TSA francophones à reproduire fidèlement différentes voyelles réparties sur l'ensemble de l'espace vocalique, y compris de nombreux timbres vocaliques non exploités en français. Les auteurs ont montré que leurs sujets TSA étaient capables de se rapprocher d'une cible vocalique autant que leurs pairs NT en moyenne, mais que par rapport à ceux-ci, ils avaient davantage recours à la stratégie consistant à sélectionner la voyelle la plus proche de la cible dans leur propre inventaire vocalique. L'étude présentée ici vise à reprendre cette étude pionnière (2021) en s'appuyant pour part sur le jeu de données de Kissine et collaborateurs, et en proposant par ailleurs plusieurs ajustements méthodologiques et prolongements, à savoir : (i) l'extension à un groupe de participants féminins ; (ii) de nouvelles mesures de formants supervisées manuellement (vs. totalement automatiques précédemment) ; (iii) une évaluation poussée des stratégies individuelles via le calcul des indices de compliance issus de Delvaux et al. (2014) et la mise en évidence des profils individuels de production.

Méthodologie

Participants et protocole

Quarante et un adultes (15 femmes et 26 hommes) ont participé à cette étude. Parmi eux, 20 personnes avec autisme dont 13 hommes et 7 femmes et 21 personnes neurotypiques dont 13 hommes et 8 femmes. Ces sujets ont pour langue maternelle le français et ont été enregistrés via un micro, intégré à un casque permettant de maintenir la distance entre la bouche du sujet et le micro constante (pour les caractéristiques des sujets et du matériel utilisé, voir Kissine et al., 2021). Trois femmes TSA ont dû être écartées du protocole du fait de la mauvaise qualité de l'enregistrement, de ce fait non exploitable. L'échantillon se compose donc de 38 participants au total.

Ces participants ont été soumis à un protocole de compliance phonique initialement décrit par Delvaux et al. (2014). Ils ont donc d'abord dû produire en lecture un ensemble de 10 voyelles (/a/, /oe/, /i/, /u/, /e/, /ϵ/, /ə/, /o/, /ɔ/, /y/), à 5 reprises (tâche 1). Ces productions sont considérées comme la ligne de base (LDB) de leurs voyelles « naturelles » (orales) du français, et ont été combinées afin de calculer les centroïdes des clusters vocaliques en français de chaque sujet. Ensuite, les participants ont été invités à répéter « le plus fidèlement possible, comme s'il s'agissait d'un son d'une langue étrangère » 4x94 voyelles synthétiques entendues via un casque (tâche 2). Les enregistrements issus des participants masculins sont communs à ceux étudiés par Kissine et al. (2021). A ces enregistrements, ont été ajoutés ceux de 12 femmes pour la présente analyse.

Stimuli, traitements et mesures

L'ensemble de 94 voyelles créé par Delvaux et al. (2014), a été construit en faisant varier les trois premiers formants par pas successifs et égaux sur une échelle Mel (F1, de 250 à 750 Hz ; F2, 800 à 2300 Hz ; F3, de 2200 à 3000 Hz) avec pour ambition de remplir tout l'espace vocalique potentiel. La fréquence fondamentale est restée constante pour tous les stimuli (patron descendant de 110 à 90

Hz) et leur durée totale est de 200 ms. Ces 94 voyelles sont représentées dans la figure 1, en violet. Dans cette figure, les points constituent des voyelles techniquement impossibles à créer puisque les valeurs de F2 et F3 seraient incompatibles, et les croix constituent des combinaisons de F1/F2/F3 impossibles à produire par une personne humaine (Delvaux et al., 2014).

Les productions de chaque sujet ont ensuite été segmentées à l'aide d'un script Praat®, puis les mesures de F1, F2 et F3 (en Hz, puis transformées en mels) ont été récoltées tous les 10% de la durée de chaque production grâce à un second script Praat®. Les paramètres des scripts ont été adaptés à chaque sujet, et les mesures des formants ont été supervisées et corrigées manuellement pour plus d'exactitude de mesure.

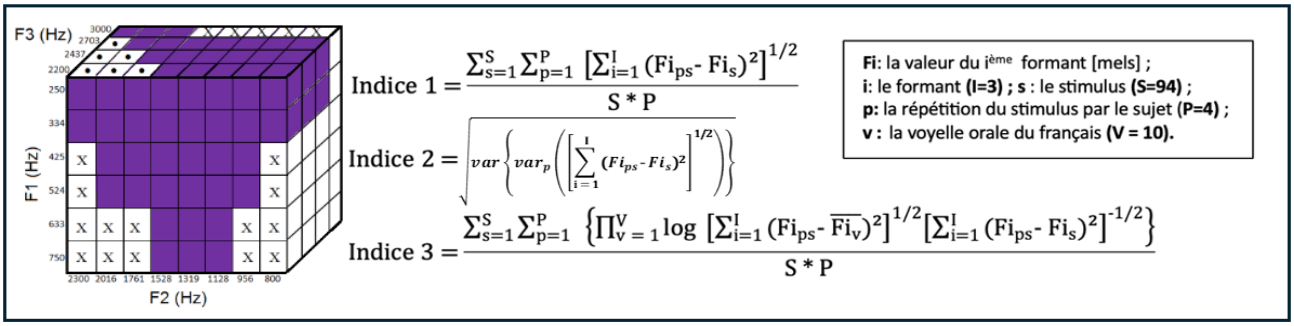


FIGURE 1 : voyelles synthétiques (Delvaux et al., 2014) et formules des indices (Huet et al., 2012)

A partir des valeurs de F1, F2 et F3 (médianes de la série des valeurs mesurées tous les 10%), trois indices ont été calculés (Huet et al., 2012 ; Figure 1, droite) à l'aide d'un script R® pour chaque participant. Ces trois indices sont destinés à évaluer la « compliance » du sujet, c'est-à-dire son niveau de performance dans la tâche de reproduction des voyelles cibles. L'indice 1 est la moyenne des distances euclidiennes entre chaque cible (stimulus) et les productions correspondantes dans l'espace F1-F2-F3. Au plus l'indice 1 diminue, au plus le sujet est jugé compliant. L'indice 2 représente la variance des variances des 4 reproductions d'une même cible. Si cette variance des variances est importante, le sujet est moins compliant dans le sens où il est plus performant (distances à la cible plus homogènes) pour certaines cibles que pour d'autres. L'indice 3 pondère la distance entre la cible et la production par un facteur exprimant l'éloignement de cette production par rapport à sa LDB (distance entre chaque production du participant et les 10 centroïdes de ses clusters vocaliques en français). Autrement dit, pour deux sujets ayant la même distance euclidienne moyenne à la cible, celui qui s'éloigne le plus de ses routines de production/voyelles du français, sera considéré comme meilleur par rapport à l'indice 3. Plus l'indice 3 augmente, plus le sujet est jugé compliant.

Résultats

Distances euclidiennes entre cibles et productions

Les distances euclidiennes entre les productions des sujets et les cibles entendues (dont la moyenne par sujet constitue l'indice 1) ont été exploitées grâce à des analyses de type Anova dans SPSS (v. 26.0) afin d'étudier l'effet du Groupe (TSA/NT), du Sexe (F/M), et du Bloc (4 répétitions des 94 voyelles). Cette analyse met en évidence un effet significatif du groupe ($F(1, 14200) = 83.611$; $p<0.0001$), du sexe ($F(1, 14200) = 525.147$; $p<0.0001$) et de l'interaction entre les deux ($F(3, 14200) = 151.964$; $p<0.0001$). La variable « bloc » et son interaction avec les autres variables ne présente pas d'effet significatif sur la variable dépendante. Le groupe NT (*M = 214.10* mels), a des distances euclidiennes plus grandes que le groupe TSA (*M = 197.24* mels*)* et ce sont principalement les participantes femmes NT qui ont des distances euclidiennes plus importantes.

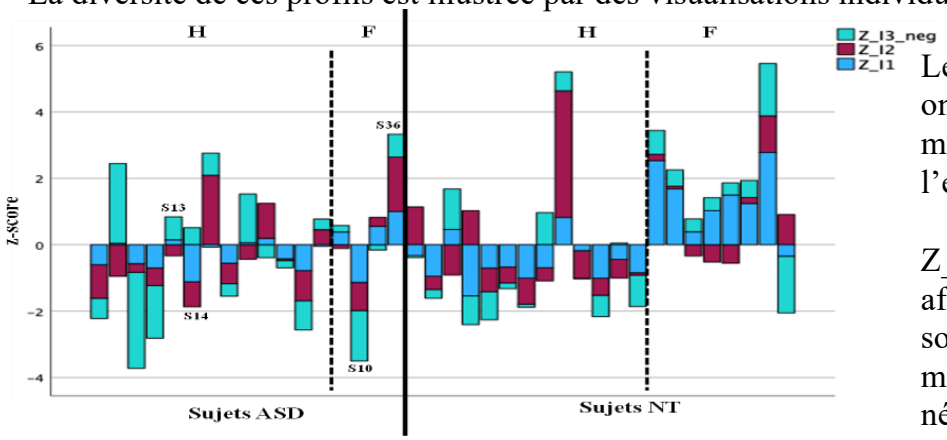
Afin d'objectiver une potentielle variabilité inter-individuelle, susceptible d'influencer les performances des groupes lorsque le nombre de participants est réduit, une nouvelle Anova a été conduite avec la variable « sujet » et le bloc comme variables indépendantes. L'analyse indique un effet significatif du participant ($F(37,14064)=49.681$; $p<0.0001$) sur la distance euclidienne entre cibles et productions. L'examen du graphique 1 (barres pleines) suggère que les différences entre groupes ne peuvent pas être imputées à un ou deux sujets particulièrement atypiques.

Etant donné que les femmes sont « désavantagées » par la tâche (stimuli compatibles avec une voix d'homme en termes de F0 et de fréquences formantiques) et que les effectifs des deux groupes ne sont pas strictement équivalents (ratio homme-femme), nous avons normalisé les données (z-score calculés par rapport à la moyenne et à l'écart type de chaque sujet) et réalisé à nouveau la 1ère analyse statistique. Les résultats de celle-ci se sont tous avérés non significatifs, tant pour le groupe que pour le sexe, ou pour le bloc.

Analyse des indices de compliance 2 et 3

La distance euclidienne, se limitant à mesurer la distance entre la production et la cible, reflète une vision très brute pour évaluer la performance des sujets. Les indices 2 et 3 permettent de nuancer cette évaluation, en examinant tant le lien entre les productions répétées et les routines de production des sujets, que l'homogénéité de la variabilité de leurs productions autour de chaque cible à répéter. Les analyses impliquant l'indice 2 et 3 ne se sont pas révélées significatives dans le sens où il n'existe pas d'effet du groupe et du sexe pour l'I2 et l'I3. Nos deux groupes de sujets performent de manière équivalente pour ces deux indices.

Une seconde analyse a été réalisée, évaluant l'effet du sujet sur ces indices, autrement dit, la variabilité interindividuelle. Comme pour nos précédentes analyses, nous pouvons mettre en évidence une grande variabilité entre nos sujets pour les trois indices (I1, I2, I3 ; cf. graphique 1). La diversité de ces profils est illustrée par des visualisations individuelles dans la figure 2.



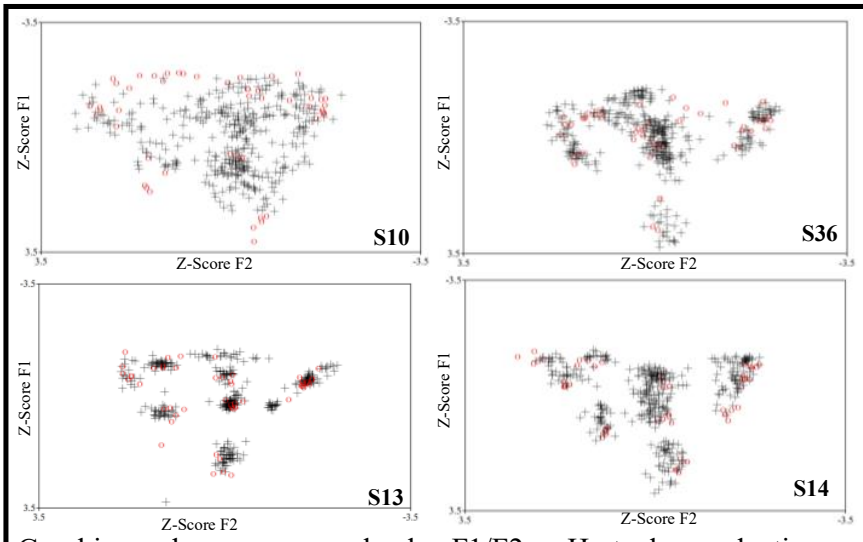
Les z-scores dans ce graphique ont été calculés par rapport aux moyennes et écarts types pour l'ensemble des sujets.

Z_I3 a été transformé en négatif afin qu'une bonne performance soit positive, et qu'une mauvaise performance soit négative, comme pour I1 et I2.

GRAPHIQUE 1 : Indices par sujet (z-scores)

Lorsque nous procédons au calcul des corrélations de Bravais Pearson (r_{bp}) entre nos différents indices, séparément pour chaque groupe, nous pouvons mettre en évidence les éléments suivants : dans le groupe NT les indices 1 et 2 d'une part et 2 et 3 d'autre part ne sont pas corrélés, tandis que les indices 1 et 3 sont significativement reliés (r_{bp} = -0.668; p=0.01) par une corrélation négative (c'est-à-dire dans la direction attendue) ; dans le groupe TSA, on observe une corrélation positive significative (r_{bp} = 0.623 ; p=0.01) entre I1 et I2 et une corrélation négative significative (r_{bp} = -0.493 ; p=0.05) entre I1 et I3.

Illustrations graphiques de différents profils en fonction des indices



Graphiques des z-scores sur le plan F1/F2 en Hertz des productions des sujet (croix noires) et de leur ligne de base (ronds rouges)

Respectivement, de gauche à droite et de haut en bas :

- S10 (TSA-F), très compliant (I1 faible, I2 faible, I3 élevé)
- S36 (TSA-F), peu compliant (I1 élevé, I2 élevé, I3 faible)
- S13 (TSA-M), compliant mais routinier : (I1 faible, I2 faible, I3 faible)
- S14 (TSA-M), peu compliant, peu variable, très routinier (I1 élevé, I2 faible, I3 faible).

FIGURE 2 : illustrations graphiques de différents profils en fonction des indices.

Discussion

Les résultats présentés offrent un aperçu des aptitudes à la flexibilité phonétique chez des adultes atteints du TSA, dans la mesure où elle est évaluée par une tâche de compliance phonique. Tout d'abord, les résultats indiquent que les distances euclidiennes cible-réponse des personnes avec TSA sont en moyenne plus faibles que celles des sujets neurotypiques, indiquant une capacité équivalente, voire légèrement meilleure, à se rapprocher des cibles présentées. Cette constatation suggère une certaine forme de flexibilité phonétique chez les individus TSA, bien que celle-ci puisse être modulée par d'autres facteurs, tels que le sexe. Dès lors, dans une tâche dépourvue de contraintes socio-pragmatiques, c'est-à-dire permettant de neutraliser en partie les deux premiers fondements sur lesquels se basent la convergence phonétique (flexibilité mentale et compétences sociales et communicatives), les individus avec TSA parviennent à se rapprocher de la cible, et même, s'en rapprochent davantage que leurs pairs NT. Nous pouvons donc mettre en évidence que les sujets TSA présentent de bonnes compétences sensori-motrices dans cette tâche très répétitive qui nécessite un traitement "local" de la parole, principalement axé sur les détails phonétiques fins (Shah & Frith, 1983 ; Frith, 1989, cités par Gepner et al., 2002). Ces résultats vont donc à l'encontre de ce que Kissine et Geelhand (2019) ont montré, soit une inflexibilité phonétique chez les personnes avec TSA en parole spontanée.

Au niveau du sexe, les analyses des distances euclidiennes ont révélé des différences significatives entre hommes et femmes dans les deux groupes. Les hommes performent mieux que les femmes, bien que la différence soit plus prononcée dans le groupe NT que dans le groupe TSA. Cela pourrait indiquer des variations dans la manière dont les hommes et les femmes traitent et reproduisent les modèles acoustiques, ce qui mérite une exploration plus poussée. Cependant, les sujets n'étaient pas tous égaux devant la tâche (locuteur modèle = voix d'homme). Dès lors, lorsque les analyses ont été reconduites en utilisant les z-scores afin de neutraliser cette limite, les analyses n'ont pas mis en évidence des différences entre les groupes, ni des différences femmes/hommes, les sujets se conduisant de manière équivalente face aux cibles à répéter.

L’analyse des performances via trois indices différents mais complémentaires permet une évaluation plus complète de la compliance phonique, avec pour objectif de mettre au jour certains mécanismes sous-jacents aux performances évaluées via les distances euclidiennes. Les corrélations entre les indices ont ainsi permis de mettre en évidence des stratégies différentes dans les deux groupes. Dans le groupe TSA, les indices 1 et 2 sont corrélés positivement. Les sujets TSA performants sont donc ceux qui se rapprochent de la cible (I1) et qui ont une variance des variances assez homogène (I2), quelle que soit la position de la cible dans l’espace vocalique. Nos sujets TSA, qui performent globalement mieux au niveau des distances euclidiennes, semblent donc s’appuyer sur des capacités au niveau sensori-moteur. Leur profil de performance, dégagé des caractéristiques spécifiques des cibles, peut faire penser à un processus de « bas niveau », très automatisé et mécanique chez eux. Ces résultats peuvent corroborer ceux de l’étude de Yu et al., (2013) ayant évalué si les « traits autistiques » dans la personnalité d’individus NT pouvaient avoir un effet sur l’imitation phonétique. Selon eux, les sujets NT avec traits autistiques pourraient être meilleurs en traitement phonétique de l’information auditive (sensibles aux différences phonétiques fines). Leurs résultats ont montré que les individus non habitués à un changement constant d’attention (*attention switching* ; cf. sujets avec traits autistiques) pourraient être plus sensibles aux fluctuations phonétiques fines dans le discours de leur interlocuteur, augmentant le risque que les attributs phonétiques soient imités.

Par ailleurs, dans nos deux groupes de participants, les indices 1 et 3 sont corrélés négativement, mais la corrélation est plus importante chez les individus NT. Chez eux davantage que chez les personnes TSA, une bonne performance à la tâche passe par la sélection d’une production de parole éloignée de leurs propres routines du français. Ce résultat fait écho à ce que Kissine et al. (2021) avaient montré dans leur étude (mais ici pour des sujets masculins et féminins), notamment que la stratégie des sujets TSA consistait à se rapprocher de leur LDB afin de se rapprocher de la cible, davantage que chez les personnes neurotypiques.

Enfin, la variabilité interindividuelle observée dans les résultats (voir graphique 1) souligne l'importance de considérer les caractéristiques individuelles dans l'évaluation de la flexibilité phonétique, chaque individu ayant des schémas et stratégies de production distincts, influençant son adaptation aux modèles acoustiques. Cette variabilité souligne la complexité du phénomène de flexibilité phonétique et suggère que des facteurs individuels (compétences sensori-motrices, cognitives, sociales), jouent dans ce processus. Dans le cadre de cette étude, par manque de place, nous n’avons pas pu prendre en considération les différents profils autistiques de nos sujets. Comme l’indique la littérature (Silleresi et al., 2020 ; Eigsti & Shuh, 2017, cités par Ferré et al., 2023), il est important de ne pas considérer la population avec TSA comme un groupe langagier unique. Chaque profil est différent et montre un tableau clinique différent. Il aurait d’ailleurs également été intéressant de considérer les éventuels traits autistiques de nos sujets NT, au vu des résultats de l’étude de Yu et al., 2013. De futures études dans ce domaine gagneraient à tenter de, relier les profils spécifiques de TSA avec les profils diversifiés observés de flexibilité phonétique.

Conclusion

En conclusion, les résultats de cette étude fournissent des éclairages précieux sur la flexibilité phonétique dans le TSA. Bien que ces résultats montrent une certaine capacité d'adaptation aux modèles acoustiques présentés et donnent des informations sur les possibles stratégies adoptées par les personnes TSA et NT, des recherches supplémentaires sont nécessaires pour comprendre pleinement les mécanismes sous-jacents à ce phénomène, ainsi que ses implications pour la prise en charge du TSA.

Références

AMERICAN PSYCHIATRIC ASSOCIATION. DSM-5: diagnostic and statistical manual of mental disorders. 5th ed. Washington (D.C.) London: American Psychiatric Publishing, a division of American Psychiatric Association.

BABEL, M. E. (2009). Phonetic and social selectivity in speech accommodation. PhD Thesis. *Department of Linguistics University of California* : Berkeley, CA: 181 pages.

BABEL, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437-456. doi:10.1017/S0047404510000400

BABEL, M. & BULATOV, D. (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, 1-18.

BALTAXE, C. A. M., & SIMMONS, J. Q. (1985). Prosodic development in normal and autistic children. In E. Schopler & G. Mesibov (Eds.), *Communication problems in autism*. 95–125. Springer: Boston.

BARON-COHEN, S., LESLIE, A. M., & FRITH, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition, 21*(1), 37-46. https://doi.org/10.1016/0010-0277(85)90022-8 ;

CARLIER, S., DUCENNE, L., COLINET, H., PONCIN, F., & DELVENNE, V. (2021). Plus-value de l'implication des enseignants dans le dépistage des troubles du spectre autistique : divergences et convergences d'observations avec les parents et les professionnels sur base de l'Autism Discriminative Tool (ADT). *Neuropsychiatrie de l'enfance et de l'adolescence, 69*(5), 211-220.

CHARTRAND, T. L., & BARGH, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893-910. https://doi.org/10.1037/0022-3514.76.6.893

CONILL, E., STILGENBAUER, J-L., MOUREN, M-C., GOUSSE, V. (2014). Rôle de la flexibilité cognitive dans la reconnaissance d'expressions émotionnelles chez les personnes atteintes de Troubles du Spectre Autistique. *Annales Médico-psychologiques, revue psychiatrique. 172*(5). 392-395. https://doi.org/10.1016/j.amp.2014.05.005

DELVAUX, V., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (06 December 2012). Assessing phonetic compliance [Paper presentation]. ISICS 2012: International Symposium on Imitation and Convergence in Speech, Aix-en-Provence, France.

DELVAUX, V., CAUCHETEUX, L., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (2017). Voice Disguise vs. Impersonation: Acoustic and Perceptual Measurements of Vocal Flexibility in Non Experts. *Interspeech*.

DELVAUX, V., HUET, K., PICCALUGA, M., & HARMEGNIES, B. (2014). Phonetic compliance: a proof-of-concept study. *Frontiers in psychology, 5*.

DEMETRIOU, E. A., LAMPIT, A., QUINTANA, D. S., NAISMITH, S. L., SONG, Y. J. C., PYE, J. E., HICKIE, I., & GUASTELLA, A. J. (2018). Autism spectrum disorders: a meta-analysis of executive function. *Molecular psychiatry, 23*(5), 1198-1204. https://doi.org/10.1038/mp.2017.75

FERRÉ, S., GASNIER, M., GRANDON, B. (2023). Caractérisation de la prosodie du mot chez des enfants avec autisme avec et sans déficit phonologique. Actes des 9èmes Journées de Phonétique Clinique : "Prendre la mesure de la parole", 51-52. *Institut de Recherche en Informatique de Toulouse, 2023.*

FRITH, U. (1989). Autism: explaining the enigma. *Basil Blackwell*, Oxford.

GEPNER, B., MASSION, J., TARDIF, C., GORGY, O., LIVET, M.O., DENIS, D., ROMAN, S., MANCINI, J., CHABROL, B., MESTRE, D., CASTET, É., RONDAN, C., DERUELLE, C., MASSON, G.S., REY, V., SCHMITZ, C., & ASSAIANTE, C. (2002). L'autisme : une pathologie du codage temporel ?

GEURTS, H. M., VAN DEN BERGH, S. F., & RUZZANO, L. (2014). Prepotent response inhibition and interference control in autism spectrum disorders: two meta-analyses. *Autism Res 2014, 7*, 407-420.
GILES, H., MULAC, A., BRADAC, J., & JOHNSON, P. (1987). Speech accommodation theory: The first decade and beyond. *Communication Yearbook*. M. L. McLaughlin. London, UK, Sage Publishers. *10*, 13-48.
GILES, H., COUPLAND, J., & COUPLAND, N. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics.* 1-68. Cambridge, UK: Cambridge University Press.
GOLDINGER, S. D. (1998). "Echoes of echoes? An episodic theory of lexical access." *Psychological Review, 105,* 251-279
GUIDETTI, M., TOURRETTE, C. (2009). ECSP - Echelle d'évaluation de la communication sociale précoce [Matériel, test, mallette pédagogique]. *Eurotests*. Paris [France]
HILL, E. L. (2004). Executive dysfunction in autism. *Trends in cognitive sciences, 8*(1), 26-32. https://doi.org/10.1016/j.tics.2003.11.003
HU, A., QI, Z., AND FRANICH, K. (2023). Accommodation to vocal pitch in children with autism. Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS), 3917-3921.
KIM, M., W. S. HORTON & BRADLOW, A. R. (2011). "Phonetic convergence in spontaneous conversations as afunction of interlocutor language distance." *Laboratory Phonology, 2,* 125-156.
KISSINE, M., GEELHAND, P. (2019). Acoustic evidence for increased articulatory stability in the speech of adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders, 49*, 2572-2580.
KISSINE, M., GEELHAND, P., PHILIPPART DE FOY, M., HARMEGNIES, B. & DELIENS, G. (2021), Phonetic Inflexibility in Autistic Adults. *Autism Research, 14*, 1186-1196.
KITAMURA, C., THANAVISHUTH, C., BURNHAM, D., & LUKSANEEYANAWIN, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior & Development, 24*(4), 372-392. https://doi.org/10.1016/S0163-6383(02)00086-3
LELONG, A. (2012). Phonetic convergence in interaction. Université de Grenoble, Retrieved from : https://tel.archives-ouvertes.fr/tel-00822871
LORD, C., RUTTER, M., & LE COUTEUR, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659-685.
MÉNARD, L., PERRIER, P. & AUBIN, J. (2016). Compensation for a lip tube perturbation in 4 year olds: Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *The Journal of the Acoustical Society of America, 139*(5), 2514-2531.
NAMY, L. L., NYGAARD, L. C. & SAUERTEIG, D. (2002). "Gender differences in vocal accommodation: The role of perception." *Journal of Language and Social Psychology, 21*, 422-432.
NIELSEN, K. (2008). The specificity of allophonic variability and its implications for accounts of speech perception. Doctoral Dissertation, University of California, Los Angeles.
NIELSEN, K. (2011). Specificity and abstractness of VOT imitation. Journal of Phonetics, 39, 132-142.
NIELSEN, K. (2011). Phonetic imitation by school-age children. Poster presented at the 162nd Meeting of the Acoustical Society of America. San Diego, CA.
PEPPÉ, S., MCCANN, J., GIBBON, F., O'HARE, A., & RUTHERFORD, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research, 50*(4), 1015-1028.

PARDO, J. S. (2006). "On phonetic convergence during conversational interaction." *Journal of the Acoustical Association of America, 119*(4), 2382–2393.
PARDO, J. S., CAJORI JAY, I., & KRAUSS, R. M. (2010). Conversational role influences speech imitation. Attention, *Perception, & Psychophysics, 72*(8), 2254-2264
PARDO, J. S., GIBBONS, R., SUPPES, A., AND KRAUSS, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40,* 190-197. doi: 10.1016/j.wocn.2011.10.001
PIAZZA, G., KALASHNIKOVA, M., & MARTIN, C. D. (2023). Phonetic accommodation in non-native directed speech supports L2 word learning and pronunciation. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-48648-7
PLOOG, B. O. (2010). Stimulus Overselectivity Four Decades Later: A Review of the Literature and Its Implications for Current Research in Autism Spectrum Disorder. *Autism Dev. Disord., 40*(11). 1332-1349.
PLOOG, B. O., BANERJEE, S., BROOKS, P. J. (2009). Attention to prosody (intonation) and content in children with autism and in typical children using spoken sentences in a computer game. *Autism Spectr. Disord., 3*(3), 743-758,
RINEHART N. J., BRADSHAW J. L., MOSS S. A., BRERETON A. V., TONGE B. J. (2001). A deficit in shifting attention present in high-functioning autism but not Asperger's disorder. *Autism 2001, 5*, 67-80.
SANCHEZ, K., MILLER, R. M., & ROSENBLUM, L. D. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research, 53*(2), 262-272.
SHAH, A., & FRITH, U. (1983). An islet of ability in autistic children: a research note. *Journal of Child Psychology and Psychiatry, 24*(4). 613-620.
YU, A. C., ABREGO-COLLIER, C., & SONDEREGGER, M. (2013). Phonetic imitation from an individual-difference perspective: subjective attitude, personality and "autistic" traits. *PloS one, 8*(9). https://doi.org/10.1371/journal.pone.0074746

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Caractérisation acoustique des réalisations approximantes du /v/ intervocalique en français spontané

Suyuan Dong[1] Nicolas Audibert[1]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle),
4 rue des Irlandais, 75005 Paris, France
suyuan.dong@sorbonne-nouvelle.fr, nicolas.audibert@sorbonne-nouvelle.fr

RESUME

Les fricatives /v/ ont tendance à se réaliser comme une variante affaiblie en français spontané. Nous nous sommes appuyés sur les données du corpus NCCFr et avons évalué 5504 occurrences de /v/ intervocalique produites par 10 hommes et 10 femmes, à partir de l'inspection des spectrogrammes. Un ensemble de mesures acoustiques dynamiques ont été relevées sur chaque exemplaire, et comparées entre les deux types de réalisation (fricatives voisées et approximantes) par des modèles GAM pour prendre en compte leur évolution temporelle. Les résultats montrent une prépondérance des réalisations approximantes, également observables en positions accentuées, et indiquent que ces deux types de réalisation divergent tant au niveau temporel que spectral, particulièrement dans les hautes fréquences. La manifestation de cet affaiblissement varie également en fonction des locuteurs. Ces observations permettent de suggérer que le /v/ intervocalique serait sujet à un processus d'affaiblissement en cours, au-delà d'une simple réduction segmentale.

ABSTRACT

Acoustic characterization of approximant realizations of intervocalic /v/ in spontaneous French

The fricative /v/ tends to manifest as a weakened variant in spontaneous French. This study relies on data from the NCCFr corpus and evaluated 5504 occurrences of intervocalic /v/ produced by 10 males and 10 females, through the inspection of spectrograms. A set of dynamic acoustic measurements was taken on each exemplar and compared between the two types of realization (voiced fricative and approximant) using GAM models to account for their temporal evolution. The results show a predominance of approximant realizations, also observable in accentuated position, and indicate that these two types of realization diverge in both temporal and spectral dimension, particularly in the high frequencies. The manifestation of this weakening also varies according to speaker differences. These observations suggest that the intervocalic /v/ might be subject to an ongoing weakening process, beyond a simple segmental reduction.

MOTS-CLÉS : fricative ; approximante ; labiodentale ; français spontané ; grand corpus ; acoustique

KEYWORDS : fricative; approximant; labiodental; spontaneous French; large corpora; acoustic

Actes des 35èmes Journées d'Études sur la Parole, pages 13–22.

1 Introduction

La variabilité phonétique est omniprésente dans la langue orale. La lénition désigne le processus phonologique par lequel un segment devient moins similaire à sa réalisation initiale (Trask, 2006, p.190). Sous une perspective synchronique, la lénition se manifeste par des alternances sonores au sein d'une langue dans une période donnée, transformant un son en un allophone « plus faible » (*ibid.*, p.201; Kirchner, 2004, p.313). Diachroniquement, elle peut entraîner une convergence progressive entre deux phonèmes, jusqu'à une perte de contraste phonologique (Trask, 2000, p.216). Selon Trask (*ibid.*, p.191) et Crystal (2008, p.274), la lénition peut engendrer des changements sonores tels que Occlusive > Fricative > Approximante > Zéro. La lénition est aussi décrite comme un phénomène positionnel, souvent observé en position faible, telles qu'en intervocalique ou coda (De Carvalho *et al.*, 2008, pp.131-172 ; Jatteau *et al.*, 2019 ; Lancien *et al.*, 2023).

Bien que les descriptions acoustiques pour les fricatives labiodentales fassent l'objet de nombreuses études, les recherches se concentrant sur leurs réalisations affaiblies sont moindres. Cependant, notre analyse préliminaire du corpus NCCFr (*The Nijmegen Corpus of Casual French*, Torreira *et al.*, 2010) révèle que ce phénomène est couramment observé dans le français spontané. Par conséquent, cette étude cherche à examiner spécifiquement les /v/ intervocaliques, et à caractériser acoustiquement leurs réalisations affaiblies en tant qu'approximantes, afin de déterminer si ce phénomène résulte d'une simple réduction segmentale en parole spontanée ou s'il pourrait s'agir d'un potentiel phénomène de lénition en cours en français spontané parmi la jeune génération.

Du point de vue articulatoire, les approximantes se distinguent des fricatives et des voyelles par le degré de fermeture entre des articulateurs (Hewlett & Beck, 2006, pp.37-39), ainsi que par le niveau d'implication du bruit de friction (Crystal, 2008, p.32 ; Trask, 2006, p.30). Les fricatives sont définies comme des consonnes produites par un rétrécissement étroit et incomplet entre deux articulateurs, générant un bruit de friction audible (e.g., Kent & Read, 2001, p.38; Vaissière, 2020, p.70). À l'inverse, les approximantes sont produites lorsque le conduit vocal est nettement rétréci, sans pour autant créer une constriction suffisamment étroite pour générer un flux d'air turbulent (Ladefoged & Johnson, 2010, p.15; Kent & Read, 2001, p.177). Cette constriction modérée les distingue des voyelles, résultant en une stabilité moindre et une énergie acoustique plus faible. Sur le plan aérodynamique, les fricatives voisées combinent deux sources d'énergie : le voisement et le bruit de friction. Cette dualité génère un conflit aérodynamique : la génération du bruit turbulent nécessite une vélocité d'air élevée, tandis que les vibrations des plis vocaux tendent à ralentir le passage du flux d'air. En conséquence, les fricatives sonores ont tendance à perdre leur bruit de friction et ainsi se réaliser comme des approximantes (Johnson, 2012, p.156).

À partir de ces caractéristiques articulatoires et aérodynamiques, les fricatives sont généralement décrites selon quatre attributs acoustiques : les propriétés spectrales et l'amplitude du bruit de friction, la durée, ainsi que les propriétés spectrales des transitions formantiques liées aux voyelles adjacentes (Reetz & Jongman, 2009, pp.227-230). Diverses méthodes ont été proposées pour mesurer et différencier les fricatives, incluant, sans s'y limiter, le pic de fréquence, les moments spectraux, la transition formantique, l'amplitude intégrale et dynamique, ainsi que la durée (e.g., Jongman *et al.*, 2000; Maniwa *et al.*, 2009; Al-Tamimi & Khattab, 2015). Récemment, la quantification de l'énergie dans les hautes fréquences (>7kHz) pour décrire les fricatives a suscité un regain d'attention (Shadle *et al.*, 2023a; Kharlamov *et al.*, 2023).

Les descriptions acoustiques des approximantes sont pour leur part moins nombreuses. Malgré leurs similitudes avec les voyelles, elles présentent une structure formantique moins intense et moins

stable, accompagnée d'une durée plus courte que celles des voyelles (Reetz & Jongman, 2009, p.225). En comparaison avec les fricatives voisées, compte tenu de l'absence de turbulence, les approximantes devraient présenter une structure formantique plus marquée et stable, un signal sonore moins bruité, une amplitude supérieure mais avec moins d'énergie dans les hautes fréquences, ainsi qu'un voisement plus constant que les fricatives voisées. De plus, si ces réalisations en tant qu'approximantes correspondent simplement à une réduction due à l'hypoarticulation et à la rapidité de la parole spontanée, elles devraient présenter une durée réduite et être rares en position accentuée (Adda-Decker & Snoeren, 2011 ; Ernestus & Warner, 2011 ; Lindblom, 1990).

2 Méthodologie

2.1 Corpus

Le corpus NCCFr (Torreira *et al*., 2010) a été élaboré pour fournir des enregistrements de parole informelle et spontanée en français, adaptés aux investigations scientifiques. Cette base de données rassemble plus de 36 heures d'enregistrements, produits par 46 jeunes locuteurs natifs du français (22 femmes et 24 hommes, âgés principalement de 18 à 27 ans). Les participants, socio-géographiquement homogènes, ont été enregistrés en binômes d'amis pour capturer leurs conversations spontanées. Les enregistrements ont été recueillis avec un enregistreur stéréo à semi-conducteur Edirol R-09, équipé de microphones unidirectionnels Samson QV et d'un préamplificateur, avec une fréquence d'échantillonnage de 48kHz. Les données collectées ont ensuite été transcrites et annotées manuellement en suivant les directives développées par LIMSI, puis segmentées par alignement automatique forcé répété dans le même système. Pour cette étude, nous avons sélectionné un sous-ensemble de 20 locuteurs avec un total de 5504 occurrences du phonème /v/ intervocalique. Ces réalisations se répartissent de manière relativement équilibrée au sein des locuteurs, comprenant 2970 /v/ produits par 10 femmes et 2534 par 10 hommes.

2.2 Annotation des données

Nous avons commencé par la classification des différentes réalisations du /v/ intervocalique à l'aide de scripts Praat. Ces scripts permettent d'extraire automatiquement des informations du son cible, telles que la durée, les voyelles adjacentes et les informations lexicales. Les différents types de réalisation du /v/ ont été manuellement catégorisés à partir de l'observation du spectrogramme dans les bandes de fréquence de 0-5kHz (structure formantique) et 0-15kHz (bruit de friction), résultant en 9 catégories : fricative voisée, approximante, élision, dévoisement, voix craquée, voix soufflée, superposition, signal faible et autres. La catégorie « autres » regroupe les sons non-exploitables tels que les rires, les erreurs de transcription, les cas d'élision d'une des voyelles adjacentes, ou encore la saturation du signal due à la voix criée. Cette catégorisation s'est basée sur des indices acoustiques visuellement et auditivement identifiés, principalement la structure formantique, la transition formantique, le bruit de friction, la barre de voisement, la périodicité et la réduction d'amplitude. Ces indices ont été notés durant la catégorisation. Des critères complémentaires ont aussi été pris en compte, incluant les voyelles bruitées ou dévoisées et différentes catégories de positions prosodiques fortes : la « position accentuée » correspondant à la proéminence à la dernière syllabe des groupes rythmiques (hésitations exclues), l'« insistance » où le locuteur accentue un mot pour un objectif pragmatique (focus), et la « syllabe accentuée » qui englobe les deux catégories précédentes, mais

incluant cette fois-ci les cas d'hésitation. Lors de cette étape, nous avons affiné manuellement la position des frontières des sons cibles. Toutes les catégorisations et segmentations ont été revérifiées.

2.3 Analyses acoustiques

Étant donné que l'objectif principal de cette étude est d'identifier le potentiel contraste acoustique entre les réalisations modales et affaiblies du /v/ intervocalique et de les caractériser, nous avons sélectionné 30 mesures proposées dans la littérature, et les avons extraites à l'aide d'un autre script Praat. Ces mesures incluent la durée du /v/, les moments spectraux, l'intensité prise au milieu des /v/ pour illustrer la forme spectrale générale, l'amplitude dynamique dans différentes gammes de fréquences, le HNR, les formants F1 à F4 convertis en Bark, ainsi que les mesures spécifiques aux non-sibilantes dans les hautes fréquences (<7kHz) proposées par Shadle *et al*. (2023a). Il est à préciser que ces dernières suivent le paramétrage proposé par les auteurs, mais sont implémentées sur un spectre DFT au lieu de la méthode *multitaper*, ce qui selon Kharlamov *et al*. (2023) n'impacte pas la classification des fricatives dans l'analyse de la parole continue.

Pour rendre compte de la dynamique de ces mesures, nous avons retenu onze points de mesure équidistants répartis sur la durée totale des réalisations de /v/, complétés par cinq points de mesures sur la seconde moitié de la voyelle précédente (V1), et cinq points sur la première moitié de la voyelle suivante (V2). Les paramètres ont été adaptés pour tenir compte des différences entre hommes et femmes pour les mesures formantiques (méthode de Burg avec une fréquence maximale respectivement de 5kHz et 5.5kHz pour la détection de 5 formants). Les mesures des moments spectraux s'appuient sur les paramètres utilisés par Al-Tamimi & Khattab (2015). La forme du spectre de 0 à 14kHz avec une résolution de 20Hz a également été extraite au milieu du /v/.

2.4 Analyses statistiques

Nous avons d'abord comparé la distribution des différentes réalisations du /v/ et des indices acoustiques relevés. Par la suite, nous avons concentré notre analyse sur les réalisations en tant que fricatives voisées ou approximantes. L'effet du type de réalisation et du sexe des locuteurs sur la durée (transformée en log) a été évalué à l'aide d'un modèle linéaire mixte (fonction lmer du package R lme4 (Bates *et al.*, 2015) avec le locuteur comme ordonnée à l'origine (*intercept*) aléatoire.

La trajectoire des mesures acoustiques sur 21 points entre le milieu de la voyelle précédente V1 et le milieu de la suivante V2 a été modélisée par des modèles additifs généralisés (GAM), séparément pour hommes et femmes, incluant comme prédicteurs le type de réalisation, la durée du /v/, ainsi que celles des V1 et V2. Les différences entre locuteurs ont été considérées comme lissage aléatoire. Suite à une première analyse qui ne révèle pas d'interaction significative entre le contexte vocalique et le type de réalisation, le contexte a été intégré comme facteur aléatoire dans ces modèles. La forme spectrale au milieu du /v/ a également été modélisée par GAM séparément pour hommes et femmes avec les mêmes prédicteurs et facteurs aléatoires, cette fois avec une modélisation en fonction de la fréquence et non du temps.

3 Résultats

3.1 Distribution des types de réalisation : effet du locuteur et du sexe

Parmi les 5504 occurrences du /v/ intervocalique examinées, les réalisations comme fricatives voisées (39.9%, n=2197) et comme approximantes (28.8%, n=1583) sont les plus fréquentes. La Figure 1 présente la distribution de chaque type de réalisation en fonction des locuteurs (10F+10H). En outre, 6.6% (n=365) des /v/ correspondent à des superpositions de parole, en raison de la nature interactive des dialogues. Quant aux réalisations en fonction du sexe, les résultats suggèrent que les locuteurs masculins ont produit plus d'approximantes (33.4%, n=846) que les locutrices (24.8%, n=737). Les élisions présentées ici correspondent aux cas où aucune trace visuelle ni audible n'est présente pour indiquer la réalisation du /v/, en complément des cas déjà pris en compte dans l'annotation initiale du corpus.

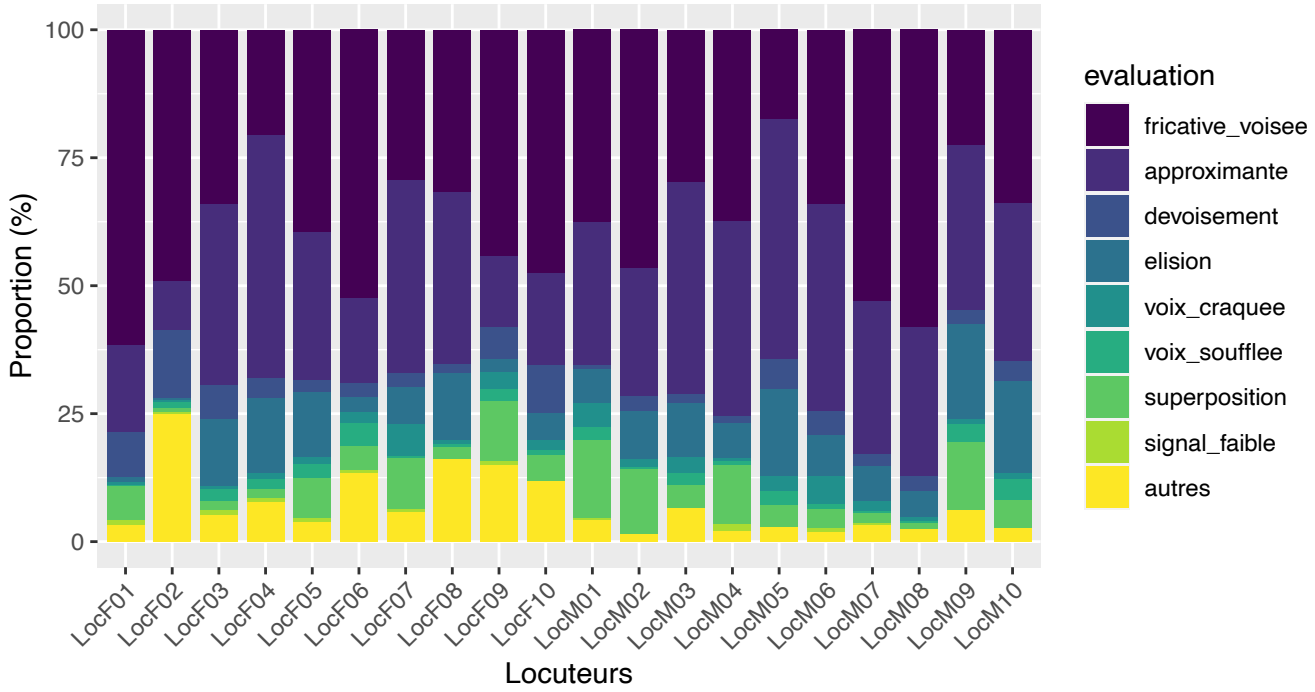


FIGURE 1 : Répartition générale des différents types de réalisations par locuteur/trice. Le sexe des participants est indiqué par « LocF » pour les femmes, et « LocM » pour les hommes.

Les indices acoustiques directement observables dans les spectrogrammes et le signal sonore sont globalement conformes aux descriptions de la littérature. Le bruit de friction et la structure formantique constituent les principaux indicateurs pour identifier visuellement les deux types de réalisation (fricatives voisées et approximantes). Une réduction d'amplitude relative aux voyelles adjacentes est plus fréquemment observée chez les fricatives voisées. Quant aux positions prosodiques, 21.8% (1200/5504) des /v/ ont été réalisés en position forte. Parmi ces réalisations, nous avons observé une présence relativement importante d'approximantes, avec 37% (311/839) en position accentuée et 11% (20/186) en focus, contre respectivement 57% (482/839) et 81% (151/186) de fricatives voisées. Une variabilité individuelle a également été constatée dans la production des approximantes en position accentuée, avec par exemple LocM09 à 71% contre seulement 2% pour LocF02. En moyenne sur l'ensemble des 20 locuteurs, 38% (σ=18%) des /v/ en position accentuée sont réalisés comme approximantes.

3.2 Résultats d'analyses acoustiques

La comparaison des durées (Figure 2, gauche) suggère que les approximantes tendent à être plus courtes que les fricatives voisées pour les deux sexes, mais un recouvrement important est présent. Afin d'éviter un éventuel impact des différences individuelles de débit de parole, nous avons normalisé la durée par locuteur (Figure 2, droite). Une différence significative de durée est bien observée en fonction du type de réalisations ($p<2.2e-16$), mais sans effet significatif du sexe des locuteurs ($p=0.52$), ni d'interaction entre le sexe et le type de réalisation ($p=0.71$).

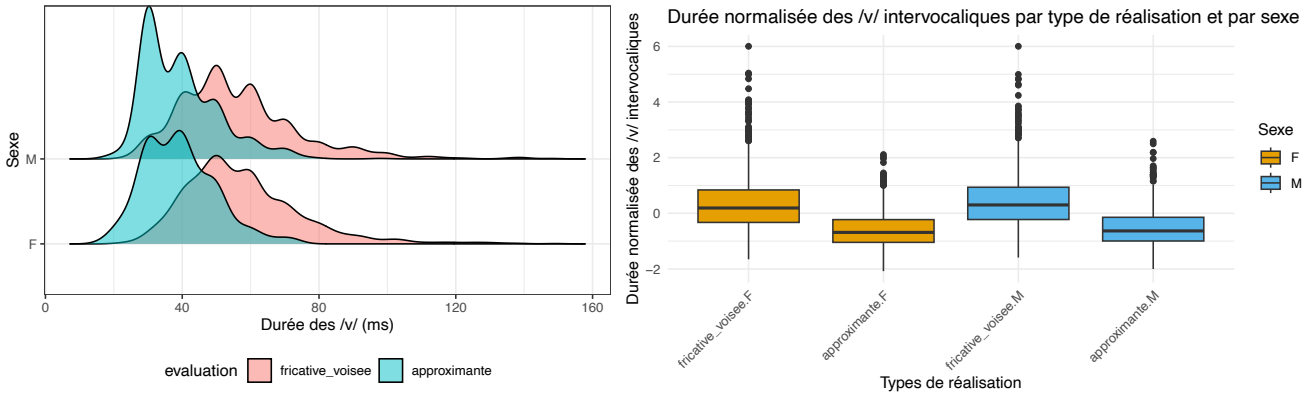


FIGURE 2 : Distribution des durées des sons cibles en fonction du sexe et du type de réalisations, avec l'échelle d'origine (gauche) et après normalisation par locuteur (droite).

Pour chacun des 60 modèles GAM correspondant aux mesures acoustiques (30 mesures * 2 sexes), les valeurs prédites au cours du temps ont été extraites pour les deux types de réalisation avec les intervalles crédibles associés. La Figure 3 illustre l'écart maximum entre fricatives voisées et approximantes prédit par les modèles pour chacune de ces mesures, après normalisation en z-scores. Parmi elles, nous pouvons distinguer les huit mesures sur la gauche de la figure qui discriminent le mieux les deux types de réalisation, avec une divergence maximum supérieure à un écart-type pour les deux sexes. Ces mesures incluent l'intensité relative (IntRel) dans les moyennes et hautes fréquences (11-13kHz > 5-8kHz > 4-5.6kHz), conformément aux observations sur la forme générale des spectres au milieu des /v/ non détaillées ici. Cette divergence entre fricatives et approximantes est également capturée par le centre de gravité spectral (CoG), la dispersion spectrale (SD), ainsi que les trois mesures (HiLevD, AmpRange, LevelH) proposées par Shadle *et al.* (2023a). Parmi ces trois dernières mesures, HiLevD et AmpRange sont suggérées par les auteurs comme les meilleurs indicateurs d'augmentation du bruit de turbulence pour les non-sibilants.

Le CoG et la SD font partie des mesures les plus couramment utilisées pour caractériser les fricatives. Un CoG élevé désigne une articulation plus antérieure, tandis qu'un SD élevé indique une dispersion spectrale plus large (e.g., Forrest *et al.*, 1988; Al-Tamimi & Khattab, 2015). Les résultats semblent correspondre à ce qui est attendu : les fricatives voisées présentent un spectre plus diffus et probablement un lieu d'articulation plus avancé que les approximantes ; ces dernières présentent une distribution spectrale plus similaires aux voyelles environnantes. Cependant, le CoG et SD peuvent être impactés par le degré de voisement, ainsi que par l'énergie dans les fréquences 11-13kHz et 5-8kHz comme le suggèrent les corrélations avec l'énergie relative dans ces bandes de fréquence (H : ρ(CoG, 11-13k)=0.67, ρ(SD, 11-13k)=0.76, ρ(CoG, 5-8k)=0.65, ρ(SD, 5-8k)=0.66 ; F : ρ(CoG, 11-13k)=0.74, ρ(SD, 11-13k)=0.77, ρ(CoG, 5-8k)=0.69, ρ(SD, 5-8k)=0.62). Parmi ces huit mesures, les corrélations en valeur absolue sont comprises entre 0.37 (ρ(CoG, AmpRange)) et 0.8 (ρ(CoG, SD)).

Le changement spectral dans les hautes fréquences est estimé par HiLevD qui mesure la différence d'intensité entre trois bandes de fréquence du spectre lissé : LevM (3-7kHz), LevH(7-11k) et LevHH(11-15k) (Shadle *et al.*, 2023a). Cette mesure de pente spectrale en hautes fréquences est suggérée comme l'indicateur le plus sensible de l'augmentation des énergies dans les hautes fréquences au milieu des fricatives. Nos résultats illustrent son efficacité pour distinguer les réalisations de /v/, marquées par une diminution notable pour les fricatives voisées et une augmentation pour les approximantes. Conformément aux observations initiales des auteurs sur les fricatives sourdes, la valeur minimale de cette mesure est très proche du milieu des fricatives /v/ (non illustré ici).

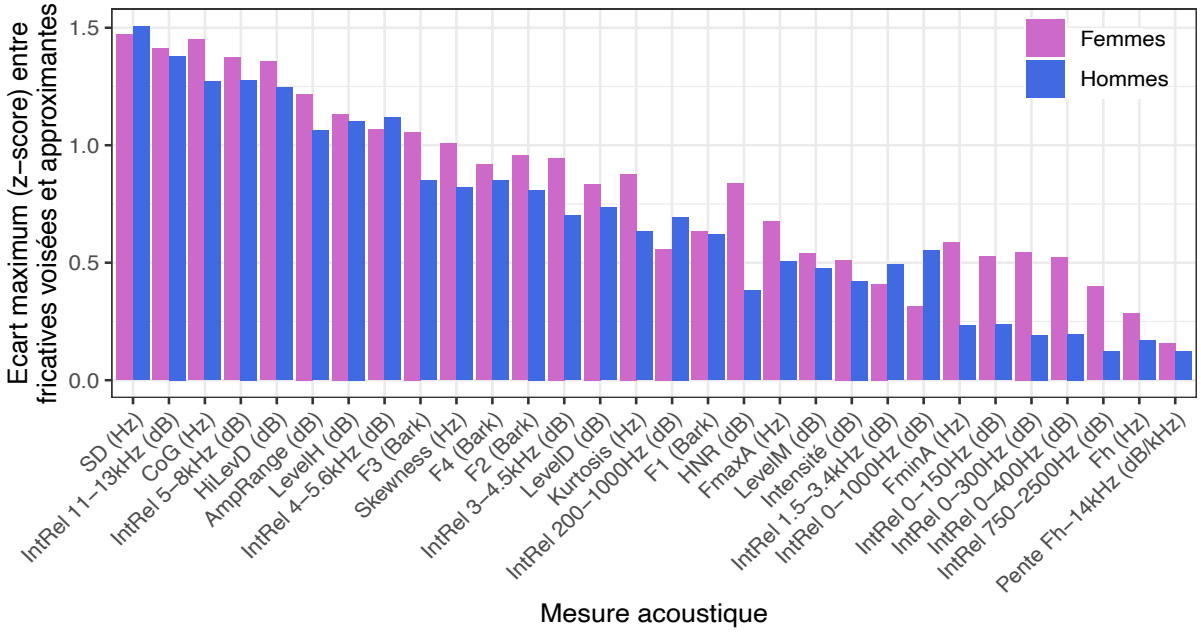


FIGURE 3 : Distance maximum modélisée par les GAM entre fricatives voisées et approximantes pour les 30 mesures acoustiques (normalisées en z-scores) en fonction du sexe des locuteurs.

La mesure AmpRange, quant à elle, est définie comme la différence d'amplitude (dB) entre le maximum en hautes fréquences (FmaxA : 2-13kHz) et le minimum en basses fréquences (FminA : 1-7kHz). Selon les auteurs, cette mesure correspond en partie à la réduction de la taille de constriction au fil du temps, dans le sens où plus la constriction est importante, plus la valeur de cette mesure est importante. Nos résultats indiquent bien une réduction de FminA et une augmentation de FmaxA qui correspondraient à une constriction plus importante pour les fricatives voisées, mais l'amplitude relative des variations sur ces mesures a pour conséquence des valeurs non conformes aux prédictions pour la mesure AmpRange.

4 Discussion

Cette étude a impliqué l'analyse acoustique de 30 paramètres, étendus jusqu'à 14kHz, pour étudier les réalisations des /v/ intervocaliques. Les résultats pour la plupart de ces mesures suggèrent que les /v/ intervocaliques ont tendance à se réaliser comme une variante affaiblie distincte des fricatives voisées. Malgré un recouvrement partiel entre gammes de fréquences capturée par les huit mesures les plus discriminantes, elles ne sont pas directement corrélées, à l'exception de relations modérées entre CoG et SD, aussi entre ces mesures et l'intensité relative dans les bandes 5-8kHz et 11-13kHz.

Cette distinction entre fricatives voisées et approximantes se manifeste d'abord au niveau spectral. Les principaux résultats acoustiques sont consistants avec les observations issues des études précédentes, indiquant un bruit de friction diffus notamment dans les hautes fréquences pour les fricatives voisées par rapport aux approximantes (Shadle *et al*., 2023a; Kharlamov *et al*., 2023). Pour la mesure AmpRange, en revanche, l'inconsistance entre nos résultats et ceux de l'étude de Shadle *et al*. (2023a) pourrait être due au voisement présent dans nos données, aux différences de contexte (exclusivement intervocalique dans notre cas), voire au style de parole (spontanée dans nos données, parole lue et mots isolés dans celles de Shadle *et al*.).

En outre, nos modèles GAM (non illustrés ici) indiquent que la divergence est maximale en un point temporel très proche du milieu du /v/ pour toutes les mesures à l'exception de l'énergie relative sur les bandes de fréquence 5-8kHz et 11-13kHz (divergence maximale légèrement plus tardive, entre 58,5% et 63% de la durée totale de la consonne), et de la mesure Fh qui est toutefois peu discriminante entre types de réalisations. Pour la majorité des mesures acoustiques prises en compte, nous pourrons donc nous appuyer sur des mesures prises au milieu de la consonne afin de caractériser les réalisations fricatives ou approximantes dans nos travaux ultérieurs. Ce constat ouvre la voie à une extension à plus grande échelle de ces analyses, avec une application envisageable à des données pour lesquelles seul un alignement forcé automatique est disponible.

En outre, les données analysées proviennent de la parole spontanée caractérisée par des conversations informelles et relâchées dans un environnement calme. Ce style de parole tend à favoriser une production réduite, typiquement marquée par des sons plus courts et moins articulés (Adda-Decker & Snoeren, 2011 ; Ernestus & Warner, 2011). Bien que nos résultats indiquent des durées significativement réduites pour les réalisations approximantes, le chevauchement important des distributions des durées suggère que les approximantes ne sont pas systématiquement plus courtes que les fricatives voisées. Tandis que les positions prosodiques accentuées impliquent souvent un effort articulatoire plus élevé (Cho, 2016), elles donnent lieu dans les données analysées à des réalisations approximantes relativement fréquentes et même majoritaires chez certains locuteurs. Ces observations suggèrent que les réalisations approximantes pourraient ne pas constituer de simples réductions segmentales, mais être plutôt une réalisation allophonique liée à la lénition. Dans cette perspective et compte tenu du fait que les données du corpus NCCFr ont été enregistrées en 2007, il serait intéressant de les comparer à des productions spontanées de locuteurs d'Ile de France de la même tranche d'âge en 2024 afin d'évaluer l'évolution de ce phénomène.

Nos résultats suggèrent également un effet du sexe sur la distinction acoustique entre ces deux types de réalisation. Par exemple, la distinction entre fricatives et approximantes est également identifiable dans les basses fréquences (0-400Hz) chez les femmes, sans pour autant être le cas chez les hommes, alors que la fréquence fondamentale est supposée incluse dans cette plage de fréquence dans les deux cas. L'observation qualitative des données suggère aussi une variabilité individuelle au-delà des importantes différences de distribution des types de réalisation, ainsi qu'une convergence au cours du temps au sein des binômes qui sera approfondie dans des recherches ultérieures.

Remerciements

Ce travail a été soutenu par le Partenariat Hubert Curien (PHC) Van Gogh n°49298RM et par le Laboratoire d'Excellence (LabEx) Empirical Foundations of Linguistics (EFL) n°ANR-10-LABX-0083. Il contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

ADDA-DECKER, M. & SNOEREN, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39(3), 261-270.

BATES D, MÄCHLER M, BOLKER B & WAKLER S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. DOI : 10.18637/jss.v067.i01.

AL-TAMIMI J. & KHATTAB G. (2015). Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *The Journal of the Acoustical Society of America*, 138(1), 344-360. DOI : 10.1121/1.4922514.

CHO, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. Language and Linguistics Compass, 10(3), 120-141.

CRYSTAL D. (2008). *A dictionary of linguistics and phonetics* (6e éd). Blackwell Pub.

DE CARVALHO J. B., SEGERAL P., SCHEER T., Éds. (2008). *Lenition & fortition*. Mouton de Gruyter.

ERNESTUS, M. & WARNER, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(SI), 253-260.

HEWLETT N. & BECK J. M. (2006). *An Introduction to the Science of Phonetics.* Routledge.

FORREST K., WEISMER G., MILENKOVIC P. & DOUGALL R. N. (1988). Statistical Analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84, 115-123. DOI : 10.1121/1.396977.

JATTEAU A., VASILESCU I., LAMEL L., ADDA-DECKER M. & AUDIBERT N. (2019). "Gra[f]e!" Word-Final Devoicing of Obstruents in Standard French: An Acoustic Study Based on Large Corpora. *Interspeech* 2019, 1726-1730. DOI : 10.21437/Interspeech.2019-2329.

JOHNSON K. (2012). *Acoustic and Auditory Phonetics*. Wiley-Blackwell.

JONGMAN A., WAYLAND R. & WONG S. (2000). Acoustic characteristics of English fricatives. The *Journal of the Acoustical Society of America*, 108(3), 1252-1263. DOI :10.1121/1.1288413.

KENT R. D. & READ C. (2001). *Acoustic Analysis of Speech (2nd edition)*. Cengage Learning.

KHARLAMOV V., BRENNER D. & TUCKER B. V. (2023). Examining the effect of high-frequency information on the classification of conversationally produced English fricatives). *The Journal of the Acoustical Society of America*, 154(3), 1896-1902. DOI : 10.1121/10.0021067.

KIRCHNER R. (2004). Consonant lenition. In HAYES B., KIRCHNER R. & STERIADE D., Éds., *Phonetically Based Phonology* (1re éd.), p. 313-345. Cambridge University Press. DOI : 10.1017/CBO9780511486401.010.

LADEFOGED P. & JOHNSON K. (2010). *A Course in Phonetics* (6e éd). Wadsworth.

LANCIEN M., HUTIN M., STUART-SMITH J., ADDA-DECKER M. & VASILESCU I. (2023). /R/ Lenition in Quebec French: Evidence from the Distribution of 9 Allophones in Large Corpora. *20th International Congress of Phonetic Sciences* (ICPhS). Prague, Czech Republic. ISBN : 9788090811423.

LINDBLOM, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403-439). Dordrecht: Springer Netherlands.

MANIWA K., JONGMAN A. & WADE T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America,* 125(6), 3962-3973. DOI : 10.1121/1.2990715.

REETZ H. & JONGMAN A. (2009). *Phonetics Transcription, Production, Acoustics, and Perception.* Wiley-Blackwell.

SHADLE C. H., CHEN W. R., KOENIG L. L. & PRESTON J. L. (2023a). Refining and extending measures for fricative spectra, with special attention to the high-frequency range. *The Journal of the Acoustical Society of America*, 154(3), 1932-1944. DOI : 10.1121/10.0021075.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3), 201-212. DOI : 10.1016/j.specom.2009.10.004.

TRASK R. L. (2000). *The Dictionary of Historical and Comparative Linguistics*. Edinburgh University Press.

TRASK R. L. (2006). *A dictionary of phonetics and phonology* (Reprinted). Routledge.

VAISSIÈRE J. (2020). *La Phonétique* (4e éd). Que sais-je ? Presses Universitaires de France.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Comment l'oreille humaine perçoit-elle la somnolence dans la parole ? Une analyse rétrospective d'études perceptuelles.

Vincent P. Martin[1] Colleen Beaumard[2,3] Jean-Luc Rouas[2]
(1) Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, L-1445, Luxembourg
(2) Univ. Bordeaux, LaBRI, CNRS UMR 5800, Bordeaux INP, Talence, F-33405, France
(3)Univ. Bordeaux, CNRS, SANPSY, UMR 6033, Bordeaux, F-33000, France
vincentp.martin@lih.lu, {colleen.beaumard, jean-luc.rouas}@labri.fr

RÉSUMÉ

La somnolence bénéficierait d'être mesurée dans des configurations écologiques, par exemple grâce à des enregistrements de parole. Pour évaluer la faisabilité de sa détection à partir de la parole par l'audition humaine, deux études perceptuelles précédentes ont produit des résultats contradictoires. Une façon de comprendre ce désaccord aurait pu être d'étudier sur quelles caractéristiques de la parole les annotateurs ont basé leur estimation, mais aucune étude n'a collecté cette information. Nous avons donc choisi d'extraire des descripteurs acoustiques des enregistrements annotés, et d'entraîner des modèles d'apprentissage automatique simples et explicables à reproduire l'annotation de chaque annotateur. Ensuite, nous mesurons la contribution de chaque caractéristique à la décision de chaque modèle, et identifions les plus importantes. Nous effectuons ensuite un regroupement hiérarchique pour dessiner les profils des annotateurs, en fonction des caractéristiques sur lesquelles ils s'appuient pour identifier la somnolence.

ABSTRACT

How does human hearing perceive sleepiness from speech ? A retrospective analysis of perceptual experiments

Excessive sleepiness would benefit from being measured in ecological settings, for example through speech recordings. To assess the feasibility of detecting sleepiness from speech by human hearing, two previous perceptual studies have yielded contradictory results. One way to investigate this disagreement would have been to look into which speech characteristics listeners based their estimation, but no study has collected this information. In this study, we extract acoustic descriptors from annotated recordings, and train simple and explainable machine learning models to reproduce the annotation of each annotator. Then, we measure the contribution of each feature to each model's decision, and identify the most important ones. We then perform hierarchical clustering to draw profiles of listeners, based on the features they rely on to identify sleepiness.

MOTS-CLÉS : Études perceptuelles, Somnolence, Modèle interprétable.

KEYWORDS: Perceptual studies, Sleepiness, Interpretable model.

Cet article est une traduction en français d'un article en cours d'évaluation par les pairs pour la conférence internationale *Speech Prosody 2024*. Il traite des éléments de la parole sur laquelle se sont basés les annotateurs d'une étude perceptuelle, que nous avions recrutés en 2021. Tous étaient francophones. Il nous semble important de promouvoir ce travail en français pour permettre un retour vers les personnes qui ont participé à l'étude (qui ne lisent en général pas l'anglais) et avoir des retours

scientifiques de la communauté de recherche en parole sur la perception de la parole en français.

1 Introduction

Contexte. L'hypersomnie est un fardeau majeur à la fois pour la santé publique (Léger *et al.*, 2012; Barnes & Watson, 2019) et la santé personnelle, en lien avec des troubles métaboliques, cardiovasculaires, neurologiques et psychiatriques, augmentant le risque d'invalidité et de mortalité (Jike *et al.*, 2018; Scott *et al.*, 2021). En raison de sa forte prévalence dans la population générale (jusqu'à une personne sur trois (Kolla *et al.*, 2020)), les cliniciens ont besoin d'outils pour mesurer le niveau de somnolence de leurs patients aussi régulièrement que possible, dans des conditions écologiques (par exemple, à domicile), de manière passive (c'est-à-dire sans tâche dédiée). À cet égard, les enregistrements vocaux et de parole sont un candidat de choix : leur collecte est implémentée dans tous les smartphones, ils peuvent être enregistrés dans des configurations passives, et ils ont déjà été liés à de multiples troubles (Fagherazzi *et al.*, 2021), y compris la somnolence.

Précédents travaux. En effet, la détection de la somnolence à l'aide d'enregistrements vocaux a déjà été au centre de deux challenges Interspeech en 2011 et 2019, reposant respectivement sur le *Sleep Language Corpus* (SLC) (Schuller *et al.*, 2011) et le corpus SLEEP (Schuller *et al.*, 2019). Les deux corpus sont étiquetés avec une mesure subjective (questionnaire d'auto-évaluation) (Martin *et al.*, 2021), l'échelle de somnolence de Karolinska (KSS) (Åkerstedt & Gillberg, 1990).

Le meilleur système du challenge Interspeech 2011 a atteint un score de Rappel Moyen Non Pondéré (*Unweighed Average Recall*, UAR) de 71.7% (Huang *et al.*, 2011) sur la classification binaire de la somnolence. Sur le corpus SLEEP, la tâche du défi Interspeech 2019 était d'estimer le degré de somnolence. Les gagnants du challenge ont atteint une corrélation de Spearman $\rho = 0.387$ entre l'estimation produite par leur système et la vérité terrain (Gosztolya, 2019). Cette approche simple n'a jamais été surpassée dans des approches plus récentes utilisant les dernières techniques d'apprentissage profond (par exemple, $\rho = 0.325$ dans (Fritsch *et al.*, 2020), $\rho = 0.367$ dans (Amiriparian *et al.*, 2020), $\rho = 0.365$ dans (Egas-López *et al.*, 2022) ou $\rho = 0.383$ dans (Campbell *et al.*, 2022)).

Plus récemment, un nouveau grand corpus enregistré dans des conditions écologiques à l'aide de smartphones a été introduit : le corpus Voiceome (Tran *et al.*, 2022). L'équipe ayant développé ce corpus a rapporté un score F1 de 81.3% sur la classification binaire de la somnolence, mesurée par l'échelle de somnolence de Stanford (Hoddes *et al.*, 1973).

Parallèlement à ce travail se concentrant sur la somnolence à court terme, il est notable de mentionner nos précédents travaux sur le *Multiple Sleep Latency Test corpus* (MSLTc), contenant des enregistrements vocaux de patients hypersomniaques de la clinique du sommeil du CHU de Bordeaux étiquetés avec à la fois la somnolence subjective à court et à long terme (questionnaires) et physiologique (latence de sommeil mesurée par électroencéphalographie). Avec ces données, nous avons atteint des scores d'UAR supérieurs à 75% sur la détection de trois symptômes liés à la somnolence dans cette population (Martin *et al.*, 2024).

Limites. Au cours de la dernière décennie, la plupart des recherches utilisant ces corpus se sont concentrées sur le développement d'algorithmes d'apprentissage automatique pour estimer la

somnolence à partir des enregistrements de parole contenus dans ces corpus. En revanche, très peu d'attention a été accordée à l'élucidation du lien entre la somnolence et le comportement vocal. Depuis le travail fondateur de Krajewski *et al.* (2009), très peu d'études ont cherché à clarifier les mécanismes sous-jacents à l'expression de la somnolence dans la parole.

Parallèlement à ce travail d'apprentissage automatique, deux études perceptuelles ont récemment été menées sur le corpus SLEEP pour déterminer si l'oreille humaine peut estimer la somnolence à partir d'échantillons de parole. Ces deux études, basées sur 99 échantillons du corpus SLEEP, ont produit des résultats contradictoires : l'étude de Huckvale et al., impliquant 26 annotateurs, a conclu qu'il était possible de reconnaître la somnolence dans les enregistrements du corpus (Huckvale *et al.*, 2020). En revanche, notre étude de réplication, basée sur les annotations de 30 annotateurs naïfs, a obtenu des résultats moins enthousiastes (Martin *et al.*, 2023c). Puisque les participants de l'étude menée par Huckvale et al. étaient anglophones natifs et ceux de notre étude de réplication parlaient français, une explication à cette divergence entre les études aurait pu être expliquée par des différences dans les caractéristiques de la parole utilisées par les annotateurs pour estimer la somnolence, mais aucune de ces études n'a collecté de tels retours.

Objectif. L'objectif de cet article est d'identifier, a posteriori, les caractéristiques de la parole sur lesquelles les annotateurs se sont appuyés pour identifier la somnolence en réanalysant les données des deux études perceptuelles précédentes sur le corpus SLEEP (Huckvale *et al.*, 2020; Martin *et al.*, 2023c). Pour ce faire, sur la base d'un ensemble minimal de descripteurs extraits des enregistrements audio, nous avons entraîné plusieurs systèmes d'apprentissage automatique pour reproduire les annotations (un système d'apprentissage automatique de "clonage" par annotateur). Les caractéristiques extraites et le système d'apprentissage automatique ont été choisis simples et parfaitement explicables, permettant l'extraction et l'interprétation de l'importance relative de chaque caractéristique de parole dans l'imitation des annotateurs. Cette technique nous permet de dresser des profils d'annotateurs, déterminés en fonction de la manière dont ils identifient la somnolence à partir des enregistrements vocaux.

2 Méthode

Un aperçu de notre méthode est représenté dans la Figure 1.

Corpus et échantillons audio. Nous nous concentrons dans cet article sur les deux études perceptuelles impliquant le corpus SLEEP (Huckvale *et al.*, 2020; Martin *et al.*, 2023c). Le corpus entier contient plus de 16.464 échantillons de 915 sujets germanophones, enregistrés sur différentes tâches inconnues (Martin *et al.*, 2021). Tous les échantillons sont inférieurs à cinq secondes, avec une durée moyenne de 3.87 secondes. Ces échantillons ont été annotés en utilisant l'échelle de somnolence de Karolinska (KSS) (Åkerstedt & Gillberg, 1990), un questionnaire mesurant la somnolence subjective instantanée (Martin *et al.*, 2023b) utilisant une échelle de Lickert à 9 points. Les deux études perceptuelles ont utilisé le même sous-ensemble de 99 échantillons du corpus SLEEP, 9 pour familiariser les annotateurs avec la tâche (un pour chaque niveau de somnolence), et 90 (dix pour chaque niveau de somnolence) pour l'expérience elle-même. Notre analyse se concentre sur les 90 échantillons utilisés pour l'expérience.

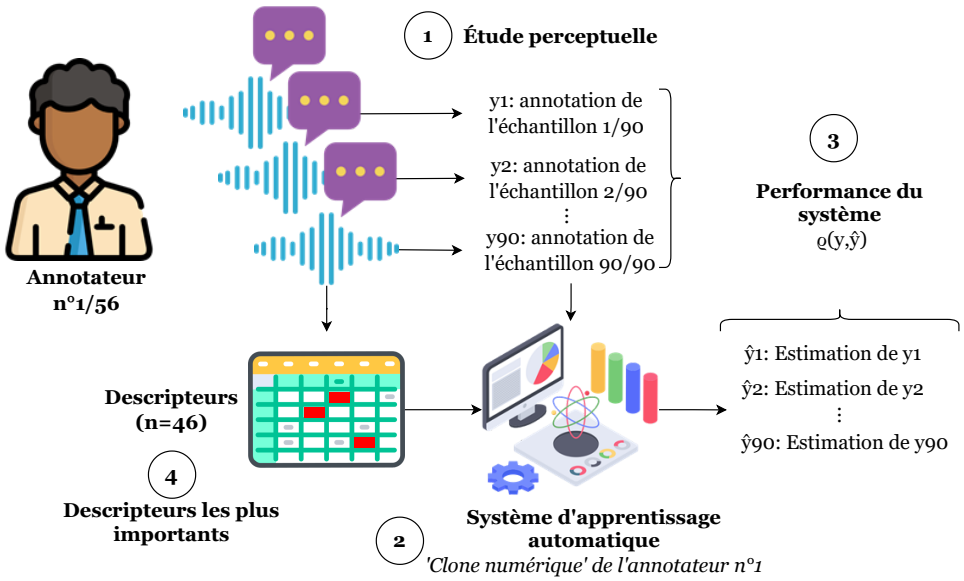


FIGURE 1 – Représentation schématique de notre méthode pour estimer les caractéristiques utilisées par les annotateurs des études perceptuelles pour estimer la somnolence à partir d'échantillons de parole

Études perceptuelles et annotateurs. Lors des deux études perceptuelles, les annotateurs étaient invités à estimer la somnolence de l'orateur à partir des enregistrements audio en utilisant une KSS à 9 points. Les échantillons étaient dans le même ordre pour les deux études, et les annotateurs ne pouvaient pas revenir en arrière. Les deux études sont arrivées à des conclusions différentes : alors que les annotations de l'étude de Huckvale et al. (Huckvale *et al.*, 2020), après application d'un algorithme de 'Sagesse des foules' (*Wisdom of the Crowd*), ont donné des performances très convaincantes ($\rho = 0.72$ entre l'estimation et la vérité terrain), les annotateurs de notre étude de réplication (Martin *et al.*, 2023c) n'ont pas atteint les mêmes performances ($\rho = 0.41$).

De plus, notre étude est la seule à avoir collecté les caractéristiques de chaque annotateur. Celles-ci incluaient leur genre (13F/17M), la sensibilité musicale (n=14 avaient des hobbies ou une profession liés à la musique ; n=16 n'en avaient pas), et leur compréhension de la langue allemande (« au moins un peu », n= 11 ; « pas du tout », n=19). Les autres caractéristiques de chaque étude sont décrites en détail dans un autre article (Martin *et al.*, 2023c).

Caractéristiques vocales. Puisque le sous-corpus sélectionné du corpus SLEEP contient peu d'échantillons par annotateur (90), et pour permettre l'interprétation des profils des annotateurs identifiés, nous nous sommes limités à 46 descripteurs extraits des enregistrements vocaux. Ils incluent la moyenne et l'écart type des caractéristiques de bas niveau (n=40) et les caractéristiques temporelles (n=6) de l'ensemble de descripteurs GEMAPS, extraits à l'aide de la boîte à outils `Opensmile` (Eyben & Schuller, 2015).

Systèmes d'apprentissage automatique. Pour pouvoir interpréter les coefficients des différentes parties du système d'apprentissage, nous avons choisi des algorithmes simples, qui ont précédemment montré leur efficacité sur des petits corpus :

(a) Lasso ($\alpha = 0.1$).

(b) Analyse en Composantes Principales (ACP, 80% de variance) + régression linéaire
(c) ACP (80% de variance) + régression à vastes marges (SVR, $C = 1$)

Un système différent a été entraîné pour chaque annotateur (n=26 pour Huckvale et al. 2020, n=30 pour Martin et al. 2023). De plus, pour comparer les performances des systèmes d'apprentissage automatique avec ceux de l'état de l'art pour la détection automatique de la somnolence à partir de la voix (cf. Introduction), nous avons également entraîné un système à reproduire les labels fournis avec le challenge IS2019. Ainsi, un total de 171 systèmes d'apprentissages (57 ensembles d'annotations $\times$ 3 systèmes) ont été entraînés.

Validation croisée et métrique de performance. Afin d'éviter un surapprentissage, les performances ont été calculées dans une procédure de validation croisée 5-fold, répétée 10 fois. En raison de la faible taille de l'échantillon, nous avons agrégé les estimations et les vérités terrain correspondantes, et calculé les performances sur les labels agrégées. De la même manière que pour le challenge IS2019, la métrique de performance choisie était la corrélation de Spearman ρ entre les labels estimés et la vérité terrain. Plus la valeur de ρ est élevée, meilleur est l'estimateur. Les étiquettes et les caractéristiques d'entrée ont été normalisées (z-score).

Contribution de chaque caractéristique. Pour chaque annotateur, nous avons mesuré la contribution de chaque caractéristique dans la chaîne de traitement entraîné pour l'imiter. Pour la chaîne de traitement utilisant uniquement un Lasso pour la classification (a), nous avons considéré les poids normalisés (norme L1) des classificateurs. Pour les autres régresseurs [ACP et régression linéaire (b) ou ACP et SVR (c)], nous avons calculé le produit croisé des coefficients de l'ACP et des coefficients du classificateur, que nous avons normalisé (norme L1). Ce faisant, nous mesurons la contribution de chaque caractéristique à une dimension donnée de l'ACP, qui est pondérée par la contribution de cette dimension de l'ACP à la classification. Pour chacun de ces coefficients, nous avons interprété séparément la valeur absolue – qui est liée à la contribution relative de la caractéristique à la classification – et le signe – qui indique la direction du lien entre la somnolence et la caractéristique vocale.

Lien entre les performances et les caractéristiques des annotateurs. Puisque les caractéristiques des annotateurs ont été collectées dans notre étude perceptuelle, nous avons calculé des tests de Mann-Whitney (MW) afin de mettre en lumière un lien possible entre le genre, la compréhension de la langue ou la sensibilité musicale des annotateurs, et les performances des systèmes entraînés pour reproduire leurs annotations.

Profils d'annotateurs. Afin de dresser les profils des annotateurs, nous avons sélectionné les caractéristiques les plus importantes, c'est-à-dire celles ayant une valeur médiane de contribution normalisée absolue supérieure à 0.05. Nous avons ensuite calculé les profils des annotateurs en utilisant le regroupement hiérarchique avec la fonction `linkage` de la bibliothèque `cluster.hierarchy` de `scipy` (Müllner, 2011). Le regroupement a été effectué en utilisant la méthode de Ward et une métrique euclidienne. Nous avons ensuite identifié les profils des annotateurs, c'est-à-dire les groupes tels que renvoyés par la fonction `linkage`. Pour chaque profil, nous avons pris en compte les performances des systèmes de régression correspondants pour être sûr qu'aucun profil n'était exclusivement constitué des descripteurs des systèmes ayant des performances faibles et que, au contraire, chaque profil était représenté par une diversité de performances.

3 Résultats

3.1 Performances des chaînes de traitement

Les moyennes et écarts-types des performances des sont rapportés dans le Tableau 1.

Ref	Modèle	Défi IS19	Huckvale et al. (n=26)	Martin et al. 2023 (n=30)
(a)	Lasso ($\alpha = 0.1$)	$\rho = 0.437$	$\rho = 0.049 \pm 0.166$	$\rho = 0.356 \pm 0.116$
(b)	ACP (0.8) + Régr. linéaire	$\rho = 0.459$	$\rho = 0.066 \pm 0.164$	$\rho = 0.323 \pm 0.100$
(c)	ACP (0.8) + SVR ($C = 1$)	$\rho = 0.447$	$\rho = 0.051 \pm 0.143$	$\rho = 0.289 \pm 0.095$

TABLE 1 – Performance des systèmes d'apprentissage automatique entraînés à imiter les annotateurs des études perceptuelles. Les valeurs sont calculées sur l'agrégation d'une validation croisée à 5-fold répétée dix fois, représentées sous la forme *Moyenne* $\pm$ *écart-type*

Sur le sous-corpus de 90 échantillons du corpus SLEEP, nos trois chaînes de traitement obtiennent des performances supérieures aux systèmes état-de-l'art sur l'ensemble du corpus (cf. Introduction), confirmant leur pertinence pour la tâche.

Sur les annotations de l'étude perceptuelle de Huckvale et al., aucun classificateur ne donne d'estimation satisfaisante des labels : tous les systèmes obtiennent des performances inférieures à ρ=0.283 et la plupart d'entre elles sont négatives, indiquant que le système n'a rien généralisé. En conséquence, nous ne les avons pas utilisés dans la suite. En revanche, le système (a) atteint un coefficient de corrélation moyen de $\rho = 0.356$ lors de la réplication des labels de notre étude perceptuelle, ce qui est dans l'ordre de grandeur des performances habituellement obtenues sur l'ensemble du corpus (cf. Introduction).

3.2 Influence des caractéristiques des auditeurs

Nous ne trouvons aucune différence dans les performances des systèmes de régression en fonction du sexe (MW, $U = 147, p = 0.132$), de la sensibilité musicale (MW, $U = 85, p = 0.271$), ou du niveau de compréhension de l'allemand (MW, $U = 145, p = 0.085$) de notre étude perceptuelle (Martin *et al.*, 2023c). Nous en déduisons donc que ces variables ne biaisent pas notre interprétation des caractéristiques de la parole impliquées dans l'estimation de la somnolence par les annotateurs.

3.3 Caractéristiques les plus saillantes

Parmi les 46 caractéristiques extraites, six sont identifiées comme les plus saillantes, c'est-à-dire ayant une valeur médiane de contribution normalisée absolue à travers les annotateurs supérieure à 0.05. Elles sont rapportées dans le Tableau 2.

3.4 Regroupement hiérarchique

Le regroupement hiérarchique a été effectué sur ces six caractéristiques pour dresser les profils des annotateurs dans notre étude perceptuelle. Nous avons identifié trois profils principaux, qui sont

représentés avec les caractéristiques les plus saillantes et la performance de chaque système de régression dans la Figure 2.

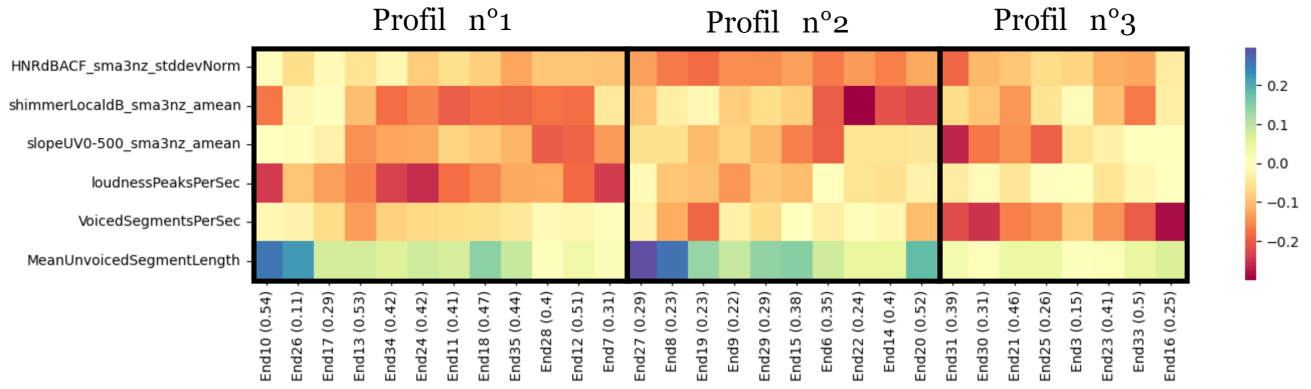


FIGURE 2 – Profils des annotateurs identifiés par le regroupement hiérarchique des systèmes entraînés à les imiter. Chaque ligne correspond à un descripteur, chaque colonne à un annotateur. La performance du système entraîné à reproduire chaque annotateur est indiquée entre parenthèse (ρ de Spearman).

Le premier groupe d'annotateurs (Profil n°1, n=12) associe la somnolence à une voix ayant des segments non voisés plus longs, et une voix plus douce (loudnessPeksPerSec) et moins expressive (slopeUV0-500 et shimmer), avec un accent particulier sur le volume. En revanche, les annotateurs du Profil n°2 (n=10) estiment la somnolence en utilisant des informations prosodiques (longueur des segments non voisés mais aussi le nombre de segments voisés par seconde), l'expressivité de la voix (slopeUV0-500, shimmer), mais aussi la pureté de la voix (variations de HNR). Enfin, les annotateurs du Profil n°3 (n=8) ne se basent pas sur la longueur du segment non voisé pour identifier la somnolence, mais se concentrent sur le nombre de segments voisés par seconde et la variabilité de la hauteur (slopeUV0-500).

4 Comparaison avec les approches automatiques

À notre connaissance, aucun système précédent travaillant sur le corpus SLEEP n'a étudié la contribution des descripteurs à l'estimation de la somnolence. Cependant, dans un de nos précédents travaux sur les tâches de lecture du *Sleepy Language Corpus* (même étiquette de somnolence que le corpus SLEEP) nous avions rapporté la corrélation entre les descripteurs acoustiques et la somno-

Nom	Description	médiane
`HNRdBACF_sma3nz_stddevNorm`	Écart-type du HNR	-0.102
`shimmerLocaldB_sma3nz_amean`	Moyenne du shimmer	-0.099
`slopeUV0-500_sma3nz_amean`	Pente de fréquence dans la bande passante [0,500Hz]	-0.095
`loudnessPeaksPerSec`	Pics d'énergie par seconde (moy.)	-0.09
`VoicedSegmentsPerSec`	Nombre de segments voisés par seconde	-0.065
`MeanUnvoicedSegmentLength`	Longueur moyenne des segments non voisés	0.075

TABLE 2 – Caractéristiques les plus saillantes dans la chaîne de traitement formé pour imiter les annotateurs dans notre étude perceptuelle (Martin *et al.*, 2023c). Les valeurs négatives signifient que la valeur de la caractéristique diminue lorsque la somnolence augmente.

lence (Martin *et al.*, 2019). Dans ce travail, les caractéristiques les plus corrélées à la somnolence étaient principalement liées à la fréquence fondamentale F0 (moyenne, max, min), la fréquence du premier formant (F1) et la plage d'énergie. À l'inverse, le HNR et la durée des segments voisés ou non voisés n'étaient pas parmi les caractéristiques les plus liées à la somnolence. De plus, en appliquant la même méthodologie à nos données, les caractéristiques les plus corrélées avec la vérité terrain donnée avec le corpus sont en partie celles identifiées comme saillantes dans l'imitation des annotateurs. En effet, alors que la pente de la fréquence fondamentale F0 ($\rho = -0.40$), le shimmer ($\rho = -0.34$) et le HNR ($\rho = -0.27$) sont fortement corrélés avec l'étiquette de somnolence, les pics de volume ($\rho = 0.10$), la durée des segments non voisés ($\rho = 0.13$) et le nombre de segments voisés ($\rho = -0.10$) ne sont pas parmi les caractéristiques les plus saillantes avec cette méthode.

Ces résultats questionnent le lien entre la vérité terrain donnée avec le corpus et ce que les annotateurs ont détecté. Le haut degré global d'accord inter-annotateurs rapporté dans notre étude perceptuelle (Martin *et al.*, 2023c) (ICC = 0.975) indique que les annotateurs semblent avoir identifié le même phénomène à travers la voix, qui n'est lui-même pas complètement représenté par l'outil de mesure utilisé pour opérationnaliser la somnolence dans le corpus SLEEP. Cependant, ce label est critiqué dans la littérature (Martin *et al.*, 2021), puisqu'il n'est pas une mesure de la somnolence validée, utilisée et reconnue en médecine du sommeil (Martin *et al.*, 2023b), et n'a jamais été utilisée ailleurs à notre connaissance que dans les deux corpus IS2011 et IS2019. De plus, une autre étude perceptuelle que nous avons menée sur le MSLTc (Martin *et al.*, 2023a), qui contient des mesures validées de la somnolence (Martin *et al.*, 2021), a conclu à la faisabilité de la détection de la somnolence par l'audition humaine à l'aide d'échantillons de parole. Nous interprétons donc cette différence entre les caractéristiques utilisées par les annotateurs et les caractéristiques corrélées avec l'étiquette fournie avec le corpus comme provenant de l'outil de mesure de la somnolence utilisé dans le corpus SLEEP.

5 Conclusion et perspectives

En entraînant des algorithmes d'apprentissage automatique à reproduire les annotations d'une étude perceptuelle sur la somnolence, nous avons pu identifier les caractéristiques sur lesquelles les annotateurs se sont appuyés pour produire cette évaluation ; et ainsi indirectement les indices qu'ils ont utilisés pour estimer la somnolence. Nous avons identifié six caractéristiques, liées à la stabilité de l'énergie (shimmer et pics d'énergie), le HNR, la variabilité de la fréquence fondamentale (pente de F0), et le ratio et la durée des segments voisés et non-voisés.

Nos prochains travaux se concentreront sur l'inclusion d'autres dimensions telles que les pauses de lecture (Martin *et al.*, 2022) ou la réalisation phonétique (Beaumard *et al.*, 2023) dans l'imitation du comportement d'annotation dans ces études perceptuelles.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche (ANR) dans le cadre de l'axe Autonom-Health du PEPR Santé Numérique, convention de subvention n°ANR-22-PESN-0009. VPM a reçu le soutien financier du programme de recherche et d'innovation européen Horizon Europe à travers le projet Marie Skłodowska-Curie MATER (No. 101106577). CB a reçu le soutien financier de la MITI du CNRS (projet PRIME 80 DSM-HEALTH).

Références

AMIRIPARIAN S., WINOKUROW P., KARAS V., OTTL S., GERCZUK M. & SCHULLER B. W. (2020). *A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech*. arXiv 2005.08722. _eprint : 2005.08722.

BARNES C. M. & WATSON N. F. (2019). Why healthy sleep is good for business. *Sleep Med. Rev.*, **47**, 112–118. DOI : 10.1016/j.smrv.2019.07.005.

BEAUMARD C., MARTIN V. P., WU Y., ROUAS J.-L. & PHILIP P. (2023). Automatic detection of schwa in French hypersomniac patients. In *Journée Santé et Intelligence Artificielle (Evènement affilié à PFIA 2023)*.

CAMPBELL E. L., DOCIO-FERNANDEZ L., GARCIA-MATEO C., WITTENBORN A., KRAJEWSKI J. & CUMMINS N. (2022). Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism. In *Workshop on Speech, Music and Mind*. DOI : 10.21437/SMM.2022-2.

EGAS-LÓPEZ J. V., BUSA-FEKETE R. & GOSZTOLYA G. (2022). On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection. In S. R. M. PRASANNA, A. KARPOV, K. SAMUDRAVIJAYA & S. S. AGRAWAL, Éds., *Speech and Computer*, Lecture Notes in Computer Science, p. 178–187, Cham : Springer International Publishing. DOI : 10.1007/978-3-031-20980-2_16.

EYBEN F. & SCHULLER B. (2015). Opensmile. *ACM SIGMultimedia Records*, **6**, 4–13.

FAGHERAZZI G., ZHANG L., ELBÉJI A., HIGA E., DESPOTOVIC V., OLLERT M., AGUAYO G. A., NAZAROV P. & FISCHER A. (2021). A Voice-Based Biomarker for Monitoring Symptom Resolution in Adults with COVID-19 : Findings from the Prospective Predi-COVID Cohort Study. *SSRN Journal*. DOI : 10.2139/ssrn.3949487.

FRITSCH J., DUBAGUNTA S. P. & MAGIMAI.-DOSS M. (2020). Estimating the Degree of Sleepiness by Integrating Articulatory Feature Knowledge in Raw Waveform Based CNNS. In *ICASSP 2020*, p. 6534–6538, Barcelona, Spain. DOI : 10.1109/ICASSP40776.2020.9053351.

GOSZTOLYA G. (2019). Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds. In *Interspeech 2019*, p. 2413–2417. DOI : 10.21437/Interspeech.2019-1726.

HODDES E., ZARCONE V., SMYTHE H., PHILLIPS R. & DEMENT W. C. (1973). Quantification of Sleepiness : A New Approach. *Psychophysiology*, **10**(4), 431–436. DOI : 10.1111/j.1469-8986.1973.tb00801.x.

HUANG D.-Y., GE S. S. & ZHANG Z. (2011). Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines. In *Interspeech 2011*, p.4̃.

HUCKVALE M., BEKE A. & IKUSHIMA M. (2020). Prediction of Sleepiness Ratings from Voice by Man and Machine. In *Interspeech 2020*. DOI : 10.21437/Interspeech.2020-1601.

JIKE M., ITANI O., WATANABE N., BUYSSE D. J. & KANEITA Y. (2018). Long sleep duration and health outcomes : A systematic review, meta-analysis and meta-regression. *Sleep Med. Rev.*, **39**, 25–36. DOI : 10.1016/j.smrv.2017.06.011.

KOLLA B. P., HE J.-P., MANSUKHANI M. P., FRYE M. A. & MERIKANGAS K. (2020). Excessive sleepiness and associated symptoms in the U.S. adult population : prevalence, correlates, and comorbidity. *Sleep Health*, **6**(1), 79–87. DOI : 10.1016/j.sleh.2019.09.004.

KRAJEWSKI J., BATLINER A. & GOLZ M. (2009). Acoustic sleepiness detection : Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods*, **41**(3), 795–804.

LÉGER D., BAYON V., LAABAN J. P. & PHILIP P. (2012). Impact of sleep apnea on economics. *Sleep Med. Rev.*, **16**(5), 455–462. DOI : 10.1016/j.smrv.2011.10.001.

MARTIN V. P., ARNAUD B., ROUAS J.-L. & PHILIP P. (2022). Does sleepiness influence reading pauses in hypersomniac patients ? In *Speech Prosody 2022*, p. 62–66 : ISCA. DOI : 10.21437/SpeechProsody.2022-13.

MARTIN V. P., FERRON A., ROUAS J.-L., SHOCHI T., DUPUY L. & PHILIP P. (2023a). Physiological vs. Subjective sleepiness : what can human hearing estimate better ? In *International Conference on Phonetic Science (ICPhS) 2023*, p. 196–200.

MARTIN V. P., LOPEZ R., DAUVILLIERS Y., ROUAS J.-L., PHILIP P. & MICOULAUD-FRANCHI J.-A. (2023b). Sleepiness in adults : An umbrella review of a complex construct. *Sleep Medicine Reviews*, **67**, 101718. DOI : 10.1016/j.smrv.2022.101718.

MARTIN V. P., ROUAS J.-L., FERRON A. & PHILIP P. (2023c). "Prediction of sleepiness ratings from voice by man and machine" : the Endymion replication perceptual study. In *International Conference on Phonetic Science (ICPhS) 2023*, p. 201–205.

MARTIN V. P., ROUAS J.-L., MICOULAUD-FRANCHI J.-A., PHILIP P. & KRAJEWSKI J. (2021). How to Design a Relevant Corpus for Sleepiness Detection Through Voice ? *Front. Digit. Health*, **3**, 686068. DOI : 10.3389/fdgth.2021.686068.

MARTIN V. P., ROUAS J.-L. & PHILIP P. (2024). Automatic detection of sleepiness-related symptoms and syndromes using voice and speech biomarkers. *Biomedical Signal Processing and Control*, **91**, 105989. DOI : https ://doi.org/10.1016/j.bspc.2024.105989.

MARTIN V. P., ROUAS J.-L., THIVEL P. & KRAJEWSKI J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania. DOI : 10.1109/SPED.2019.8906577.

MÜLLNER D. (2011). Modern hierarchical, agglomerative clustering algorithms. Publisher : arXiv Version Number : 1, DOI : 10.48550/ARXIV.1109.2378.

SCHULLER B., BATLINER A., BERGLER C., POKORNY F. B., KRAJEWSKI J., CYCHOCZ M., VOLLMAN R., ROELEN S.-D., SCHNIEDER S., BERGELSON E., CRISTIA A., SEIDL A., WARLAUMONT A., YANKOWITZ L., NÖTH E., AMIRIPARIAN S., HANTKE S. & SCHMITT M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech 2019*. DOI : 10.21437/Interspeech.2019-1122.

SCHULLER B., STEIDL S., BATLINER A., SCHIEL F. & KRAJEWSKI J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *Interspeech 2011*, p. 3201–3204. DOI : 10.21437/Interspeech.2011-801.

SCOTT A. J., WEBB T. L., MARTYN-ST JAMES M., ROWSE G. & WEICH S. (2021). Improving sleep quality leads to better mental health : A meta-analysis of randomised controlled trials. *Sleep Med. Rev.*, **60**, 101556. DOI : 10.1016/j.smrv.2021.101556.

TRAN B., ZHU Y., LIANG X., SCHWOEBEL J. W. & WARRENBURG L. A. (2022). Speech Tasks Relevant to Sleepiness Determined With Deep Transfer Learning. In *ICASSP 2022*, p. 6937–6941. ISSN : 2379-190X, DOI : 10.1109/ICASSP43922.2022.9747000.

ÅKERSTEDT T. & GILLBERG M. (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci*, **52**, 29–37. DOI : 10.3109/00207459008994241.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Disfluences en parole continue en français : paramètres prosodiques des répétitions

Ivana Didirková[1] Yaru Wu[2] Anne-Catherine Simon[3]
(1) UR 1569 TransCrit, Université Paris 8 Vincennes - Saint-Denis, France
(2) UR 4255 CRISCO, Université de Caen Normandie, France
(3) Centre VALIBEL, Institut Langage et Communication,
Université catholique de Louvain, Belgique
ivana.didirkova@univ-paris8.fr, yaru.wu@unicaen.fr,
anne-catherine.simon@uclouvain.be

RÉSUMÉ

Cette étude s'intéresse aux caractéristiques acoustiques des répétitions à l'identique à travers différents genres de parole en français. Le corpus d'étude LOCAS-F inclut 42 échantillons de parole représentant 14 activités de parole (ou genres). Ces échantillons sont caractérisés en fonction du degré de préparation, d'interactivité, etc. Le nombre de fois que les éléments sont répétés ne varie pas sensiblement en fonction du degré de préparation ou d'interactivité du discours. Du point de vue des durées, les résultats montrent que la durée totale des répétitions augmente avec le degré de préparation du discours. Cela vaut aussi pour la durée des séquences de répétitions avec les insertions. Du point de vue de la fréquence fondamentale, les écarts mélodiques au début d'une séquence de répétition sont plus étendus dans la parole spontanée (non préparée).

ABSTRACT

Disfluencies in continuous speech in French : prosodic parameters of repetitions

This study focuses on the acoustic characteristics of identical repetitions across different speech genres in French. The LOCAS-F study corpus includes 42 speech samples representing 14 speech activities (or genres). These samples are characterised according to the degree of preparation, interactivity, etc. The number of times elements are repeated does not vary significantly according to the degree of preparation or interactivity of the speech. From the point of view of duration, the results show that the total duration of repetitions increases with the degree of preparation of the speech. This is also true for the duration of sequences of repetitions with insertions. In terms of fundamental frequency, the melodic gaps at the start of a repetition sequence are wider in spontaneous (unprepared) speech.

MOTS-CLÉS : répétitions, disfluences, genres de parole, durée, fréquence fondamentale.

KEYWORDS: repetitions, disfluencies, speech genres, duration, pitch.

1 Introduction

Cette contribution traite des caractéristiques acoustiques des répétitions dans un corpus multigenre en français parlé. Elle vise à mieux décrire les phénomènes de répétitions en comparant différentes situations de parole. L'objectif est donc triple : décrire la forme des répétitions présentes dans le corpus dans son ensemble, analyser leur distribution en fonction des paramètres situationnels (comme le degré de préparation) et enfin mesurer les caractéristiques acoustiques des répétitions, en

particulier leur durée et leur fréquence fondamentale. Nous nous intéresserons particulièrement aux traits distinguant la parole préparée/non préparée et interactive/non interactive. En effet, nous nous attendons à ce que ces paramètres influencent la fréquence ou la forme des répétitions : les répétitions devraient être moins fréquentes dans un discours préparé (moins d'effort de planification) et plus courtes dans un discours peu interactif (moins de risque d'interruption). À plus long terme, ce type d'études permet de décrire quels types de répétitions sont "normales" (norme objective) en fonction de différentes situations de parole.

L'introduction permet de définir différents types de répétitions (1.1), d'expliciter la terminologie retenue pour l'annotation des données (1.2) et de présenter les principales caractéristiques acoustiques des répétitions relevées dans la littérature existante (1.3).

1.1 Différents types de répétitions

Traditionnellement, les répétitions sont considérées comme des disfluences ou des marques d'hésitation. Cependant, des travaux ont montré que toute répétition qu'on observe dans la langue parlée n'est pas disfluente. Blanche-Benveniste *et al.* (1990, p. 18-22) sont parmi les premiers à décrire de manière systématique les différentes formes de piétinement qui peuvent jalonner la progression de la parole. « Le déroulement syntagmatique est brisé [...] lorsque le locuteur répète [...]. Aucune règle du français ne pourrait nous inciter à enchaîner deux occurrences de *hier* pour y trouver une relation de dépendance » (p. 18). La répétition est un piétinement sur une seule et même place syntaxique. Si la plupart des répétitions sont disfluentes, certaines visent cependant un effet stylistique. Dans les exemples (1) à (3), les répétitions de *nous* et de *non* illustrent respectivement (1) une répétition à fonction syntaxique, chaque élément occupant une fonction différente dans la dépendance du verbe ; (2) une répétition disfluente, actualisant une hésitation ; (3) une répétition qu'on peut potentiellement qualifier d'intensive, et donc d'intentionnelle.

1. une fois par année **nous nous** réunissons et nous célébrons notre patrie sans fausse modestie (corpus LOCAS-F, pol-3)
2. **nous nous** vou/ **nous** vivons une époque catastrophique dans tous les domaines (corpus LOCAS-F, intlib-3)
3. **non non** il avait l'air hyper clean le gars (corpus LOCAS-F, conv-i-3)

Si les répétitions à fonction syntaxique (1) doivent être écartées, il n'est pas toujours aisé de distinguer entre les répétitions disfluentes et les répétitions stylistiques ou expressives (Blanche-Benveniste *et al.*, 1990, p. 21). Par conséquent, la plupart des études annotent l'ensemble de ces répétitions dans une seule catégorie (Crible *et al.*, 2015).

1.2 Terminologie

Une répétition est une disfluence complexe, composée de plusieurs éléments. Le modèle de Shriberg (1999) permet de décrire de manière générale la structure de la plupart des disfluences qui s'observent dans la parole. Le premier élément de la zone disfluente est le *reparandum*, le matériel verbal qui sera ultérieurement remplacé. La fin de cet élément marque l'interruption de la fluence verbale, dans la mesure où le locuteur a détecté un problème qui empêche le déroulement fluide de la production en cours. La phase d'édition se trouve entre l'élément à corriger (*reparandum*) et la réparation elle-même (*repair*), et elle peut contenir des éléments comme une pause silencieuse, une pause pleine ou des termes de type *enfin, je veux dire*, etc. La phase d'édition peut également être vide. Enfin, la réparation signale la reprise d'une parole fluente et la continuation du processus de production de la parole.

Ce modèle permet de distinguer, au sein d'une séquence formant une répétition, l'élément *répétable* et un deuxième élément, identique au premier, qui est nommé *répété* (Candea, 2000, p. 315). L'élément répétable peut être répété deux ou trois fois et ces différents éléments peuvent être séparés par d'autres marques du travail de formulation, comme des pauses silencieuses ou des pauses pleines de type *euh*.

1.3 Caractéristiques des hésitations

Comme d'autres auteurs, Grosjean & Deschamps (1975) constatent que les répétitions ne sont pas les marques d'hésitation les plus fréquentes parmi les « variables secondaires » de la parole, à savoir les pauses remplies, les syllabes allongées, les répétitions et les faux départs. Les répétitions représentent 16,80% de ces marques secondaires et deux tiers d'entre elles (64,22%) touchent des mots grammaticaux. En outre, 24,77% des répétitions concernent plusieurs mots. Le fait que les répétitions touchent davantage les mots grammaticaux que les mots lexicaux est rapporté par la plupart des études empiriques. Candea (2000, p. 316) observe des répartitions similaires à celles rapportées par Grosjean & Deschamps (1975) : les répétitions concernent 2,94% des mots outils et 1,43% des ligateurs (élément qui précise le lien avec ce qui précède, selon Morel & Danon-Boileau 1998, p. 20, cités par Candea 2000, p. 68) font l'objet d'une répétition, pour seulement 0,56% des mots lexicaux. Dans le corpus de Grosjean & Deschamps (1975, p. 180), on observe que 49% des répétitions sont précédées par une pause (silencieuse, pleine ou allongement). Candea (2000, p. 325) montre que la plupart des répétitions se combinent avec un allongement de l'élément répétable (53,93%). En outre, dans 82,42% des cas, la durée du mot répété est supérieure à celle du mot répétable (p. 335). Lorsqu'elles sont simples, sans combinaison avec une autre marque d'hésitation, les répétitions de mots outils sont très brèves (max. 90 s). Cela incite l'auteure à les considérer comme une forme de retard articulatoire de la production vocale, une forme de léger bégaiement (p. 327-328). Lorsqu'elles se combinent avec d'autres marques du travail de formulation, comme une pause silencieuse, un allongement ou une pause pleine de type euh, les répétitions forment des séquences plus longues qui peuvent approcher les 3 secondes. La combinaison avec un euh représente 39,6% des cas. Grosman (2018, p. 252) constate que les répétitions à l'identique forment la troisième marque la plus fréquente de disfluence, après les pauses silencieuses et les pauses pleines. Elles se produisent à une fréquence de 18,9 répétitions pour 1000 mots et représentent 5,5% du temps d'articulation. Analysant un corpus multigenre, elle observe que le degré de préparation est une caractéristique qui prédit un plus faible taux de répétitions (p. 254). Près de la moitié (45,83%) des répétitions sont produites de manière simple, sans autre terme d'édition. Lorsqu'une combinaison se produit, c'est avec une pause silencieuse qu'elle est la plus fréquente (29,7%), puis avec un marqueur de discours (10,1%) et enfin avec une troncation (9,6%) (Grosman, 2018, p. 259) . Les données de Grosman ne permettent pas de confirmer le fait que l'élément répétable est généralement plus long que l'élément répété, c'est le cas dans seulement 11% des cas (p. 270).

2 Méthode

2.1 Corpus et annotation

Le corpus LOCAS-F contient 42 échantillons de parole représentant 14 activités de parole pour une durée totale de 3h59 minutes et un nombre de 41 322 tokens. Les répétitions ont été annotées manuellement selon un protocole mis au point par Crible *et al.* (2015). Dans cette étude, les échantillons contenant des narrations conversationnelles ont été écartés en raison des nombreux chevauchements qui auraient pu fausser les mesures de f0. Nos analyses portent donc sur 2h38min35sec d'enregistrements (durée moyenne des enregistrements : 4min53). Les répétitions font partie des marques de

disfluence composées, comportant plus d'un élément. Tous les types de répétitions (cf. 1.1), disfluents ou stylistiques, sont annotés. Les répétitions à l'identique (RI) couvrent « un mot ou une séquence de mots (quasi-)contigus répétés formellement à l'identique » (p. 14). La quasi-contigüité rend compte de la possibilité d'insérer un ou plusieurs éléments disfluents entre l'élément répétable et l'élément répété, comme une pause vide (UP), une pause pleine de type *euh* (FP), un marqueur de discours (DM), un terme d'édition (ET) voire une insertion parenthétique (IP). Les éléments répétés sont numérotés de 0 (élément répétable) à N (en fonction du nombre de répétés). Les répétitions avec modulation (RM) font varier un élément et les répétitions grammaticales impliquent une forme différente appartement à la même catégorie grammaticale.

La fiabilité de l'annotation a été vérifiée au moyen d'un score d'accord interannotateurs portant sur 25,7 minutes de parole et incluant des parties du corpus LOCAS-F. « L'accord interannotateurs obtenu sur l'ensemble des données s'élève à K = 0,67 (raters = 2 ; z = 98,60 ; p < 0,001 ; n = 6892), ce qui correspond à un accord fort » (Grosman, 2018, p. 120).

Chaque échantillon de parole est également annoté en fonction de critères situationnels : le degré de préparation, le degré d'interactivité, le caractère public ou privé, médiatique ou non médiatique, etc. Concernant le degré de préparation, qui est un paramètre important pour notre étude, l'annotation comprend trois degrés : un discours est non préparé ou spontané lorsque le locuteur l'improvise au fur et à mesure et n'en prépare aucune partie par écrit ; un discours est semi-préparé lorsque le locuteur prépare le contenu global du discours sans le rédiger in extenso (la production orale peut inclure un support écrit partiel) ; le discours est préparé lorsqu'il a été entièrement rédigé par écrit, qui soit lu ou récité lors de la production orale.

2.2 Traitement des données

Différents types de répétitions sont disponibles dans nos données. Il s'agit plus spécifiquement de répétitions dites à l'identique (RI), où tous les éléments répétés sont les mêmes (par ex. « j'arrive dans j'arrive dans 5 minutes »), qu'elles soient ou non interrompues par un / plusieurs autres éléments (« j'arrive dans euh j'arrive dans 5 minutes »), et de répétitions dites avec modulation (RM), où la structure des éléments répétés est préservée mais au moins un élément sera modifié (« j'ai vu ton j'ai vu ta cousine »).

Dans cette étude, nous avons fait le choix de nous concentrer sur les séquences de mots répétés à l'identique (RI) et ce, afin d'éviter l'influence de la modification sur la longueur de l'élément répété et son comportement en termes de f0.

La première partie des analyses porte sur les répétitions à l'identique, avec ou sans éléments insérés, et vise notamment à décrire ces dernières en termes de durée (en tenant compte ou non des insertions) et en termes de nombre d'éléments répétés. Ces mesures sont d'abord analysées en fonction du degré de préparation du discours, puis en fonction du degré d'interactivité. En effet, nous supposons qu'un discours préparé devrait être moins sujet aux répétitions que des discours semi- ou non-préparés, puisque nécessitant moins de planification de la part du locuteur. D'un autre côté, les discours avec un degré d'interactivité plus important devraient favoriser des répétitions longues, afin de signaler à l'interlocuteur la volonté de continuer le discours et éviter ainsi une interruption par l'interlocuteur.

La seconde partie de nos analyses porte sur la fréquence fondamentale des répétitions. Pour mieux se concentrer sur la comparaison de la fréquence fondamentale entre les séquences de répétition et les syllabes environnantes, nous avons décidé de nous concentrer sur les séquences de répétition avec des mots qui ne sont répétés qu'une seule fois. Plus précisément, nous nous intéressons aux

intervalles mélodiques entre la première syllabe du mot avant sa répétition et la dernière syllabe du mot qui le précède (« RI0 - PrecSyl ») et aux intervalles mélodiques entre la première syllabe du mot suivant et la dernière syllabe du mot répété pour la première fois (« FollSyl - RI1 »), par rapport aux intervalles mélodiques entre les autres syllabes (« Autres »). Par exemple, la séquence de mots « bah la question » a été produite avec le mot « la » répété une seule fois (« $bah\ la_{(R0)}\ la_{(RI)}\ question$ »). Nous nous intéressons donc aux intervalles mélodiques entre la première syllabe du mot avant sa répétition « $\text{la}_{(R0)}$ » ([la]) et la dernière syllabe de « bah » ([ba]) et aux intervalles mélodiques entre la première syllabe du mot « question » ([kɛs]) et la dernière syllabe du mot répété pour la première fois « $\text{la}_{(RI)}$ » ([la]). Nous avons analysé environ 29k tokens pour chacune des analyses d'intervalles mélodiques (dont 225 « RI0 - PrecSyl » et 216 « FollSyl - RI1 »).

Les valeurs de fréquence fondamentale ont été extraites en utilisant Praat (Boersma & Weenink, 2024) et les valeurs mesurées en Hz ont été converties en demi-tons (DT) par rapport à 50 Hz.

Les syllabes ont été exclues de toute analyse lorsqu'elles sont indéfinies en termes de mesure de fréquence fondamentale (« undefined ») ou lorsque la syllabe qui les précède ou les suit est une pause remplie (« UP »).

3 Analyses et résultats

3.1 Nombre de répétitions

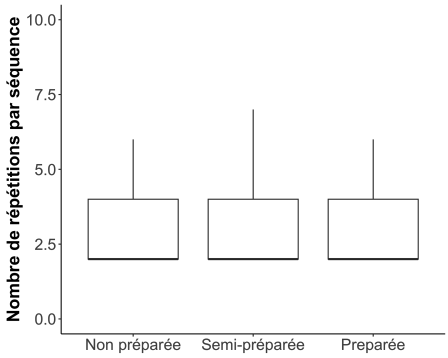


FIGURE 1 – Nombre de répétitions par séquence

La figure 1 montre qu'il n'y a pas d'effet du degré de préparation sur le nombre total d'éléments répétés dans chaque séquence analysée. En effet, dans les discours préparés, en moyenne, les locuteurs répètent 3,2 fois chaque élément, alors qu'ils les répètent en moyenne 2,92 fois dans les discours semi-préparés et 3,04 fois dans les discours non-préparés. Le constat est le même avec le degré d'interaction où aucun effet de ce dernier n'est observé sur la même variable, avec des éléments répétés 3,25 fois en moyenne dans les discours interactifs, 2,92 fois dans les discours semi-interactifs et enfin 2,91 fois en moyenne dans les discours non-interactifs.

3.2 Durée

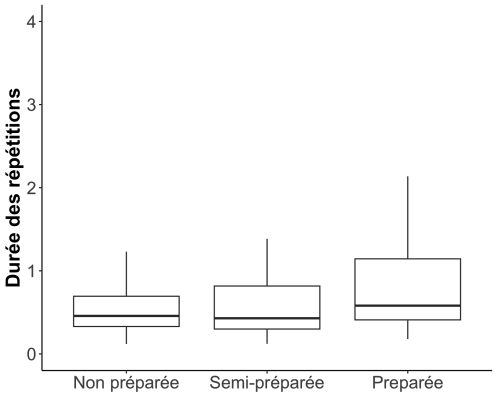


FIGURE 2 – Durée des répétitions en fonction du degré de préparation.

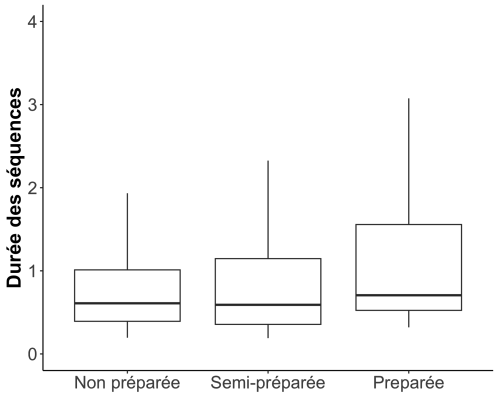


FIGURE 3 – Durée des séquences de répétition en fonction du degré de préparation.

Lorsque l'on s'intéresse à la durée totale des répétitions (cf. Figure 2, hors insertions éventuelles, c'est-à-dire uniquement la durée des éléments répétés à l'identique), l'on s'aperçoit que cette durée varie de manière significative en fonction du degré de préparation. Plus concrètement, les répétitions sont plus longues dans les discours préparés (0,92s en moyenne, ET = 0,75s, p = 0,004, t = 2,981, SE = 0,112) que dans les discours non-préparés (moyenne = 0,58s, ET = 0,43s). Il n'y a pas de différence significative entre les discours semi-préparés (moyenne : 0,64s, ET = 0,53s) et non-préparés. En revanche, ce même paramètre, à savoir la durée totale des répétitions, ne semble pas être influencé par le caractère interactif ou non des discours (non-interactif : durée moyenne 0,72s, ET = 0,58s ; semi-interactif : durée moyenne 0,58s, ET = 0,47s ; interactif : durée moyenne 0,59s, ET = 0,43s).

Enfin, comme le montrent les LMM, il existe également une variation significative de la durée totale de la séquence (répétitions + éventuelles insertions) en fonction du degré de préparation d'un discours (Figure 3). En effet, la séquence répétée est significativement plus longue dans les discours préparés (moyenne = 1,22s, ET = 1s) que dans les discours non-préparés (moyenne = 0,81s, ET = 0,64s, p = 0,03, t = 2,233, SE = 0,174). Il n'y a pas de différence significative avec les discours semi-préparés (moyenne = 1s, ET = 0,98s). Si l'on s'intéresse à la durée totale de la séquence répétée en fonction du degré d'interactivité, l'on s'aperçoit que les discours non-interactifs entraînent une durée de la séquence disfluente plus longue (moyenne = 1,13s, ET = 1,06s) que les discours interactifs (moyenne = 0,83s, ET = 0,66s, p = 0,04, t = 2,107, SE = 0,125). Ici encore, il n'y a pas de différence significative avec les discours semi-interactifs (moyenne = 0,83s, ET = 0,69s).

3.3 Intervalle mélodique

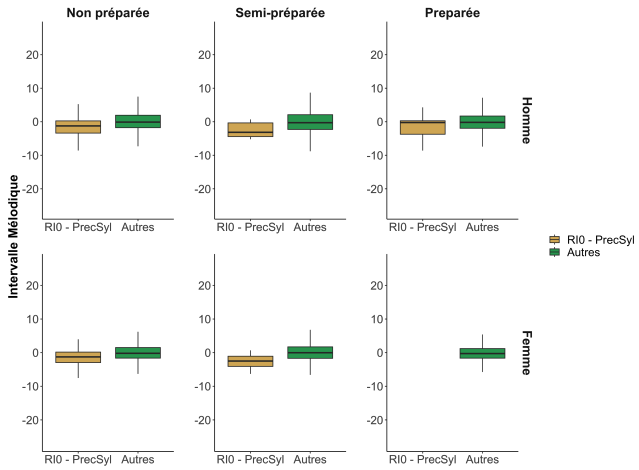


FIGURE 4 – Intervalle mélodique (en DT) entre la dernière syllabe du mot précédent et la première syllabe du mot en RI (RI0), en fonction du degré de préparation et du genre (de gauche à droite : parole non préparée, semi-préparée et préparée ; partie en haut pour les hommes, partie en bas pour les femmes).

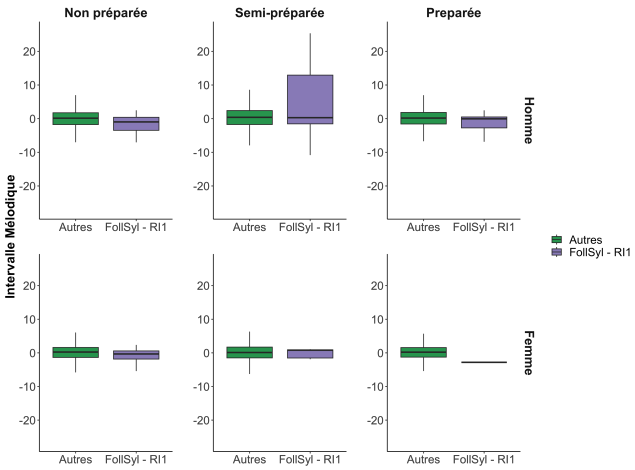


FIGURE 5 – Intervalle mélodique (en DT) entre la dernière syllabe du mot répété (RI1) et la première syllabe du mot suivant, en fonction du degré de préparation et du genre (de gauche à droite : parole non préparée, semi-préparée et préparée ; partie en haut pour les hommes, partie en bas pour les femmes).

Les résultats sur les intervalles mélodiques sont présentés dans la figure 4 et la figure 5, pour les hommes (en haut) et les femmes (en bas) et pour les trois degrés de préparation (de gauche à droite : parole non préparée, semi-préparée et préparée).

Comme nous pouvons le voir dans la figure 4, il n'y a pas de données de locutrice en parole préparée dans l'ensemble de données sélectionné. L'intervalle mélodique est plus faible pour « RI0-PrecSyl » que pour « Autres » en général, pour les locuteurs masculins et féminins, ainsi que pour les trois différents degrés de préparation. Les résultats basés sur le modèle LMM confirment que l'intervalle mélodique est significativement plus faible pour « RI0-PrecSyl » (β = -5,671e-01 ; t = 2,120 ; SE = 2,675e-01) par rapport à celui observé pour « Autres ». L'intervalle mélodique est significativement plus élevé pour la parole non-préparée (β = 3,076e-01 ; t = 3,755 ; SE = 8,193e-02), comparé à celui observé pour la parole préparée. Aucune différence significative n'est observée entre les paroles préparée et semi-préparée, ni entre les hommes et les femmes.

Des tendances similaires sont observées pour les hommes et les femmes pour la parole non préparée (première colonne à partir de la gauche) et pour la parole préparée (troisième colonne à partir de la gauche) dans la figure 5 : le changement d'intervalle mélodique est plus faible pour « FollSyl-RI1 » que pour « Autres ». Pour la parole semi-préparée, aucune différence significative n'est observée. Les résultats basés sur le modèle LMM confirment que l'intervalle mélodique est significativement plus faible pour « FollSyl-RI1 » (β = -1,470e+00 ; t = 5,107 ; SE = 2,879e-01) par rapport à celui observé pour « Autres ». L'intervalle mélodique est significativement plus élevé pour la parole non préparée (β = -1,734e-01 ; t = 2,620 ; SE = 6,620e-02), comparé à celui observé pour la parole préparée. La différence n'est pas significative entre les paroles préparée et semi-préparée, ni entre les hommes et les femmes.

4 Discussion, limites et perspectives

L'objectif de cette étude était de mieux décrire les répétitions identiques en parole continue en français du point de vue de leur durée, du nombre d'éléments répétés et de la fréquence fondamentale, le tout en corrélation avec le degré de préparation du discours (préparé, semi-préparé, non-préparé) et le degré d'interactivité (interactif, semi-interactif, non-interactif).

Les résultats portant sur le nombre d'éléments répétés ne montrent aucune corrélation significative entre cette variable et le degré de préparation ou d'interactivité d'un discours.

En ce qui concerne la durée totale des répétitions, celle-ci est plus importante lors d'un discours préparé que lorsque le discours présente davantage de spontanéité. Ce résultat peut paraître surprenant : l'on pourrait s'attendre à ce qu'un discours préparé ne nécessite pas de séquences disfluentes particulièrement longues, a fortiori si l'on considère qu'une disfluence résulte d'un manque de préparation de la séquence à suivre. Cependant, Simon (2021) a montré que les répétitions à l'identique sont une manière fréquente d'hésiter dans des discours extrêmement solennels. De même, on pourrait expliquer cette durée quasiment doublée de la répétition dans les discours préparés par le fait que le locuteur, seul, n'a pas besoin d'accélérer sa production de la parole puisqu'il n'est pas concurrencé par un interlocuteur qui pourrait l'interrompre (Hirsch *et al.*, 2016). Enfin, ces durées accrues des répétitions et des séquences de répétitions dans les discours préparés et non-interactifs pourraient s'expliquer par une corrélation avec la variation du débit, plus lent dans les genres plus formels (Simon *et al.*, 2010).

Enfin, l'étude sur l'intervalle mélodique montre que la différence entre la syllabe précédant le premier élément répété et la première syllabe du mot répété est plus faible qu'entre des syllabes qui ne sont pas concernées par les répétitions. La même tendance est observée entre la dernière syllabe de la répétition et la première syllabe du mot suivant. Ce résultat montre que les premières et dernières syllabes des éléments répétés seraient davantage intégrées prosodiquement, autorisant moins de sauts de f0 que d'autres éléments présents dans les discours. Si une tendance similaire a été observée dans notre précédente étude sur les allongements et les pauses pleines (Wu *et al.*, 2022) entre la disfluence et la syllabe suivante, le résultat montrant que la première syllabe de la répétition est plus intégrée prosodiquement que la moyenne indique un comportement différent des répétitions comparé aux deux autres types de disfluences. Ce constat serait à approfondir en analysant les intervalles mélodiques entre les différents éléments répétés (par exemple entre deux « je suis je suis »).

Nos analyses présentent un certain nombre de limites. Tout d'abord, nous avons, pour cette étude, exclu les répétitions comprenant des modifications d'un ou plusieurs éléments. Or, il serait intéressant de comparer leur fréquence fondamentale à celle des types de disfluences semblables, à savoir les reprises et les répétitions à l'identique. Cette comparaison permettrait de savoir si les répétitions modifiées se rapprochent davantage de l'un ou de l'autre type, ou si elles se situent à mi-chemin entre les deux. En effet, les répétitions avec modification (auto-correction) montrent une capacité de contrôle de la production orale, dont le degré varie probablement selon la situation de parole. Nous avons également exclu les cas complexes de répétitions à l'identique. Il s'agit notamment de cas où plusieurs répétitions d'un élément se retrouvent entrecoupés par des insertions et sont modifiées avant d'être reprises à l'identique ("je je euh je euh je suis euh j'étais je suis") et pour lesquelles une catégorie spécifique devrait être créée.

Enfin, nous entendons poursuivre nos analyses notamment en incluant d'autres types de disfluences (faux-départs, reprises, etc.) et en comparant les paramètres étudiés dans ces différents types de disfluences, dans l'objectif de déterminer s'il existe des différences saillantes en fonction du type de la disfluence (voir aussi l'étude de Wu *et al.*, 2022).

Références

BLANCHE-BENVENISTE C., MERTENS P. & WILLEMS D. (1990). Le français parlé : études grammaticales. *Éditions du CNRS*.

BOERSMA P. & WEENINK D. (2024). Praat : doing phonetics by computer [computer program]. *http ://www. praat. org/*.

CANDEA M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits" d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français.* Thèse de doctorat, Université de la Sorbonne nouvelle-Paris III.

CRIBLE L., DUMONT A., GROSMAN I. & NOTARRIGO I. (2015). " annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs v. 1.0.

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

GROSMAN I. (2018). *Évaluation contextuelle de la (dis) fluence en production et perception : pratiques communicatives et formes prosodico-syntaxiques en français*. Thèse de doctorat, UCL-Université Catholique de Louvain.

HIRSCH F., MARSAC F., DIDIRKOVA I., BECHET M. & MESSAOUD M. A. B. (2016). Spécificités du rythme de la parole politique. le cas de françois hollande. *Romanica Wratislaviensia*, p. p–145.

MOREL M.-A. & DANON-BOILEAU L. (1998). *Grammaire de l'intonation l'exemple du français*. Editions Ophrys.

SHRIBERG E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, p.2̃.

SIMON A.-C. (2021). De l'écrit à l'oral dans les discours solennels : le cas de l'académie française. *Linguistique de l'écrit*, **2**, éparants.

SIMON A.-C., AUCHLIN A., AVANZI M., GOLDMAN J.-P. *et al.* (2010). Les phonostyles : une description prosodique des styles de parole en français. *Les voix des Français : en parlant, en écrivant, Bern : Lang*, p. 71–88.

WU Y., DIDIRKOVÁ I. & SIMON A.-C. (2022). Disfluences en parole continue en français : paramètres prosodiques des pauses pleines et des allongements vocaliques. In *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, p. 900–909. DOI : 10.21437/JEP.2022-95.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Effet de la tâche sur le débit articulatoire d'enfants et adolescents avec et sans trouble du spectre de l'autisme en français.

Cwiosna Roques[1,] Fanny Guitart-Ivent[1,] Christelle Dodane[2,] Fabrice Hirsch[1]

(1) Praxiling UMR 5267, 34 000 Montpellier, France
(2) Clesthia EA 7345, 75 005 Paris, France
cwiosna.roques@univ-montp3.fr, fanny.guitart-ivent@univ-montp3.fr, christelle.dodane@sorbonne-nouvelle.fr, fabrice.hirsch@univ-montp3.fr

RESUME

Cette étude comparative propose d'observer le débit de parole de 8 enfants de 10 à 16 ans avec un trouble du spectre de l'autisme (TSA), et celui de 8 enfants tout-venants appariés, dans deux activités extraites du module 3 de l'ADOS-2, la première consistant à raconter une histoire à partir d'images puis de la mimer, la seconde, à parler librement sur le thème de l'amitié. Nos résultats montrent que les enfants avec TSA parlent plus lentement que les locuteurs contrôles appariés, dans les deux tâches et qu'ils parlent davantage entre deux pauses en parole libre qu'en description d'histoire.

ABSTRACT

Effect of task on articulatory speed of children and adolescents with and without autism spectrum disorder in French.
This comparative study observes the speech rate of 8 children aged 10 to 16 with an autism spectrum disorder (ASD), and that of 8 matched normal children, in two activities extracted from module 3 of the ADOS-2. The first task consists in telling a story from pictures and then miming it, the second consists in free speech on the theme of friendship. Our results show that children with ASD speak more slowly than matched control speakers in both experimental contexts, and that they speak more between pauses in free speech than in story description.

MOTS-CLES : Trouble neurodéveloppemental, autisme, enfants, adolescents, débit, parole, motricité verbale.
KEYWORDS: Neurodevelopmental disorder, autism, children, adolescents, flow, speech, verbal motor skills.

1 Introduction

Le Trouble du Spectre de l'Autisme (TSA) est un Trouble du Neurodéveloppement (TND) dont les marqueurs diagnostiques se répartissent selon deux groupes de symptômes : des niveaux d'altération de la communication et des interactions sociales d'un côté, et des stéréotypies, qui se manifestent par des centres d'intérêts et d'activités restreints de l'autre (DSM-5, 2013 ; ICD-11, World Health Organization, 2022). De nombreuses recherches ont été menées afin de tenter de définir les comportements de communication propres à l'autisme. Certaines d'entre elles ont

retenu, sur l'ensemble du spectre, des atypicités prosodiques pouvant constituer un élément sémiologique et donc participer au diagnostic. Celles-ci apparaissent autour des 3 ans et se manifestent par une accentuation et un rythme inaccoutumé (Peeters, 1996), un timbre marqué par une voix nasale ou enrouée, une intonation (Frith, 1989) souvent qualifiée de « monotone » (Perrin & Maffre, 2013), notamment car la personne avec TSA a du mal à percevoir (Baltaxe & Gutbrie, 1987) et à reproduire la prosodie émotionnelle, à découper les mots et à marquer l'accentuation d'emphase (Paul et al., 2008). D'autres chercheurs relèvent également des variations d'intensité atypiques (avec des cris et/ou des chuchotements sans lien avec le contenu sémantique de l'énoncé) et un débit de parole trop rapide ou trop lent (Globerson et al., 2015 ; Loveall et al., 2021). Parmi ces différents phénomènes, nous allons nous focaliser sur les aspects temporels de la parole. Le débit est particulièrement intéressant à étudier chez des enfants avec TSA car il est lié non seulement à la structure rythmique et phonotactique des langues (Rouas et al. 2004), mais également au traitement cognitif. Ainsi, selon certains auteurs, la parole spontanée serait plus exigeante sur le plan cognitif que la parole lue en raison d'efforts de planification plus importants (Tasko et McLean, 2004 ; Van-Lancker-Sidtis et Rallong, 2004). Une autre étude (O'Keefe et al., 2022) échoue à montrer des différences significatives entre des tâches de parole spontanée et de parole lue, mais montre en revanche des différences en fonction de la vitesse de traitement de l'information et des fonctions exécutives.

La nature de la tâche langagière paraît donc affecter les aspects temporels (parole spontanée, description d'images, etc.) et selon sa nature, le débit peut varier de façon conséquente (Colletta et al., 2016). Ainsi, on relève de nombreuses variations de résultats obtenus sur des mesures de débit en fonction des méthodes adoptées comme des tâches proposées. Dans l'étude de Colletta et al. (2016), les résultats varient en fonction de l'âge et de la tâche. Dans la tâche narrative, le débit passe de 3,8 syll/sec. chez les enfants de 3-4 ans à 4,10 syll/sec. chez les enfants de 9-10 ans et dans la tâche explicative, de 3,61 syll/sec. chez les enfants de 3-4 ans à 4,40 syll/sec. chez les enfants de 9-10 ans. Les auteurs observent bien une augmentation du débit de parole liée à l'âge dans les deux tâches, mais cette progression n'est pas statistiquement significative. La plupart des études relève également une augmentation régulière du débit de parole des enfants en fonction de l'âge (Martins et al., 2007), même s'ils parlent plus lentement que des adultes (Koopmans-van Beinum, 1993 ; Colletta et al., 2016), dont le débit moyen est de 5 à 7 syll/sec. (Koopmans-van Beinum, 1993). Chez des enfants et adolescents de 5 à 17 ans dans une tâche narrative, Martins et al. (2007) trouvent une moyenne de 91,2 mots par minute avec peu de pauses.

Bien que des atypicités prosodiques aient été signalées depuis les premières observations de TSA par Kanner et Asperger (Asperger et Frith, 1991 ; Kanner, 1943), la méta-analyse de McCann et Peppé (2003) portant sur 16 recherches menées sur la prosodie chez des personnes avec un TSA a révélé de nombreux résultats non significatifs ou contradictoires. C'est notamment le cas pour les études sur le débit de parole, avec soit une augmentation, soit une diminution chez les sujets TSA en comparaison des sujets neurotypiques ou tout-venants (Baron-Cohen et Staunton, 1994 ; Shriberg et al., 2001 ; Patel et al., 2020). La variabilité de ces résultats s'expliquerait par la taille insuffisante des échantillons de population étudiés, un faible nombre de données sur les locuteurs contrôles, une absence de protocole standardisé et des méthodologies non détaillées (MacCann et Peppé, 2003 ; Patel et al., 2020). En outre, la plupart de ces recherches a utilisé des mesures subjectives (c'est-à-dire des jugements perceptifs) qui, bien que valables sur le plan clinique, n'offrent pas de catégorisation fine et précise des différences entre les groupes de sujets. Des travaux plus récents ont cependant apporté des éléments de réponse au sujet de l'atypicité prosodique des sujets avec un TSA en fonction des situations et notamment lors d'une tâche consistant à raconter une histoire émotionnelle (Edelson et al., 2007 ; Patel et al., 2020). Seule l'étude de Patel et al. (2020) fait des mesures de débit de parole en activité de narration et obtient une vitesse articulatoire (en syllabes par secondes) plus lente chez les sujets avec TSA (hommes et

femmes confondus) que les sujets contrôles, mais sans comparaison avec une activité de conversation libre. Notre objectif est donc de rechercher si la tâche de narration a un effet sur la vitesse articulatoire avant de pouvoir déterminer si les sujets avec TSA parlent plus lentement que leurs pairs contrôles sans que ce résultat ne soit la conséquence de cette tâche narrative.

Le débit reste cependant relativement peu étudié chez les enfants avec TSA et les résultats tendent à se contredire entre eux. Pour Patel et al. (2020), il serait significativement plus faible que chez l'enfant apparié neurotypique, tandis que deux autres études n'aboutissent pas à des écarts significatifs (Nadig et Shaw, 2012 ; Ochi et al. 2019). Par ailleurs, il existe très peu de travaux menés en français. De plus, chez les enfants avec TSA, les troubles fonctionnels de la motricité sont décrits depuis longtemps (Morange-Majoux et Adrien, 2016) et l'habileté concernant les mouvements orofaciaux reste, parmi les trois habiletés motrices souvent traitées dans les batteries de test (les deux autres étant les actions sur les objets d'un côté, les gestes des mains et la posture globale de l'autre), souvent la plus sévèrement altérée, notamment en situation d'imitation (Rogers et Benetto, 2002). Partant de ces éléments, notre étude a pour objectif d'apporter des connaissances supplémentaires sur le débit de parole des enfants avec un TSA. Notre hypothèse est que le débit est ralenti chez l'enfant autiste en raison d'une motricité orofaciale dysfonctionnelle. De même, nous postulons que le débit est plus diminué lorsque l'enfant doit élaborer seul son propos, sans pouvoir se reposer sur de l'observation.

Outre le débit articulatoire, nous avons également choisi d'étudier la longueur des unités inter-pausales (désormais IPU) qui sont des blocs de parole encadrés par des pauses silencieuses d'un minimum de 200 ms en français (Bertrand et al., 2008). En effet, certains travaux (Georgeton et Meunier, 2015) ont montré que les IPU étaient intéressantes à étudier dans le cadre d'une comparaison entre sujets typiques et atypiques, en l'occurrence chez des malades de Parkinson. Par ailleurs, leur caractère objectif, formel et identifiable automatiquement (Koiso et al., 1998) facilite le découpage du signal de parole et la transcription de grands corpus, comme celui de la thèse en cours dont l'échantillon de cette étude est extrait. Nous avons ainsi décidé de comparer le débit et la durée des IPU chez une population de 8 enfants avec TSA et une population de 8 enfants contrôles appariés en sexe et en âge lors de deux tâches langagières, une tâche de parole spontanée et une tâche de description d'une histoire présentée sous forme d'une suite d'images. Nous postulons que les sujets avec TSA parlent plus entre deux pauses dans la tâche de description, car celle-ci est plus coûteuse cognitivement sur le plan de l'élaboration, de la planification.

2 Méthodologie

Notre corpus est composé de 16 entretiens audiovisuels (8 enfants avec TSA âgés de 10 à 16 ans et 8 enfants contrôles appariés) de l'ADOS-2, l'échelle diagnostique internationale de référence dans la détection des TSA. L'ADOS-2 se divise en 4 modules distincts, avec une évolution graduelle adaptée au niveau de communication et à l'âge du sujet. Nous travaillons à partir d'extraits du module 3, qui concerne des enfants et adolescents avec TSA ou suspicion de TSA, au langage verbal fluide ou plutôt fluide. Ce module se décompose en 14 thèmes et activités, parmi lesquels nous avons sélectionné un extrait de conversation spontanée sur le thème de l'amitié (parole libre : PL) et une activité dont la consigne est de décrire une histoire à partir d'images (DI).

Les enfants du groupe TSA ont été diagnostiqués au CHU Lapeyronie de Montpellier et sont intégrés à la cohorte de recherche sur les TSA nommée ELENA[1]. Le groupe TSA sélectionné pour notre étude est constitué de 5 garçons et 3 filles, pour un âge moyen de 12,5 ans (avec un écart

[1] Cohorte ELENA du CeAND de Montpellier (Baghdadli et al., 2018)

type de 1,9). La durée moyenne des extraits de PL est de 247 sec. et celle des extraits de DI de 135,25 sec.

Le groupe contrôle (CTRL), recruté au laboratoire Praxiling UMR 5267 CNRS comprend 6 garçons et 2 filles, pour un âge moyen de 12,8 ans (avec un écart-type de 1,8). La durée moyenne des extraits de PL est de 135,8 sec. et celle des extraits de DI de 244,5 sec.

Afin d'analyser les paramètres de débit articulatoire des sujets du corpus, les données ont d'abord été exportées au format .wav pour permettre la transcription orthographique manuelle sur Praat (Boerma et Weenink, 2018). Ensuite la transcription a été segmentée de manière automatique avec le script EasyAlign (Goldman, 2011), de façon à obtenir 3 niveaux de plus que celui de la transcription orthographique des énoncés préalablement réalisée : une segmentation phonétique, par mots et par syllabes. L'alignement des énoncés a ensuite été vérifié manuellement et les erreurs de transcription automatique en SAMPA ont été rectifiées.

Nous avons extrait le nombre de syllabes et la durée (en secondes) des IPU en fixant un seuil de pause silencieuse à 200 ms, selon les critères utilisés par Campione et Veronis (2002) parmi d'autres. À partir de ces mesures, nous avons calculé le débit articulatoire en nombre de syllabes par seconde (pauses exclues). A partir d'une base de 1232 IPUs, nous avons écarté celles de moins de quatre syllabes pouvant contenir des marques d'hésitation, des pauses pleines et des éléments constitués d'un seul mot. Finalement, nous avons analysé le débit à partir de 812 observations dont la répartition entre nos Groupes de sujets (TSA vs. CTRL) et Tâches (PL vs. DI) est donnée dans le Tableau 1.

	TSA	CTRL
PL	298 (37%)	203 (25%)
DI	114 (14%)	197 (24%)

TABLEAU 1 : Effectifs de nos deux groupes de sujets pour chaque Tâche de parole.

Afin de comparer le débit des enfants avec TSA avec celui du groupe CTRL dans nos deux tâches de parole, un modèle mixte a été conduit sur R (R Core Team, 2008) avec la fonction *lmer* du package lme4 (Bate et al., 2015). Ainsi nous avons testé les relations entre le débit articulatoire et le groupe de sujet (TSA vs. CTRL) en interaction avec la tâche (PL vs. DI). Un intercept aléatoire par sujet a été modélisé ainsi qu'une pente aléatoire par locuteur pour le facteur Tâche. Un modèle similaire a été effectué sur la durée des IPUs sans inclure de pente aléatoire pour des raisons de convergences. Les valeurs de p ont été obtenues par approximations de type Satterthwaite à l'aide de la fonction *lmerTest*. Le seuil de référence a été fixé à $p<.05$. Les effets de chaque facteur fixe et de leur interaction ont été testés par comparaison de modèles avec la fonction anova. Les valeurs de R2 associées à chaque modèle ont été obtenues à l'aide de la fonction *r.squaredGLMM* intégrée dans la bibliothèque 'MuMIn'. Enfin, l'analyse a posteriori des contrastes a été réalisée à l'aide de la fonction *lsmeans* de la bibliothèque "emmeans" (Lenth et al., 2018) avec des ajustements de la valeur p de Tukey.

3 Résultats

L'ensemble des analyses statistiques sont résumés dans le Tableau 2 (panel du haut : débit articulatoire ; panel du bas : durée des IPU).

Facteurs & interactions	χ2	Effet débit articulatoire	β	SE	t
Groupe	6.48^{*}	CTRL > TSA	-0.90	0.31	-2.92
Tâche	0.48^{ns}	PL = DI	-0.02	0.17	-0.14
Groupe*Tache	1.02^{ns}	-	0.26	0.25	1.06
Facteurs & interactions	**χ2**	**Effet durée d'IPU**	**β**	**SE**	**t**
Groupe	2.04^{ns}	TSA = CTRL	0.48	0.23	2.12
Tâche	24.07^{***}	PL > DI	-0.27	0.12	-2.27
Groupe*Tache	4.64^{*}	TSA PL > TSA DI	-0.39	0.18	-2.16

TABLEAU 2. Test du rapport de vraisemblance des modèles linéaires mixtes testant les effets principaux des prédicteurs Groupe et Tâche et leurs interactions sur le débit articulatoire (haut) et la durée des IPUs (bas). Les valeurs χ2 (avec un degré de liberté) sont indiquées comme estimations de l'ampleur de l'effet (p<.05 = *, p<.001 = **, p<.0001 = ***). Estimations : les β-coefficients, les erreurs standard (SE) et les valeurs t du modèle sont également rapportés.

3.1 Analyse du débit

Le modèle mixte (R^2m = 0.01, R^2c = 0.28) révèle un effet du Groupe sur le débit articulatoire. En revanche, aucun effet de la Tâche, ni d'interaction entre le Groupe et la Tâche ne ressort. Comme nous pouvons le voir Figure 1, les sujets CTRL ont un débit articulatoire plus élevé que les sujets avec TSA (β = 0.93, p = 0.02) quelle que soit la Tâche de parole.

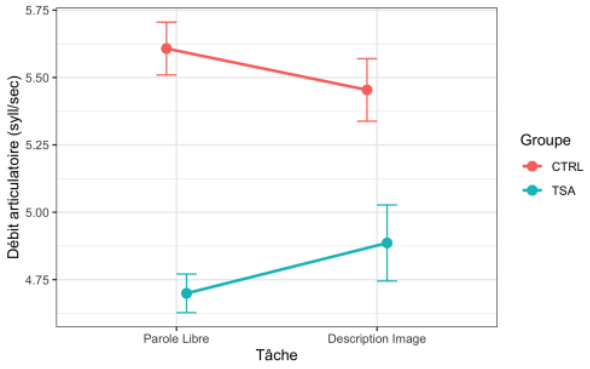


Figure 1. Débit articulatoire (syll./sec) de nos de groupes de sujets (TSA et CTRL) pour chaque Tâche (PL et DI).

3.2 Analyse de la durée d'IPU

Concernant la durée d'IPU, le modèle mixte (R^2m = 0.06, R^2c = 0.16) ne révèle pas d'effet du Groupe mais un effet de la Tâche ainsi qu'une interaction entre nos deux facteurs Groupe et Tâche. Ainsi, en PL, les IPU sont plus longues qu'en DI (β = 0.44, p <0.0001) et cette différence est significative uniquement pour les sujets avec TSA (β = 0.66, p <0.0001) comme illustré sur la Figure 2.

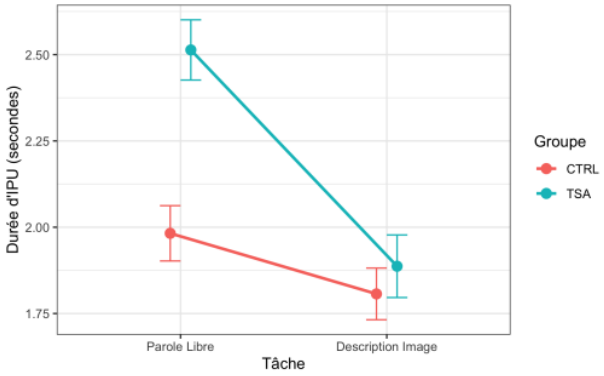


Figure 2. Durée des IPU en secondes de nos groupes de sujets (TSA et CTRL) pour chaque tâche de parole (PL et DI).

4 Discussion et conclusion

Cette étude examine le débit articulatoire d'enfants avec un TSA en parole libre et en activité semi-contrôlée de narration d'histoire courte à partir d'images, en comparant les résultats à ceux d'enfants contrôles appariés ayant suivi le même protocole de l'ADOS-2. Nos résultats révèlent que les enfants avec TSA parlent avec un débit articulatoire plus lent que leurs pairs contrôles, en DI comme en PL. Les données traitées n'ont pas permis d'observer de variations intra-groupes significatives en fonction de la tâche.

Ces résultats confirment ceux de l'étude menée par Patel et al. (2020) et pourraient donc indiquer que les enfants avec TSA parlent plus lentement que les enfants tout-venant. Il faudrait néanmoins augmenter la taille l'échantillon de la population pour affiner cette observation, ce qui est en cours dans le cadre de la thèse menée sur les données utilisées dans cet article. Si ces résultats se maintiennent en agrandissant le corpus de travail, cela pourrait nous amener à nous interroger sur les habiletés motrices oro-bucco-faciales chez les enfants autistes, et sur leur implication dans le débit articulatoire. Des habiletés réduites pourraient notamment être à l'origine de la réduction du débit articulatoire.

L'étude perceptive menée par Redford et al. (2018) postulait déjà que l'intelligibilité de la parole des personnes autistes était affectée par ce que les auditeurs percevaient comme des troubles de l'élocution. Nos résultats pourraient également révéler des difficultés d'élaboration du discours dans la mesure où des recherches antérieures sur des populations typiques ont montré un lien direct

entre une charge cognitive accrue et une baisse de la vitesse d'élocution (Griffin et Williams, 1987 ; Huttunen et al., 2011).

Concernant la durée des IPU, nous observons que celles-ci sont plus longues dans la tâche DI par rapport à la tâche PL chez les deux groupes d'enfants. Notons que cette différence est particulièrement significative dans le groupe TSA, dont les sujets ont tendance à parler davantage entre deux pauses dans le cas de la tâche narrative que dans la parole spontanée. Ceci tend à confirmer l'hypothèse qu'un effort cognitif supplémentaire a un impact sur le débit articulatoire, comme indiqué supra. L'élaboration en situation contrôlée serait plus coûteuse qu'en situation spontanée chez les enfants avec un TSA.

En résumé, ce travail a révélé des différences de débit articulatoire et de longueur d'IPU entre les sujets avec un TSA et les sujets contrôles, en utilisant des mesures acoustiques. Il semblerait que les enfants avec un TSA parlent plus lentement que les contrôles et que l'activité leur coûte un effort de communication supplémentaire. Le débit de parole étant un élément majeur d'identification de modèles de parole atypique, des analyses complémentaires sur un plus vaste corpus de données permettraient de confirmer ces résultats et d'approfondir la question de la cause (neurologique et/ou motrice).

Des recherches supplémentaires sont nécessaires pour comprendre les différents facteurs cognitifs et socio-pragmatiques qui peuvent influencer le débit de parole chez les enfants et préadolescents avec un TSA. Il serait notamment intéressant d'observer comment le débit interagit avec d'autres caractéristiques prosodiques décrites dans la littérature comme atypiques, telles que la variation de la fréquence fondamentale ou l'étendue vocale (Paul et al., 2008), et de créer des sous-groupes par âge afin d'obtenir des profils prosodiques plus précis et plus complets.

Dans l'ensemble, ces résultats soulignent l'importance de cibler la vitesse articulatoire dans les interventions orthophoniques et le fait que de telles interventions pourraient être bénéfiques pour les capacités langagières pragmatiques globales. Pourtant, il existe peu d'interventions basées sur la prosodie pour les personnes avec un TSA (Diehl et Paul, 2009), bien que celles-ci tendent à se développer depuis les dernières années (Sicard et Menin-Sicard, 2021). En effet, les différences relevées dans la présente étude peuvent servir de repères importants pour les intervenants en orthophonie, ce qui souligne l'importance d'une collaboration entre les cliniciens et les chercheurs dans ce domaine.

Références

AMERICAN PSYCHIATRIC ASSOCIATION. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

ASPERGER, H., & FRITH, U. (1991). "Autistic psychopathy" in childhood. In Autism and Asperger syndrome (pp. 37–92). Cambridge: Cambridge University Press.

BAGHDADLI, A., MIOT, S., RATTAZ C., et al. (2019). Investigating the natural history and prognostic factors of ASD in children: the multicEntric Longitudinal study of childrEN with ASD - the ELENA study protocol. BMJ Open, 9:e026286. doi:10.1136/bmjopen-2018-026286.

BALTAXE, C., Use of contrastive stress in normal, aphasic, and autistic children, Journal of speech and Hearing Research, 27(1), 1984, 97-105.

BALTAXE, C., GUTHRIE, D., The use of primary sentence stress by normal, aphasic, and autistic children, Journal of Autism and Developmental Disorders, 17(2), 1987, 255-271

BARON-COHEN, S., & STAUNTON, R. (1994). Do children with autism acquire the phonology of their peers? An examination of group identification through the window of bilingualism. First Language, 14(42–43), 241–248. https://doi.org/10.1177/014272379401404216.

BATES, D., MÄCHLER, M., BOLKER, B., & WALKER, S. (2014). Fitting linear mixed-effects models using lme4. ArXiv Preprint ArXiv:1406.5823.

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRE, G., MEUNIER, C. et al. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. Revue TAL : traitement automatique des langues, 2008, 49 (3), pp.105-134. hal-00349893.

CAMPIONE, E., VERONIS, J. Pauses et hésitations en français spontané. JEP 2004. Actes des XXVe Journées d'Etudes sur la Parole, Fès, Maroc, 19-22 avril, 2004.

COLLETTA, J. M., PELLENQ, C., ROUSSET, I. Evolution du débit de parole chez l'enfant francophone dans des tâches narrative et conversationnelle. 27èmes Journées d'Etudes sur la Parole, Association Francophone de la Communication Parlée, Jun 2008, Avignon, France. hal-01292877

DIEHL, J. J., & PAUL, R. (2009). The assessment and treatment of prosodic disorders and neurological theories of prosody. International Journal of Speech-Language Pathology, 11(4), 287–292. https://doi.org/10.1080/17549500902971887

DUEZ, D. & NISHINUMA, Y. (1985). Le rythme en français. Travaux de l'Institut de Phonétique d'Aix, 10, 151-169.

EDELSON, L., GROSSMAN, R., & TAGER-FLUSBERG, H. (2007). Emotional prosody in children and adolescents with autism. Poster session presented at the annual international meeting for Autism Research, Seattle, WA.

FRITH, U., A new look at language and communication in autism, International Journal of Language & Communication Disorders, 24, 1989, 123-150. Globerson et al., 2015

GEORGETON, L. & MEUNIER, C. (2015). Spontaneous speech production by dysarthric and healthy speakers: temporal organization and speaking rate. ICPHs 2015, 310.

GRIFFIN, G. R., & WILLIAMS, C. E. (1987). The effects of different levels of task complexity on three vocal measures. Aviation, Space and Environmental Medicine, 58(12), 1165–1170.

GROSJEAN, F. ET DESCHAMPS, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica, 31, 144-184.

HUTTUNEN, K. H., KERÄNEN, H. I., PÄÄKKÖNEN, R. J., PÄIVIKKI ESKELINEN-RÖNKÄ, R., & LEINO, T. K. (2011). Effect of cognitive load on articulation rate and formant frequencies during simulator flights. The Journal of the Acoustical Society of America, 129(3), 1580–1593. https://doi.org/10.1121/1.3543948.

ICD-11 : International Classification of Diseases, World Health Organization, 2022

KANNER, L. (1943). Autistic disturbances of affective contact. Nervous Child, 35(2), 217–250. https://doi.org/10.1105/tpc.11.5.949.

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

F.J. KOOPMANS-VAN BEINUM. Cyclic effects of infant speech perception, early sound production, and maternal speech. IFA (Institut de Phonétique d'Amsterdam) Proceedings, 17 : 65-78, 1993.

LENTH, R., LOVE, J., & LENTH, M. R. (2018). Package 'lsmeans'. The American Statistician, 34(4), 216-221.

LORD, C., RUTTER, M., DILAVORE, P., et al., ADOS, Autism diagnostic observation schedule. Manual. Los Angeles: WPS, 1999.

LOVEALL, S-J., HAWTHORNE, K., GAINES, M., A meta-analysis of prosody in autism, Williams syndrome, and Down syndrome, Journal of Communication Disorders, Volume 89, 2021

MAFFRE, T., PERRIN, J., Autisme et psychomotricité, De Boeck-Solal, 2013.

Martins IP, Vieira R, Loureiro C, Santos ME. Speech rate and fluency in children and adolescents. Child Neuropsychol. 2007 Jul;13(4):319-32. doi: 10.1080/09297040600837370. PMID: 17564849.

MCCANN, J., PEPPE, S., Prosody in autism spectrum disorders: a critical review, International Journal of Language and Communication Disorders, 38, 2003, 235-350.

MORANGE-MAJOUX Françoise, ADRIEN Jean-Louis, « Motricité et préférence manuelle chez les enfants avec troubles du spectre de l'autisme : une nouvelle voie d'exploration des troubles, à partir d'une revue de la littérature », Devenir, 2016/4 (Vol. 28), p. 213-227. DOI : 10.3917/dev.164.0213. URL : https://www.cairn.info/revue-devenir-2016-4-page-213.htm

NADIG, A., & SHAW, H. (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners. Journal of Autism and Developmental Disorders, 42(4), 499–511.

O'KEEFFE C, YAP SM, DAVENPORT L, COGLEY C, CRADDOCK F, KENNEDY A, TUBRIDY N, LOOZE C, SULEYMAN N, O'KEEFFE F, REILLY RB, MCGUIGAN C. (2002) Association between speech rate measures and cognitive function in people with relapsing and progressive multiple sclerosis. doi: 10.1177/20552173221119813. PMID: 36003923; PMCID: PMC9393591.

OCHI, K. et al. Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. PLoS ONE 14, e0225377. https://doi.org/10.1371/journal.pone.0225377 (2019).

PATEL, S.P., NAYAR, K., MARTIN, G.E. et al. An Acoustic Characterization of Prosodic Differences in Autism Spectrum Disorder and First-Degree Relatives. J Autism Dev Disord 50, 3032–3045 (2020). https://doi.org/10.1007/s10803-020-04392-9

PAUL, R., BIANCHI, N., AUGUSTYN, A., KLIN, A., VOLKMAR, F., Production of syllable stress in speakers with autism spectrum disorders, Research in Autism Spectrum Disorders, 2(1), 2008, 110–124.

PEETERS, T., ROGÉ, B., FRANCO, G., L'autisme: de la compréhension à l'intervention, Dunod, 1996.

R CORE TEAM (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

REDFORD, M. A., KAPATSINSKI, V., & CORNELL-FABIANO, J. (2018). Lay listener classification and evaluation of typical and atypical children's speech. Language and Speech,61(2), 277–302. https://doi.org/10.1177/0023830917717758.

ROGERS, S., BENETTO, L., Le fonctionnement moteur dans le cas de l'autisme, Enfance, 2002/1 (Vol. 54). DOI : 10.3917/enf541.0063.

ROQUES, C. (en cours) Description et évaluation de la prosodie et de la gestualité chez des enfants et adolescents au langage fluide avec un Trouble du Spectre de l'Autisme (TSA). Université Paul-Valéry Montpellier 3.

ROUAS, J.L., FARINAS, J. & PELLEGRINO, F. (2004). Évaluation automatique du débit de la parole sur des données multilingues spontanées/ XVèmes Journées d'Études sur la Parole, 437-440.

SHRIBERG, L., PAUL, R., MCSWEENY, J., KLIN, A., & VOLKMAR, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. Journal of Speech, Language, and Hearing Research,44(5), 1097–1115.

SICARD, E., MENIN-SICARD, A., Analyse acoustique de la prosodie dans le cadre de la clinique orthophonique. 2021. ⟨hal-03177645⟩

TASKO, SM, MCCLEAN, MD (2004). Variations in articulatory movement with changes in speech task. J Speech Lang Hear Res, 47, 85–100.

VAN LANCKER-SIDTIS D. & RALLON, G. (2004). Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification. Lang Commun, 24: 207–240.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude de la qualité vocale dans la parole professionnelle des aides-soignants français

Jean-Luc Rouas[1] Yaru Wu[2] Takaaki Shochi[1, 3]
(1) Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
(2) CRISCO/UR4255, Université de Caen Normandie, 14000 Caen, France
(3) CLLE CNRS UMR 5263, Bordeaux, France
rouas@labri.fr, yaru.wu@unicaen.fr, takaaki.shochi@labri.fr

RÉSUMÉ

Cet article présente une méthodologie complète pour étudier les attributs vocaux des aides-soignants travaillant dans des maisons de retraite en France. L'objectif était d'analyser les modèles de parole de 20 aides-soignants dans deux établissements distincts. Les aides-soignants ont été équipés de microphones-casque connectés à des smartphones pour garantir une qualité audio optimale. Les données enregistrées comprenaient la lecture de texte, des entretiens informels et des jeux de rôle professionnels avec des patients fictifs. Le traitement des données a été effectué à l'aide d'un système de reconnaissance automatique de la parole de pointe, permettant de générer des séquences de mots ou de phonèmes avec leurs frontières. L'analyse s'est concentrée sur la détection des variations de la qualité vocale dans divers contextes de parole spontanée. L'objectif final est le développement d'outils de formation automatisés pour les aides-soignants, afin de capturer et reproduire leurs caractéristiques vocales uniques, améliorant ainsi leurs capacités professionnelles.

ABSTRACT

Voice quality in French Caregivers' Professional Speech.

This paper presents a comprehensive methodology for investigating the vocal attributes of caregivers in retirement homes in France. The aim was to analyze the speech patterns of 20 proficient caregivers across two facilities. Caregivers were equipped with headset microphones connected to smartphones for high-quality audio. Data included reading text, informal interviews, and professional role-play scenarios with fictitious patients. Data processing involved a state-of-the-art automatic speech recognition system, generating word or phone sequences with timestamps. Analysis focused on detecting nuanced variations in voice quality in diverse spontaneous speech contexts. The ultimate goal is to develop cutting-edge automated training tools tuned to capture and replicate caregivers' unique vocal characteristics, thereby enhancing their caregiving capabilities.

MOTS-CLÉS : Qualité de voix, aides-soignants, styles de parole, parole spontanée.

KEYWORDS: Voice quality, caregivers voice, speaking styles, spontaneous speech.

1 Introduction

Une communication verbale efficace est essentielle lors de la prise en charge de personnes âgées dépendantes. Cependant, dans les milieux hospitaliers, les aides-soignants sont souvent accaparés par leurs tâches, ce qui entraîne des interactions brèves avec les patients. Un rapport alarmant a révélé que

la communication verbale avec les patients atteints de démence alités dans les établissements de soins de longue durée durait seulement deux minutes par jour (Gineste *et al.*, 2008). Les aides-soignants peuvent trouver décourageant que les patients ne réagissent pas ou de manière non pertinente.

Des recherches ont montré que la communication volontaire et positive des aides-soignants professionnels peut avoir un fort impact sur les patients âgés atteints de démence (Gineste & Pellissier, 2007). La méthode "Humanitude" est conçue pour promouvoir des interactions positives et continues grâce au développement de compétences de communication efficaces, basées sur le contact visuel, la communication verbale et l'interaction tactile. De nombreuses études ont démontré que cette méthode entraîne une réduction significative (88,5 %) des comportements agressifs des patients et une diminution de la nécessité de médicaments neuroleptiques (Honda *et al.*, 2013, 2016; Ito & Honda, 2015).

En ce qui concerne les compétences en communication verbale, la méthode "Humanitude" s'appuie sur des éléments phonétiques et lexicologiques, ainsi que sur une technique appelée "Auto-Feedback". Dans cette technique, les aides-soignants se doivent de parler sans interruption, même lorsque les bénéficiaires de soins fournissent des réponses inadéquates. Par conséquent, il existe deux principales catégories de paramètres qui nécessitent une investigation : les paramètres prosodiques (tels que l'intensité, le débit et la mélodie) qui devraient être en accord avec la voix douce, calme et mélodieuse recommandée, et les éléments lexicaux destinés à véhiculer des émotions positives.

L'étude des paramètres prosodiques de la parole professionnelle des aides-soignants a été menée dans Rouas *et al.* (2023b) tandis que les attributs affectifs ont été examinés dans Rouas *et al.* (2023a). L'analyse acoustique réalisée dans Rouas *et al.* (2023b) a montré que des valeurs plus élevées pour $F0$ sont observées pour les enregistrements professionnels des aides-soignants. Bien que cela semble contraire à ce qui peut être attendu en ce qui concerne la voix "douce" recommandée par "Humanitude", nous formulons ici l'hypothèse que les changements de qualité vocale, comme les modifications de la quantité de souffle, peuvent avoir un impact sur la perception de la douceur ou de la proximité de la voix. L'objectif de cet article est donc l'étude des attributs de qualité vocale de la parole des aides-soignants.

Nous résumons brièvement dans la Section 2 les informations qui peuvent être véhiculées à l'aide des attributs de qualité vocale. Le protocole d'enregistrement du corpus produit par les aides-soignants professionnels français avec la description de l'équipement spécifique utilisé pour permettre la liberté de mouvement, les tâches effectuées et les paramètres d'enregistrement sont décrits dans la Section 3. Ensuite, la Section 4 décrit comment nous avons prétraité les fichiers pour obtenir la transcription phonétique qui est ensuite utilisée pour extraire les caractéristiques de qualité vocale sur les Unités Inter-Pausales. L'analyse des caractéristiques est réalisée dans la Section 5 et nous discutons des résultats dans la Section 6.

2 Qualité de la voix

La variabilité de la qualité de la voix laryngée, telle que le souffle (« breathiness ») et le craquement (« creakiness »), est observée de manière étendue dans la parole. Ces variations peuvent servir à diverses fins linguistiques, comme l'utilisation contrastée de la qualité de la voix dans les langues ou en tant que caractéristiques prosodiques comme le craquement phrastique final dans plusieurs langues. Plus important encore pour notre étude actuelle, il existe des preuves de distinctions sociolinguistiques

dans la qualité de la voix (Stuart-Smith, 1999).

Les travaux de Gobl & Ní Chasaide (2003) et de Yanushevskaya *et al.* (2006) montrent que la qualité de la voix seule peut évoquer des associations affectives, même si celles-ci n'existent pas sur une base univoque ; une qualité de voix particulière, par exemple le craquement, peut être associée à plusieurs états affectifs. Cependant, la perception des affects exprimés par les qualités vocales varie selon les langues (Yanushevskaya *et al.*, 2018). En anglais, une voix chuchotée peut être associée à la peur (de Mareüil *et al.*, 2002), tandis que dans d'autres langues, elle peut véhiculer d'autres émotions. De même, en anglais, la voix soufflée est traditionnellement associée à l'intimité (Laver, 1980), alors qu'en japonais, elle est plus souvent liée à la formalité et à la politesse (Ito, 2004; Ishi *et al.*, 2008). De manière similaire, la voix craquée tend à évoquer différents états émotionnels en fonction du contexte linguistique. En français, Grichkovtsova *et al.* (2012) montrent que la perception des attitudes repose principalement sur le contour prosodique, tandis que les émotions reposent à la fois sur la qualité de la voix et sur le contour prosodique.

En plus de la relation entre qualité de voix et émotions, certains chercheurs ont également essayé d'établir des liens entre différentes qualités de voix et la personnalité perçue d'un locuteur en anglais américain (Pearsell & Pape, 2023). Alors que la voix craquée tend généralement à être perçue de manière négative, des recherches suggèrent que la voix soufflée influence principalement la perception de la voix d'une locutrice en termes de sexualité et de sensualité (Laver, 1980). De plus, la voix soufflée peut être liée à la perception de la solidarité (Pittman, 1985).

3 Protocole d'enregistrement

3.1 Équipement

Pour garantir une mobilité totale lors des enregistrements, nous avons créé un dispositif entièrement autonome. Nous avons équipé nos sujets d'un microphone directionnel haut de gamme, le casque DPA 4288 CORE, connecté à un préamplificateur iRIG PRO. Ce préamplificateur et un smartphone Samsung Galaxy A51 étaient placés dans un sac banane, offrant une liberté de mouvement totale, tout en maintenant un enregistrement de haute qualité.

3.2 Tâches

En utilisant cet équipement, nous avons enregistré nos sujets sur trois tâches différentes :

Lecture de texte : lecture à voix haute de *La bise et le soleil*. Le but de cette tâche est d'enregistrer un échantillon vocal contrôlé dans un contexte très structuré.

Entretien informel : réponses à des questions ouvertes axées sur le travail du soignant et sa routine quotidienne. Le but de cette tâche est de collecter des données spontanées dans une interaction "naturelle". Cet exercice aide également à renforcer la confiance du locuteur.

Tâche de soin professionnelle : réaliser une tâche de soin sur un patient fictif non réactif. La tâche de soin sélectionnée était l'habillage, ce qui comprenait le boutonnage d'une chemise et des techniques de réveil du corps le matin. Après l'habillage, le soignant aide le patient (simulé) à se lever et à marcher. Pour évaluer la technique de « Auto-Feedback », le patient reste totalement silencieux tout

en permettant au soignant d'administrer les soins. Le cadre a été intentionnellement conçu pour être familier au soignant, ressemblant à une chambre typique, avec des cloisons pour un sentiment d'intimité avec le patient simulé.

3.3 Données collectées

Les enregistrements audio ont été collectés dans deux établissements d'hébergement pour personnes âgées dépendantes (EHPAD) dans le sud-ouest de la France : « Les Balcons du Lot » à Prayssac et « Les Résidences du Quercy Blanc » à Castelnau-Montratier. Trois sessions d'enregistrement ont eu lieu : deux à l'établissement de Prayssac le 24 septembre 2021 et le 25 mars 2022, et une à Castelnau-Montratier le 24 novembre 2021.

Au total, 25 participants ont été enregistrés lors de ces sessions, dont 21 femmes et 4 hommes. Pour l'analyse, nous avons exclu les enregistrements des 4 participants masculins et 1 enregistrement d'une participante féminine en raison de la mauvaise qualité d'enregistrement. Nous avons donc conservé 20 participants pour une durée combinée de 2 heures et 30 minutes. Les durées spécifiques des tâches et les durées moyennes par locuteur par tâche sont spécifiées dans le Tableau 1.

Tâche	durée moyenne	durée totale (s)
Lecture de texte	45,0 s.	15 min 03 s.
Entretien	134,8 s.	44 min 56 s.
Tâche de soin	188,9 s.	62 min 58 s.

TABLE 1 – Durée moyenne par locuteur et par tâche et durée totale par tâche. L'ensemble des 20 sujets enregistrés a participé à chaque tâche.

4 Paramètres

4.1 Transcription orthographique et phonétique automatique

Nous avons utilisé un système automatisé, basé sur le framework Kaldi (Povey *et al.*, 2011), pour générer des transcriptions orthographiques et phonétiques. Ce système a été entraîné en utilisant la base de données ESTER (Galliano *et al.*, 2009) et utilise un réseau neuronal à retard temporel (Time Delayed Neural Network - TDNN) couplé à un modèle de Markov caché. Le TDNN est composé de 7 couches, chacune avec 1024 unités. Le modèle acoustique prend en entrée un vecteur MFCC haute résolution de 40 dimensions concaténé avec un i-vecteur de 100 dimensions (Gupta *et al.*, 2014). Ce système atteint un taux d'erreur de 13,7% sur l'ensemble de test du corpus ESTER (Boyer, 2021), ce qui est proche des performances de l'état de l'art sur le même corpus (légèrement inférieur à 12% WER (Heba, 2021)). Les symboles phonétiques et leur alignement sont obtenus à l'aide de la commande *lattice-align-phones*, ce qui permet de segmenter et d'annoter 35 phonèmes. La transcription phonétique de sortie est utilisée pour calculer les caractéristiques moyennes sur chaque unité phonétique.

4.2 Mesures de la qualité de la voix

Au cours des deux dernières décennies, la qualité de la voix laryngée a suscité un intérêt croissant dans les domaines de l'articulation, de l'acoustique et de la perception. Les phonéticiens ont activement recherché des attributs acoustiques capables de distinguer entre les voix modales, craquées et soufflées. Les dimensions articulatoires fondamentales qui contribuent à la qualité de la voix laryngée sont le degré de constriction des plis vocaux et la présence de bruit d'aspiration. Ces dimensions peuvent être quantifiées à l'aide de mesures telles que la pente spectrale et le rapport harmonique-sur/bruit (Harmoics to Noise Ratio - HNR) (Kane & Gobl, 2011).

La pente spectrale peut être mesurée de différentes manières, la plus courante étant le calcul de H1–H2 (Bickley, 1982) qui représente la différence d'amplitude entre le premier et le deuxième harmonique. De plus, des mesures alternatives de la pente spectrale comprennent la mesure de la différence d'amplitude entre les formants et le premier harmonique (H1–A1, H1–A2 et H1–A3) qui peuvent être liées à des types spécifiques de phonation (Kane & Gobl, 2011). Dans un travail récent, Chai & Garellek (2022) a proposé de remplacer le H1-H2 par un calcul corrigé de H1. Ici, en plus des différences, nous calculons donc également H1 et H1*, la correction du calcul de H1 avec la méthode décrite dans Iseli & Alwan (2004) pour éliminer l'influence des résonances du conduit vocal.

En plus des mesures d'inclinaison spectrale et de HNR, plusieurs autres indicateurs acoustiques de la qualité de la voix laryngée ont été proposés dans la littérature. Le plus prévalent est le Pic Cepstral Prominent (Cepstral Prominence Peak - CPP) (Hillenbrand *et al.*, 1994) . Les développements récents dans l'analyse acoustique de la voix ont constamment renforcé l'importance du CPP comme indicateur objectif de l'existence de souffle et de la dysphonie globale (Patel *et al.*, 2018). Des valeurs moins importantes pour la mesure du CPP suggèrent plus de souffle dans la voix.

La qualité de la voix soufflée est corrélée avec l'énergie du bruit, principalement à des fréquences élevées, et avec la hauteur relative du premier harmonique par rapport au reste (Hillenbrand *et al.*, 1994). Cela suggère que des valeurs plus faibles pour le HNR à haute fréquence et des valeurs plus élevées pour H1, H1-H2 et H1-Ax sont attendues pour des voix plus soufflées. De manière analogue, les corrélats acoustiques les plus saillants pour la voix craquée sont un H2 plus élevé et un HNR plus élevé en dessous de 500 Hz (Xu *et al.*, 2023).

Afin d'extraire automatiquement les paramètres de qualité de voix, nous avons utilisé la boîte à outils "Snack" (Sjölander, 2004) pour obtenir les valeurs de fréquence fondamentale et les fréquences des formants. A partir de ces valeurs, nous avons calculé le CPP, le HNR ainsi que les valeurs des amplitudes des formants et des harmoniques avec une implémentation python inspirée des méthodes utilisées dans "VoiceSauce" (Shue *et al.*, 2009). Ces paramètres sont ensuite moyennés sur chaque phonème issu de la transcription automatique avant analyse.

5 Résultats

Afin d'analyser les différences dans les caractéristiques de la qualité de la voix entre les styles de parole, nous avons analysé nos données à l'aide de Modèles linéaires mixtes (Linear Mixed Models - LMM) calculés grâce au package R *lme4* (R Development Core Team, 2019). Un modèle a été produit pour chaque paramètre acoustique présenté dans la section 4. Le style de parole a été inclus comme effet fixe pour tous les modèles. En ce qui concerne les effets aléatoires, des intercepts ont été inclus

pour les sujets.

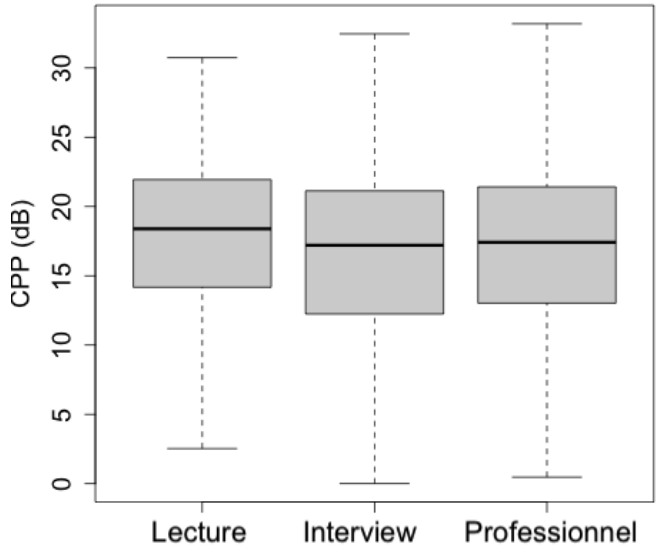


FIGURE 1 – Diagramme en boîte des valeurs normalisées de CPP pour les trois styles de parole.

Un CPP plus faible est observé (Figure 1) chez les soignants lorsqu'ils sont interviewés que lorsqu'ils s'occupent d'un patient [β = -1.26721 ; t = -2.240 ; SE = 0.56567]. Aucune différence significative n'est mesurée entre la lecture et la parole professionnelle des soignants. Ces résultats suggèrent que la parole des soignants est plus soufflée lorsqu'ils sont interviewés que lorsqu'ils accomplissent d'autres tâches.

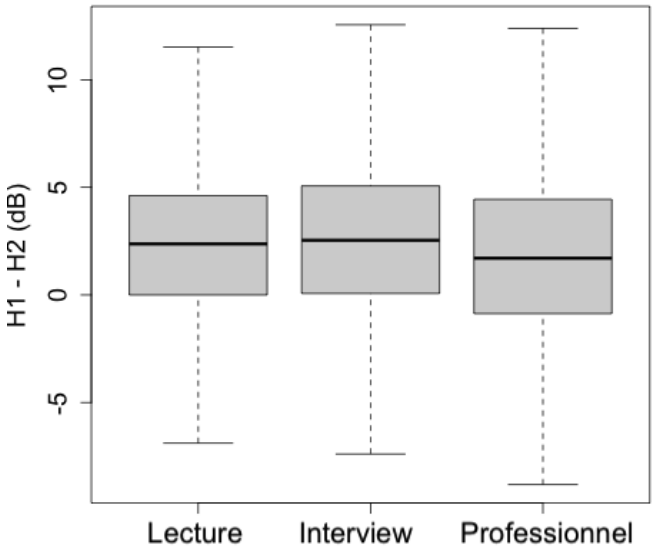


FIGURE 2 – Diagramme en boîte des valeurs normalisées de H1-H2 pour les trois styles de parole.

Le H1-H2 est plus élevé en interview [β = 0.8989 ; t = 2.841 ; SE =0.3164] que dans la parole professionnelle des soignants (Figure 2). Aucune différence significative n'est mesurée entre la lecture et la parole professionnelle des soignants. Les valeurs de H1 et H2 ont également été analysées séparément et montrent un comportement similaire (valeurs plus élevées pour l'interview et la voix professionnelle que pour la lecture).

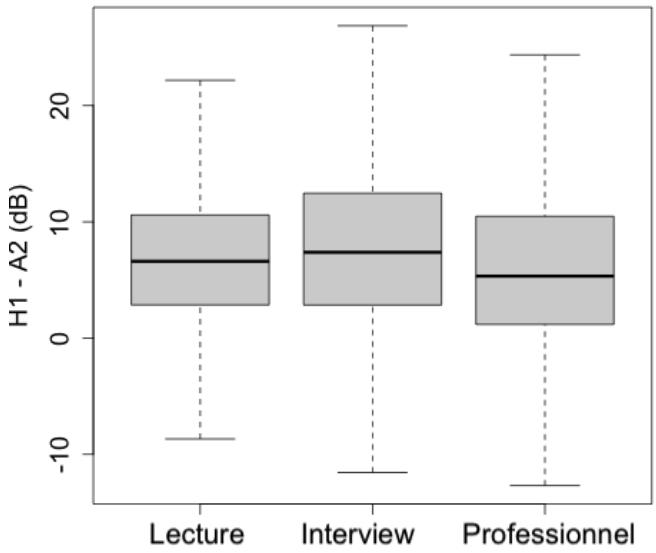


FIGURE 3 – Diagramme en boîte des valeurs normalisées de H1-A2 pour les trois styles de parole.

Les observations sur H1-A2 montrent également des valeurs plus élevées en interview [β = 2.1232 ; t = 3.066 ; SE =0.6925] que dans la parole professionnelle (Figure 3). Aucune différence significative n'est observée entre la lecture et la parole professionnelle.

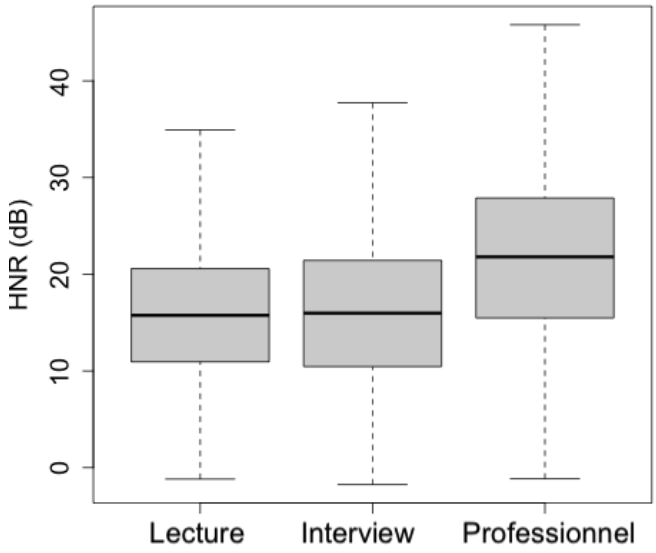


FIGURE 4 – Diagramme en boîte des valeurs normalisées de HNR pour les trois styles de parole.

Des valeurs de HNR plus faibles sont observées en interview [β = -5.5048 ; t = -5.170 ; SE = 1.0648] et en lecture [β = -5.6953 ; t = -5.337 ; SE = 1.0672], par rapport à celles observées dans la parole professionnelle (Figure 4). Ces résultats indiqueraient que les soignants utilisent moins de voix craquée dans les soins professionnels que lors de la lecture ou de l'interview.

Pour compléter ces analyses, nous fournissons les résultats LMM pour tous les paramètres dans le Tableau 2. Des différences significatives sont observées pour les paramètres H1, H2, H1-A1 et H1-A2 dans les comparaisons Interview/Professionnel et Lecture/Professionnel. Aucune différence n'est

	CPP	H1	H2	H1-H2	H1-A1	H1-A2	H1-A3	HNR
IP	*	***	*	**	NS	**	***	***
LP	NS	***	***	NS	NS	NS	**	***

TABLE 2 – Résultats LMM sur les caractéristiques de voix évaluées. « IP » fait référence à l'interview (I) comparée à la parole professionnelle (P) ; « LP » indique la lecture (L) comparée à la parole professionnelle (P). NS : Non significatif ; $* : p < .05; ** : p < .01; *** : p < .001$

trouvée pour la comparaison LP en utilisant CPP et HNR.

6 Discussion

Nous avons étudié dans cet article les styles de parole des soignants dans trois conditions différentes : la lecture de textes, les entretiens et la voix professionnelle produite lors de soins. Les résultats de l'analyse de la qualité de la voix indiquent que, lors de la prise en charge professionnelle, les soignants ont tendance à avoir des valeurs plus élevées pour le HNR et des valeurs plus faibles pour le CPP, H1, H2, H1-H2 et H1-A2.

Dans l'ensemble, ces observations ne nous fournissent malheureusement pas de conclusions claires sur la quantité de souffle utilisée dans la voix professionnelle. Les résultats sur CPP semblent illustrer une tendance vers plus de craquement dans la voix pendant les soins professionnels, tandis que les résultats du HNR sont contradictoires. L'analyse de l'amplitude des harmoniques (H1-H2, H1-A2, H1-A3) semble montrer que les soignants ont moins de souffle dans la voix lors des situations professionnelles. Toutefois, l'estimation de l'amplitude des harmoniques peut ne pas être très fiable, comme discuté dans Chai & Garellek (2022), où il est mentionné que ces mesures sont impactées par le niveau de pression sonore.

Ces résultats montrent que la proportion de souffle dans la voix, et plus largement tous les aspects de qualité de la voix, sont difficiles à mesurer et que les résultats obtenus à partir de ces mesures peuvent varier en fonction des ensembles de données et de l'objectif des études. En particulier, il nous semble évident que l'utilisation d'une seule mesure de souffle (par exemple, la CPP) n'est pas toujours suffisante.

Remerciements

Les auteurs remercient les fondateurs de l'"Humanitude" Yves Gineste et Rosette Marescotti pour leur aide dans la construction de ce projet. Un grand merci à Jean-Yves Nou, Hervé Tomassi et surtout Aurélie Rives pour nous avoir permis d'enregistrer en milieu professionnel. Nous sommes profondément redevables à tous les personnels que nous avons enregistrés pour leur confiance et leur temps.

Références

BICKLEY C. (1982). *Acoustic Analysis and Perception of Breathy Vowels*. Speech Communication Group Working Papers I, Research Laboratory of Electronics, MIT, Cambridge, MA.

BOYER F. (2021). *Reconnaissance de Parole Pour Le Français et Intégration Dans Un Système de Compréhension Du Langage Parlé*. Thèse de doctorat, Université de Bordeaux.

CHAI Y. & GARELLEK M. (2022). On H1–H2 as an acoustic measure of linguistic phonation typea). *The Journal of the Acoustical Society of America*, **152**(3), 1856–1870. DOI : 10.1121/10.0014175.

DE MAREÜIL P. B., CÉLÉRIER P. & TOEN J. (2002). Generation of Emotions by a Morphing Technique in English, French and Spanish. In *Speech Prosody*.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech, Brighton (United Kingdom*.

GINESTE Y., MARESCOTTI R. & PELLISSIER J. (2008). L'humanitude dans les soins. *Recherche en soins infirmiers*, **94**(3), 42–55. DOI : 10.3917/rsi.094.0042.

GINESTE Y. & PELLISSIER J. (2007). *Humanitude*. Nouvelle édition : Armand colin édition.

GOBL C. & NÍ CHASAIDE A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, **40**(1), 189–212. DOI : 10.1016/S0167-6393(02)00082-1.

GRICHKOVTSOVA I., MOREL M. & LACHERET A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, **54**(3), 414–429.

GUPTA V., KENNY P., OUELLET P. & STAFYLAKIS T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP*. DOI : 10.1109/ICASSP.2014.6854823.

HEBA A. (2021). *Reconnaissance Automatique de La Parole à Large Vocabulaire : Des Approches Hybrides Aux Approches End-to-End*. Theses, Université toulouse 3 Paul Sabatier.

HILLENBRAND J., CLEVELAND R. A. & ERICKSON R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, **37**(4), 769–778.

HONDA M., ITO M., ISHIKAWA S., TAKEBAYASHI Y. & TIERNEY L. (2016). Reduction of Behavioral Psychological Symptoms of Dementia by Multimodal Comprehensive Care for Vulnerable Geriatric Patients in an Acute Care Hospital : A Case Series. *Case Reports in Medicine*, **2016**, 4813196. DOI : 10.1155/2016/4813196.

HONDA M., MORI M., HAYASHI S., MORIYA K., MARESCOTTI R. & GINESTE Y. (2013). The effectiveness of French origin dementia care method ; Humanitude to acute care hospitals in Japan. *European Geriatric Medicine*, **4**, S207. DOI : 10.1016/j.eurger.2013.07.689.

ISELI M. & ALWAN A. (2004). An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. In *ICASSP 2004*, volume 1, p. I–669. DOI : 10.1109/ICASSP.2004.1326074.

ISHI C. T., ISHIGURO H. & HAGITA N. (2008). The roles of breathy/whispery voice qualities in dialogue speech. In *Speech Prosody*.

ITO M. (2004). Politeness and Voice Quality – The Alternative Method to Measure Aspiration Noise. In *Speech Prosody*, Nara, Japan.

ITO M. & HONDA M. (2015). An examination of the influence of Humanitude caregiving on the behavior of older adults with dementia in Japan. In *Proceedings of the 8th International Association of Gerontology and Geriatrics European Region Congress*, volume 2018.

KANE J. & GOBL C. (2011). Identifying regions of non-modal phonation using features of the wavelet transform. In *INTERSPEECH*, p. 177–180.

LAVER J. (1980). *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics.

PATEL R. R., AWAN S. N., BARKMEIER K. J., COUREY M., DELIYSKI D., EADIE T., PAUL D., ŠVEC J. G. & HILLMAN R. (2018). Recommended Protocols for Instrumental Assessment of Voice : American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *American Journal of Speech-Language Pathology*, **27**(3), 887–905. DOI : 10.1044/2018_AJSLP-17-0009.

PEARSELL S. & PAPE D. (2023). The effects of different voice qualities on the perceived personality of a speaker. *Frontiers in Communication*, **7**.

PITTMAN J. (1985). *Voice Quality : Its Measurement and Functional Classification - UQ eSpace*. Thèse de doctorat, The University of Queensland, School of English, Media Studies and Art History.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society.

R DEVELOPMENT CORE TEAM (2019). *R : A La nguage and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROUAS J.-L., WU Y. & SHOCHI T. (2023a). Affective attributes of French caregivers' professional speech. In *INTERSPEECH 2023*, p. 1239–1243 : ISCA. DOI : 10.21437/Interspeech.2023-848.

ROUAS J.-L., WU Y. & SHOCHI T. (2023b). A study on caregivers speech in retirement homes. In *ICPhS 2023*.

SHUE Y.-L., KEATING P. & VICENIK C. (2009). VOICESAUCE : A program for voice analysis. *The Journal of the Acoustical Society of America*, **126**(4_Supplement), 2221–2221. DOI : 10.1121/1.3248865.

SJÖLANDER K. (2004). The snack sound toolkit.

STUART-SMITH J. (1999). Glasgow : Accent and voice quality. p. 201–222. Leeds, UK : Arnold.

XU C., FOULKES P., HARRISON P., HUGHES V. & WORMALD J. H. (2023). Contributions of acoustic measures to the classification of laryngeal voice quality in continuous English speech. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)* : ASSTA.

YANUSHEVSKAYA I., GOBL C. & CHASAIDE A. N. (2006). Mapping Voice to Affect : Japanese listeners. In *Speech Prosody*, Dresden, Germany.

YANUSHEVSKAYA I., GOBL C. & NÍ CHASAIDE A. (2018). Cross-language differences in how voice quality and f contours map to affecta). *The Journal of the Acoustical Society of America*, **144**(5), 2730–2750. DOI : 10.1121/1.5066448.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude des liens acoustico-moteurs après cancer oral ou oropharyngé, via la réalisation d'un inventaire phonémique automatique des consonnes

Mathieu Balaguer[1, 2] Lucile Gelin[1, 3] Clémence Devoucoux[2] Camille Galant[4] Muriel Lalain[4] Alain Ghio[4] Jérôme Farinas[1] Julien Pinquier[1] Virginie Woisard[2, 5]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) Hôpital Larrey, Toulouse, France
(3) Lalilo, Paris, France
(4) Aix-Marseille Université, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(5) Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France
mathieu.balaguer@irit.fr

RESUME

En cancérologie ORL, le lien entre anatomie et déficit de parole est étroit en raison de l'impact de la pathologie et de son traitement sur les structures anatomiques en jeu dans la production de parole. Pourtant, les corrélations entre scores moteurs et évaluation perceptive restent faibles. L'utilisation de systèmes automatiques dédiés à la reconnaissance de phonèmes pourrait permettre d'obtenir de nouveaux résultats. L'objectif est d'étudier les liens entre scores moteurs et production phonémique via un système de reconnaissance automatique de phonèmes appliqué à une tâche de production de pseudo-mots. Après réalisation d'un inventaire phonémique par sujet, le taux d'occlusives reconnues est significativement plus faible en cas d'atteinte des structures. Certains mécanismes de compensation ont également pu être mis en évidence, notamment au niveau de la production de consonnes labiodentales, plus élevée en cas d'atteinte de la langue ou de la mâchoire.

ABSTRACT

Analysis of acoustico-motor links after oral or oropharyngeal cancer, using an automatic phonemic inventory of consonants
In ENT oncology, the link between anatomy and speech deficit is close, due to the impact of the pathology and its treatment on the anatomical structures involved in speech production. However, correlations between motor scores and perceptual evaluation remain weak. The use of automatic systems dedicated to phoneme recognition could lead to new results. The objective is to analyze the links between motor scores and phonemic production using an automatic phoneme recognition system applied to a pseudoword production task. After completing a phonemic inventory for each subject, the rate of occlusives recognized was significantly lower in cases of structural impairment. Compensatory mechanisms were also demonstrated, notably in the production of labiodental consonants, which was higher in cases of tongue or jaw damage.

MOTS-CLES : trouble de la parole, cancérologie, évaluation automatique, phonèmes, capacités motrices
KEYWORDS : speech disorder, oncology, automatic assessment, phonemes, motor skills

Actes des 35èmes Journées d'Études sur la Parole, pages 61–70.

1 Contexte

Les cancers de la cavité buccale ou de l'oropharynx sont des cancers fréquents (Lapôtre-ledoux et al., 2023). La question de la vie après cancer est essentielle, et il convient de s'intéresser aux fonctions des voies aérodigestives supérieures (VADS) particulièrement dégradées en raison de la localisation de ces cancers. La fonction de parole devient ainsi un sujet crucial en termes de réhabilitation, à cause de l'impact fonctionnel et psychosocial qu'un trouble de parole va avoir sur les personnes (Mlynarek et al., 2008; Reich, 2009).

Les évaluations cliniques de la parole sont majoritairement menées par les orthophonistes (Pommée et al., 2022). En cancérologie des VADS, peu d'outils d'évaluation existent à l'heure actuelle (Ghio et al., 2016), alors qu'il s'agit pourtant du symptôme le plus fréquent (Plisson et al., 2017). L'évaluation clinique de la parole comprend habituellement deux volets. La partie analytique consiste en un bilan des structures anatomiques en vue de mettre en évidence un déficit d'amplitude, de tonus, de force motrice ou de sensibilité. La partie fonctionnelle concerne l'évaluation perceptive pour de caractériser, sur des tâches de production de parole, les répercussions des déficits moteurs ou sensitifs sur les productions orales (Middag, 2013). Ici, l'intelligibilité est particulièrement ciblée dans les évaluations. Elle est définie comme la capacité à « *reconstruire un énoncé au niveau acoustico-phonétique* » (Pommée et al., 2021), en d'autres termes à décoder un signal de parole en éléments phonémiques sans mise en œuvre des mécanismes cognitifs de restauration du message par l'auditeur (Ghio et al., 2018).

Ainsi, même si en cancérologie le lien entre anatomie et déficit fonctionnel sur la parole est étroit, la corrélation reste faible entre les scores fonctionnels d'intelligibilité attribués perceptivement et les scores analytiques issus du bilan moteur (Lazarus et al., 2013). Les avancées récentes dans le champ de l'analyse automatique de la parole peuvent désormais permettre d'envisager une application clinique de ces techniques, en palliant certaines caractéristiques inhérentes à l'évaluation perceptive (notamment la variabilité inter et intra-juges (Fex, 1992; Middag et al., 2008)). Des études ont par exemple déjà montré l'intérêt d'une mesure de performance de systèmes de reconnaissance automatique de parole, comme mesure d'intelligibilité (Christensen et al., 2012; Doyle et al., 1997; Maier et al., 2010). L'utilisation de systèmes de reconnaissance automatique de phonèmes pour dresser un inventaire des phonèmes reconnus peut permettre ainsi d'ouvrir de nouvelles perspectives dans l'étude du lien entre parole et scores moteurs.

L'objectif est d'étudier les liens entre scores moteurs (cavité buccale, oropharynx) et production phonémique (via la constitution d'un inventaire phonémique des consonnes par utilisation de système de reconnaissance automatique de phonèmes) après traitement d'un cancer de la cavité buccale ou de l'oropharynx.

2 Matériel et méthodes

2.1 Schéma d'étude

Cette étude prospective observationnelle s'inscrit dans le cadre du projet PHRIP DAPADAF-E (PHRIP-19-0004). Toutes les procédures ont été effectuées conformément à la déclaration d'Helsinki de 1964 et à ses amendements ou à des normes éthiques comparables. Chaque sujet a été informé à l'avance de l'objectif de l'étude et a reçu une fiche d'information. Les sujets ont confirmé leur non-opposition à la collecte de données et l'utilisation de la recherche dans le projet DAPADAF-E. Les fichiers audio enregistrés ont été récupérés via le GIS Parolothèque, de même que les données individuelles, cliniques et de traitement.

2.2 Population

Les patients venant en consultation ORL ou dans un service de soins de suite et de réadaptation ORL entre novembre 2021 et décembre 2023, sur les sites d'inclusion des Hôpitaux de Toulouse ou de l'Assistance Publique – Hôpitaux de Marseille, ont été invités à participer à cette étude.

Les critères d'inclusion étaient : être majeurs, francophones natifs, avoir été traités pour un cancer de la cavité buccale ou de l'oropharynx (chirurgie et/ou radiothérapie et/ou chimiothérapie) et être en rémission clinique depuis au moins 6 mois (caractère chronique et stable des troubles). N'ont pas été inclus les patients dont la fatigabilité ne permet pas la passation des épreuves, ou présentant une pathologie associée potentiellement responsable de trouble de parole (bégaiement, trouble neurologique…).

2.3 Corpus

2.3.1 Scores analytiques moteurs

En raison de l'absence de test exhaustif et validé en cancérologie ORL, le protocole DAPADAF-E s'appuie sur le bilan clinique issu de la BECD (Batterie d'Évaluation Clinique de la Dysarthrie, (Auzou & Rolland-Monnoury, 2019)), et particulièrement sur l'épreuve d'examen moteur.

Ainsi, pour chaque structure anatomique d'intérêt (lèvres, joues, mâchoire, langue et vélopharynx), deux scores sont obtenus : un score de synthèse analytique (correspondant aux épreuves du domaine non-verbal : tâches motrices hors production de parole) et un score de synthèse fonctionnel (domaine verbal : tâches associées à une production de parole). Chacun de ces scores est coté sur le principe d'une échelle de Likert à 5 niveaux : 0 correspondant à l'absence d'anomalie, 1 à une anomalie discrète ou rare, 2 à une anomalie modérée ou occasionnelle, 3 à une anomalie marquée ou fréquente et 4 à une anomalie sévère ou quasi permanente. Pour les besoins de cette étude et pour obtenir un nombre plus important de sujets dans chaque classe, les scores moteurs (analytiques et fonctionnels) ont été dichotomisés, de façon à obtenir deux valeurs pour chacun : « *pas d'anomalie, anomalie discrète ou modérée* » (0 : scores entre 0 et 2) et « *anomalie marquée, sévère ou quasi permanente* » (1 : scores 3 ou 4).

2.3.2 Enregistrements de parole

Les sujets ont été enregistrés sur une tâche de production de pseudo-mots respectant les règles phonotactiques du français (par exemple : *crerquin*, *ruflu*…). Ils étaient installés devant un écran affichant le pseudo-mot à produire et portaient un casque dans lequel une voix de synthèse prononçait le pseudo-mot, de façon à ce qu'ils aient accès à la double modalité lecture et répétition pour produire le pseudo-mot cible. Chaque sujet a été enregistré sur deux listes de 52 pseudo-mots, différentes entre elles et pour chaque patient. Ces listes ont été mises au point par l'équipe du Laboratoire Parole et Langage d'Aix-en-Provence (Ghio et al., 2022, 2018).

Les enregistrements ont été réalisés au moyen d'un micro Neumann TLM 102, protégé par une bonnette et un filtre anti pop, et connecté à une interface audio RME Fireface UC. L'interface était ensuite reliée à un ordinateur portable sur lequel le logiciel LiveIntel (également mis au point par le LPL) était installé et qui gérait l'enregistrement de la tâche. Les sujets étaient enregistrés dans une

salle de consultation calme. Les enregistrements étaient au format 48 kHz 16 bits, puis ils ont été rééchantillonnés en 16 kHz car il s'agit de l'usage dans le traitement automatique.

2.4 Analyse automatique de la parole

Chaque fichier, comprenant la production de 52 pseudo-mots, a été ensuite segmenté par le détecteur d'activité vocale WebRTC-VAD[1], afin d'obtenir des segments de parole courts (de durée inférieure à 25 secondes) pouvant être gérés par le système de reconnaissance de phonèmes.

Les fichiers segmentés ont ensuite été donnés à un système de reconnaissance automatique de phonèmes de type Transformer CTC, développé par les auteurs de cette étude et entraîné sur le corpus francophone CommonVoice, ici constitué de 148,9 heures d'enregistrements de parole lue par 1 276 locuteurs (avec 420 secondes d'enregistrements par locuteur en moyenne). Le décodage a abouti à la reconnaissance de la séquence des phonèmes produits, parmi 33 possibilités pour chaque phonème sur lequel le Transformer CTC a été entraîné :

- 18 consonnes : /p/, /t/, /k/, /b/, /d/, /g/, /f/, /s/, /ʃ/, /v/, /z/, /ʒ/, /m/, /n/, /ɲ/, /ŋ/, /l/, /ʁ/) ;
- 12 voyelles, dont /ɑ̃/, /ɛ/, /e/, /o/, /ɔ/, /i/, /ɔ̃/, /y/, /u/, et les archiphonèmes /A/ (/a/, /ɑ/), /Œ/ (/ə/, /œ/, /ø/) et /Ẽ/ (/ɛ̃/, /œ̃/) ;
- et trois semi-consonnes : /j/, /w/, /ɥ/.

Puis, sur l'ensemble des deux listes produites par sujet (deux fois 52 soit 104 pseudo-mots par sujet), un inventaire phonémique ciblé sur les consonnes a été dressé afin d'obtenir le nombre total d'occurrences pour chaque phonème.

Enfin, dix indicateurs ont été calculés suite à cet inventaire phonémique, concernant les taux de :

- Consonnes reconnues (*csn*) : nombre total de consonnes / nombre total de phonèmes ;
- Occlusives reconnues (*occ*) : nombre total d'occlusives / nombre total de consonnes ;
- Bilabiales (*bilabial*) : nombre total de consonnes bilabiales / nombre total de consonnes ;
- Labiodentales (*labiodent*) : nombre total de consonnes labiodentales / nombre total de consonnes ;
- Alvéolaires (*bilabial*) : nombre total de consonnes alvéolaires / nombre total de consonnes ;
- Post-alvéolaires (*postalv*) : nombre total de consonnes post-alvéolaires / nombre total de consonnes ;
- Palatales (*palatal*) : nombre total de consonnes palatales / nombre total de consonnes ;
- Vélaires (*velaire*) : nombre total de consonnes vélaires / nombre total de consonnes ;
- Sourdes (*sourde*) : nombre total de consonnes sourdes / nombre total de consonnes ;
- Consonnes orales (*corale*) : nombre total de consonnes orales / nombre total de consonnes.

2.5 Analyses statistiques

Les analyses statistiques ont été réalisées au moyen du logiciel Stata 16.1 (StataCorp. 2019. Stata statistical software: release 16. College Station, TX: StataCorp LLC). Un seuil de significativité à 5 % a été choisi pour l'ensemble des analyses. En raison du caractère non-paramétrique des données, le test de Mann-Whitney a été réalisé pour tester l'absence de différence significative de proportion des différentes catégories phonémiques selon le niveau d'anomalie motrice.

[1] Site : https://github.com/wiseman/py-webrtcvad, consulté le 02/02/2024

3 Résultats

3.1 Description de la population

Soixante-dix-sept sujets ont été inclus dans cette étude, majoritairement de sexe masculin (50/77, 65 %). La proportion de sujets avec tumeurs de petit et de grand volume est équilibrée (T1+T2 = 32/64, 50 % ; T3+T4 = 32/64, 50 % ; 13 données manquantes). La majorité des localisations tumorales se situent dans la cavité buccale (langue : 38, plancher buccal : 18, bouche : 14, gencives : 5 et glandes salivaires : 1) avec certains patients présentant plusieurs sites tumoraux. 30 sujets ont présenté un cancer de l'oropharynx. 63 sujets ont été traités par chirurgie (82 %), 66 (86 %) par radiothérapie et 41 (53 %) par chimiothérapie.

3.2 Reconnaissance automatique phonémique

Dans un premier temps, le nombre de phonèmes différents reconnus, sans tenir compte de leur nombre d'occurrences, a été étudié. En moyenne, 28,9 phonèmes différents sont reconnus sur les 32 phonèmes cibles (écart-type = 4,6, voir Figure 1). Un sujet n'a que trois phonèmes différents reconnus au cours de la production des 104 pseudo-mots attendus.

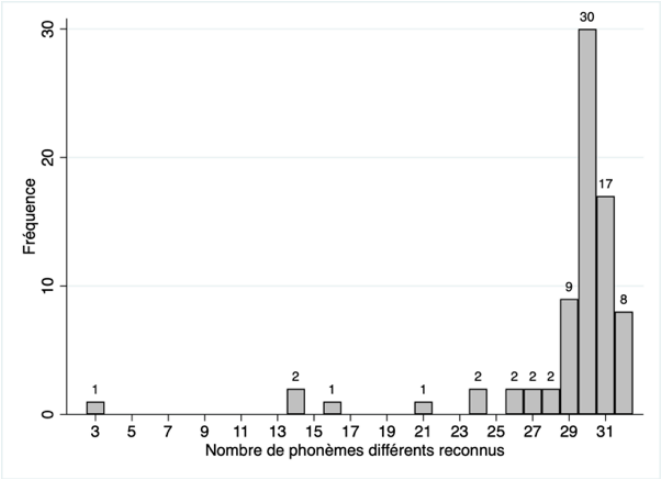


FIGURE 1 : Distribution du nombre de phonèmes différents reconnus pour chaque sujet par le système de reconnaissance automatique

Puis, nous avons dressé un inventaire phonémique global par sujet, en comptant toutes les occurrences de tous les phonèmes (cf. Table 1). Nous retrouvons alors une légère prépondérance des consonnes reconnues dans la parole des sujets par rapport aux voyelles, avec un taux moyen de consonnes à 50,9 % (la proportion de consonnes en français en population générale étant de 56,6 %, Wioland, 1991).

3.3 Liens articulatoires et moteurs

La table 2 représente les taux des différentes classes phonémiques significativement différents entre les sujets ne présentant pas d'anomalie (anomalie absente ou faible) et ceux présentant une anomalie lors de l'examen moteur (cf. section 2.3.1). Les classes « taux de consonnes bilabiales », « taux de consonnes alvéolaires », « taux de consonnes palatales » et « taux de consonnes orales » ne sont pas représentées, car elles ne montrent aucune différence significative, quelle que soit la localisation anatomique testée.

Consonnes	Moyenne (ET)	Minimum ; Maximum
csn	50,87 % (3,19)	33,33 % ; 60,98 %
occ	19,34 % (8,43)	0,00 % ; 40,00 %
bilabial	12,74 % (4,00)	0,00 % ; 24,00 %
labiodent	14,94 % (5,52)	0,00 % ; 32,47 %
alveol	38,50 % (9,44)	18,18 % ; 100,00 %
postalv	5,66 % (3,11)	0,00 % ; 13,53 %
palatal	21,75 % (5,67)	0,00 % ; 47,37 %
velaire	6,41 % (3,09)	0,00 % ; 13,01 %
sourde	24,60 % (8,07)	0,00 % ; 38,32 %
corale	86,68 % (5,87)	69,94 % ; 100,00 %

TABLE 1 : Tableau de distribution des taux de consonnes reconnues

		csn	occ	labiodent	postalv	velaire	sourde
Lèvres	Score A (p)	*N.S.*	*0,03*	*0,01*	*0,02*	*N.S.*	*0,049*
	0		20,17%	**14,29%**	6,03%		25,36%
	1		**13,81%**	19,26%	**3,24%**		**19,53%**
	Score F (p)	*N.S.*	*0,003*	*0,01*	*0,02*	*N.S.*	*N.S.*
	0		20,15%	**14,30%**	5,99%		
	1		**13,25%**	19,78%	**3,24%**		
Joues	Score A (p)	*N.S.*	*0,02*	*N.S.*	*N.S.*	*N.S.*	*0,001*
	0		20,31%				25,94%
	1		**14,58%**				**17,99%**
	Score F (p)	*0,02*	*0,006*	*N.S.*	*N.S.*	*N.S.*	*0,0002*
	0	51,56%	21,74%				27,44%
	1	**49,85%**	**15,78%**				**20,40%**
Mâchoire	Score A (p)	*N.S.*	*N.S.*	*N.S.*	*0,03*	*N.S.*	*N.S.*
	0				6,02%		
	1				**3,73%**		
	Score F (p)	*N.S.*	*N.S.*	*0,01*	*0,02*	*N.S.*	*N.S.*
	0			**14,28%**	6,03%		
	1			18,91%	**3,45%**		
Langue	Score A (p)	*0,004*	*0,0001*	*0,0002*	*N.S.*	*0,02*	*0,0001*
	0	51,30%	24,29%	**12,20%**		7,46%	28,98%
	1	**50,60%**	**16,19%**	16,68%		**5,73%**	**21,80%**
	Score F (p)	*0,001*	*0,0002*	*0,005*	*N.S.*	*0,04*	*0,001*
	0	51,41%	23,15%	**12,94%**		7,18%	27,94%
	1	**50,46%**	**16,48%**	16,44%		**5,83%**	**22,10%**
Vélopharynx	Score A (p)	*0,0475*	*0,01*	*N.S.*	*N.S.*	*N.S.*	*0,03*
	0	51,31%	21,06%				26,07%
	1	**50,01%**	**15,97%**				**21,72%**
	Score F (p)	*N.S.*	*N.S.*	*N.S.*	*N.S.*	*N.S.*	*0,01*
	0						26,46%
	1						**20,95%**

* 0 = pas d'anomalie, anomalie discrète ou modérée ; 1 = anomalie marquée, sévère ou quasi permanente ; N.S. : non significatif

* Score A = score de synthèse analytique ; Score F = score de synthèse fonctionnel. Le p correspond à la p-value du test de Mann-Whitney (entre anomalie absente ou faible et marquée).

Seules les proportions significativement différentes selon l'anomalie sont indiquées, avec le taux le plus faible indiqué en gras.

TABLE 2 : Moyenne des taux de consonnes reconnues par sujet, selon les scores moteurs analytiques et fonctionnels et par segment anatomique testé

Le taux de consonnes est ainsi significativement plus faible en cas d'atteinte de la langue, de même que le taux d'occlusives. Ce taux d'occlusives est également significativement plus faible en cas d'atteinte analytique ou fonctionnelle des joues et des lèvres. Le taux de consonnes sourdes est également moindre en cas d'atteinte analytique des lèvres, et d'atteinte analytique et fonctionnelle des joues, de la langue et du vélopharynx.

Dans le détail des catégories, aucune différence significative de proportion de taux de consonnes post-alvéolaires n'est retrouvée au niveau de la langue (en fonctionnel : anomalie absente ou faible à 5,56 % ; anomalie marquée à 5,74 %, p=0,95), alors que cette différence est retrouvée au niveau de la mâchoire sur la même classe phonémique (en fonctionnel : anomalie absente ou peu marquée à 6,03 % ; anomalie marquée à 3,45 %, p=0,02). Le taux de consonnes vélaires n'est significativement différent qu'en cas d'atteinte de la langue, avec un taux significativement plus important en cas d'absence d'anomalie. Enfin, en cas d'atteinte de la mâchoire, des lèvres et de la langue, le taux de consonnes labiodentales reconnues devient plus élevé.

4 Discussion

4.1 Résultats principaux

Cette étude montre l'effet que peut avoir une atteinte motrice, analytique ou fonctionnelle, sur la qualité de la production phonémique de la parole.

Les anomalies linguales décelées lors du bilan analytique sont ainsi liées à un taux plus faible de consonnes reconnues par un système de reconnaissance phonémique, d'occlusives, de postalvéolaires et de consonnes sourdes. Ces éléments sont cohérents avec l'altération des structures anatomiques en jeu dans la production de ce type de phonèmes. Seul le taux de consonnes post-alvéolaires reconnues n'est pas significativement associé à une altération de la langue, probablement lié au fait que les consonnes post-alvéolaires sont majoritairement des fricatives (/ʃ/, /ʒ/) et que le déficit est davantage lié à une atteinte de la mâchoire que l'on retrouve dans nos résultats. Enfin, une atteinte de la langue est également significativement associée à un taux plus important de consonnes labiodentales reconnues, pouvant être lié à un mécanisme de compensation où l'occlusion est délocalisée du niveau lingual (occlusion dégradée ou impossible) au niveau labio-dental.

Contrairement à De Bruijn et al., 2012, nous ne retrouvons pas dans notre corpus de différence significative de taux de consonnes alvéolaires selon l'altération motrice, et ce quel que soit le segment anatomique. Ceci pourrait être dû à la composition de notre échantillon qui présente très majoritairement une atteinte orale (76 sujets, 30 sujets ayant une atteinte de l'oropharynx avec des sujets présentant une atteinte mixte orale et oropharyngée), ce qui ne permet pas de mettre en évidence ce type de dégradation.

De façon plus globale, nous retrouvons un taux d'occlusives significativement plus faible en cas d'anomalie analytique ou fonctionnelle au niveau des lèvres, des joues, de la langue et du vélopharynx (analytique seulement). En raison de la localisation de la pathologie tumorale et du traitement subi, majoritairement chirurgical (63/77, 82 %), le manque de mobilité des segments anatomiques combiné au défect structurel limite la capacité d'occlusion des sujets qui ont alors tendance à produire des approximantes, reconnues comme fricatives par notre système de reconnaissance phonémique (taux d'occlusives reconnues plus faible). Le rôle du défect est également retrouvé au niveau des anomalies analytiques et fonctionnelles labiales dans une population pour laquelle aucun sujet n'a présenté de localisation tumorale au niveau des lèvres. Ainsi,

les différences de taux d'occlusives, de labiodentales, postalvéolaires et sourdes retrouvées peuvent être liées à l'atteinte labiale secondaire relative aux conséquences du traitement chirurgical et radiothérapique, pouvant réduire la capacité de contention labiale ou justifier la mise en place d'une compensation plus locale labiodentale.

4.2 Limites de l'étude

Notre étude s'intéresse aux liens acoustico-moteurs après cancer oral ou oropharyngé, via la réalisation d'un inventaire phonémique automatique des consonnes. Toutefois, les recrutements étant encore en cours à l'heure de publication de ce papier, les résultats présentés ici ne peuvent être considérés que comme des tendances et non des résultats définitifs. Un recrutement d'un plus grand nombre de sujets permettra également de s'intéresser de façon plus précise aux liens pouvant exister entre la localisation de la pathologie, ses impacts analytiques et fonctionnels et ses répercussions sur la production de parole, notamment au niveau phonémique. Il est en effet possible que les résultats obtenus montrent des tendances différentes selon la région anatomique en jeu dans le cancer ou son traitement, notamment entre la zone orale et la zone oropharyngée.

De plus, notre étude s'est intéressée à l'impact du cancer oral ou oropharyngé sur les consonnes. L'observation du comportement d'un système de reconnaissance phonémique au niveau des voyelles pourrait ainsi permettre de compléter l'analyse, notamment via les effets de modification des volumes intra-oraux induits principalement par le traitement chirurgical.

4.3 Perspectives

L'étude des liens acoustico-moteurs est importante en cancérologie ORL. Un lien entre la production de parole et le score moteur permettrait de simplifier et d'optimiser les évaluations cliniques chez des patients souvent fatigables, en ne proposant aux sujets que de réaliser une tâche de répétition de pseudo-mots. C'est par une analyse automatique des enregistrements de ces productions que seraient déterminées à la fois les unités linguistiques déficitaires, mais également les altérations dynamiques habituellement retrouvées lors d'un bilan analytique orthophonique complet.

Ce type d'analyse permettrait également de fournir aux cliniciens des éléments de compréhension plus fins des mécanismes en jeu dans les liens articulatoires et acoustiques de bas niveau de la parole. Les stratégies thérapeutiques de patients seront améliorées, grâce à une combinaison d'approches ciblant l'intelligibilité de la parole et d'autres plus écologiques ciblant la communication et la compréhensibilité.

5 Conclusion

L'utilisation de systèmes de reconnaissance automatique de phonèmes aboutissant à la réalisation d'un inventaire phonémique par sujet permet de mettre en évidence une influence de la pathologie cancérologique orale ou oropharyngée et de ses traitements sur la reconnaissance de phonèmes. Notamment, l'altération motrice de la langue a un impact large sur la reconnaissance des consonnes (occlusives et vélaires notamment), et certains mécanismes de compensation ont pu être retrouvés (avec des taux de reconnaissance de consonnes labiodentales plus élevés en cas d'atteinte des lèvres, de la mâchoire ou de la langue). Cette étude doit donc se poursuivre par une analyse de la reconnaissance automatique des voyelles, et la mise en perspective de ces résultats avec les données cliniques et de traitement des sujets.

Remerciements

Cette étude a bénéficié d'un financement du ministère de la Santé (PHRIP, 2019, PHRIP-19-0004).

Références

AUZOU, P., & ROLLAND-MONNOURY, V. (2019). BECD : Batterie d'Évaluation Clinique de la Dysarthrie (*Ortho Édit*).

CHRISTENSEN, H., CUNNINGHAM, S., FOX, C., GREEN, P., & HAIN, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Interspeech 2012*, 2, 1776–1779. https://doi.org/10.21437/Interspeech.2012-484

DE BRUIJN, M. J., BOSCH, L. TEN, KUIK, D. J., WITTE, B. I., LANGENDIJK, J. A., RENE LEEMANS, C., & VERDONCK-DE LEEUW, I. M. (2012). Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. *Speech Communication*, 54(5), 632–640. https://doi.org/10.1016/j.specom.2011.06.005

DOYLE, P. C., LEEPER, H. A., KOTLER, A. L., THOMAS-STONELL, N., O'NEILL, C., DYLKE, M. C., & ROLLS, K. (1997). Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development*, 34(3), 309–316. http://www.ncbi.nlm.nih.gov/pubmed/9239624

FEX, S. (1992). Perceptual evaluation. *Journal of Voice*, 6(2), 155–158.

GHIO, A., GIUSTI, L., BLANC, E., PINTO, S., LALAIN, M., ROBERT, D., FREDOUILLE, C., & WOISARD, V. (2016). Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ? *Journées d'Étude Sur La Parole*, 589–596. https://hal.archives-ouvertes.fr/hal-01372037

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. *XXXIIe Journées d'Études Sur La Parole*, 285–293. https://doi.org/10.21437/JEP.2018-33

GHIO, A., LALAIN, M., REBOURG, M., MARCZYK, A., FREDOUILLE, C., & WOISARD, V. (2022). Validation of an Intelligibility Test Based on Acoustic-Phonetic Decoding of Pseudo-Words: Overall Results from Patients with Cancer of the Oral Cavity and the Oropharynx. *Folia Phoniatrica et Logopaedica*, 74(3), 209–222. https://doi.org/10.1159/000519427

LAPOTRE-LEDOUX, B., REMONTET, L., UHRY, Z., DANTONY, E., GROSCLAUDE, P., MOLINIE, F., WORONOFF, A.-S., LECOFFRE-BERNARD, C., LAFAY, L., DEFOSSEZ, G., D'ALMEIDA, T., & FRANCIM, R. français des registres de cancers. (2023). Incidence Des Principaux Cancers En France Métropolitaine En 2023 Et Tendances Depuis 1990 / Main Cancers Incidence in Metropolitan France in 2023 and Trends Since 1990. *Bulletin Épidémiologique Hedomadaire*, 12–13, 188–204. http://beh.santepubliquefrance.fr/beh/2023/12-13/2023_12-13_1.html

LAZARUS, C. L., HUSAINI, H., ANAND, S. M., JACOBSON, A. S., MOJICA, J. K., BUCHBINDER, D., & URKEN, M. L. (2013). Tongue Strength as a Predictor of Functional Outcomes and Quality of Life after Tongue Cancer Surgery. *Annals of Otology, Rhinology & Laryngology*, 122(6), 386–397. https://doi.org/10.1177/000348941312200608

MAIER, A., HADERLEIN, T., STELZLE, F., NÖTH, E., NKENKE, E., ROSANOWSKI, F., SCHÜTZENBERGER, A., & SCHUSTER, M. (2010). Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(April 2014), 1–7. https://doi.org/10.1155/2010/926951

MIDDAG, C. (2013). Automatische analyse van pathologische spraak Automatic Analysis of Pathological Speech.

MIDDAG, C., VAN NUFFELEN, G., MARTENS, J. P., & DE BODT, M. (2008). Objective intelligibility assessment of pathological speakers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1745–1748.

MLYNAREK, A., RIEGER, J., HARRIS, J., O'CONNELL, D., AL-QAHTANI, K., ANSARI, K., CHAU, J., & SEIKALY, H. (2008). Methods of functional outcomes assessment following treatment of oral and oropharyngeal cancer: review of the literature. *Journal of Otolaryngology - Head & Neck Surgery*, 37(1), 2–10. https://doi.org/10.2310/7070.2008.1001

PLISSON, L., PILLOT-LOISEAU, C., & CREVIER-BUCHMAN, L. (2017). Intelligibilité de la parole après le traitement d'un cancer de l'oropharynx : étude descriptive chez sept patients en pré-traitement et en post-traitement précoce. *7èmes Journées de Phonétique Clinique* (JPC7).

POMMEE, T., BALAGUER, M., MAUCLAIR, J., PINQUIER, J., & WOISARD, V. (2021). Intelligibility and comprehensibility: A Delphi consensus study. *International Journal of Language & Communication Disorders*, 1–44. https://doi.org/10.1111/1460-6984.12672

POMMEE, T., BALAGUER, M., MAUCLAIR, J., PINQUIER, J., & WOISARD, V. (2022). Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, 47(2), 92–108. https://doi.org/10.1080/14015439.2020.1870245

REICH, M. (2009). Cancer et image du corps : identité, représentation et symbolique : Le corps retrouvé (French). *Cancer and Body Image : Identity, Representation* (English), 85(3), 247–254. https://doi.org/10.3917/inpsy.8503.0247

WIOLAND, F. (1991). *Prononcer les mots du français – Des sons et des rythmes*. Hachette. ISBN : 2-01-017482-8

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude en *temps réel* de la fusion des /a/ ~ /ɑ/ en français depuis 1925

Juliusz Cecelewski[1], Cédric Gendrot[1], Martine Adda-Decker[1], Philippe Boula de Mareüil[2]
(1) Laboratoire de Phonétique et Phonologie (CNRS, U. Sorbonne-Nouvelle), 4 rue des Irlandais, 75005 Paris, France
(2) Laboratoire Interdisciplinaire des Sciences du Numérique (CNRS, Paris-Saclay), rue Raimond Castaing bâtiment 650, 91190 Gif-sur-Yvette, France
juliusz.cecelewski@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr, martine.adda-decker@sorbonne-nouvelle.fr, philippe.boula.de.mareuil@limsi.fr

RESUME

Cette étude explore la variation diachronique de la réalisation des voyelles /a/ ~ /ɑ/ du français en position finale de mot dans la parole déclamatoire/journalistique de 1925 à 2023. Nos données comprennent deux corpus préexistants – le corpus d'archives INA (1940–1997) et le corpus ESTER (2000–2004) – ainsi que deux nouveaux corpus composés d'enregistrements issus des Archives de la Parole d'Hubert Pernot (1925–1929), de Radio France et de YouTube (2020–2023).

Nos résultats indiquent une postériorisation du /a/ vers une position plus centrale et, dans une moindre mesure, une antériorisation du /ɑ/, qui ont abouti à la neutralisation et la fusion acoustique des deux phonèmes au cours du XX[e] siècle. Les résultats sont discutés à la lumière de l'évolution globale du système des voyelles à double timbre en français.

ABSTRACT

Real-time study of the fusion of vowels /a/~/ɑ/ in French since 1925. This study explores the diachronic variation in the realization of French vowels /a/ ~ /ɑ/ in final word positions in declamatory/journalistic speech from 1925 to 2023. Our corpora include two pre-existing corpora - the INA archive corpus (1940–1997) and the ESTER corpus (2000–2004) - as well as two new corpora consisting of recordings from Hubert Pernot's Archives de la Parole (1925–1929), from Radio France and YouTube (2020–2023).

Our findings indicate a backing of /a/ towards a more central position and, to a lesser extent, a fronting of /ɑ/, leading to the neutralization and acoustic merger of these two phonemes over the course of the 20th century. The results are discussed in light of the overall evolution of the system of two-quality vowels in French.

MOTS-CLES : phonétique diachronique, phonologie du français, /ɑ/ postérieur

KEYWORDS : diachronic phonetics, French phonology, back /ɑ/

1 Introduction

Les traités de prononciation, de même que de nombreux ouvrages didactiques font état, traditionnellement, de deux voyelles A à valeur distinctive en français : un /a/ *antérieur* et un /ɑ/ *postérieur* ou *grave*, ce dernier étant « produit par un résonateur dont le volume est plus grand, ou l'orifice plus petit » (Rousselot & Laclotte, 1902). À en croire les témoignages impressionnistes des grammairiens d'époques antérieures, un timbre spécifique de /ɑ/ n'apparaît pas avant le XVII[e] siècle. Pour ces auteurs, la prononciation du A français n'était « point beaucoup differente de celle des Latins » (Estienne, 1557). C'est au XVII[e] siècle que remontent les premiers témoignages admettant que l'on « prononce differemment *male* une eſpece de coffre, & *mâle maſculus* » (Lamy, 1688). Même si, encore au XVIII[e] siècle, il est des grammairiens qui soutiennent, à propos de l'accent

circonflexe, qu'il « ne change point le ſon de l'a, il ne ſert qu'à le rendre long » (De La Touche, 1730) ; c'est dès cette époque qu'une différence de timbre aura été définitivement reconnue et caractérisée en termes articulatoires (Boindin, 1753).
Les premiers traités phonétiques du XX^e^ siècle restreindront la présence du /ɑ/ en français standard ou parisien à la syllabe finale de mot : *pâte* [pɑt], la syllabe pénultième ne pouvant accueillir qu'un timbre intermédiaire, dit « a moyen » : *pâté* [pɑ̝te]. À tous les /ɑ/ étymologiques au-delà de l'avant-dernière syllabe sera alors assigné le timbre antérieur : *pâtisserie* [patisχi] (Rousselot & Laclotte, 1902 ; Malmberg, 1969).

Dès les années 1930, des voix se feront pourtant entendre affirmant que les deux A étaient en passe de s'acheminer vers un seul timbre (Pernot, 1928 ; Fouché, 1935). Pour Delattre, ce rapprochement des timbres aurait été « l'effet d'une réaction contre la divergence profonde qui existe dans l'accent faubourien, où [a] est presque [æ] et [ɑ] presque [ɔ] » (Delattre, 1957). Quoi qu'il en soit, un rendement fonctionnel faible de cette opposition aura sans doute contribué à entamer la neutralisation (Léon, 1992).
Un changement en cours semble être corroboré par différents témoignages concernant un effet d'âge du locuteur sur la prononciation (Fouché, 1935). Dans la même lignée, les résultats d'une enquête de Walter (1977) ont confirmé que la répartition des deux phonèmes dans les unités lexicales, sensiblement différente d'un sujet à l'autre, n'était stable que chez les locuteurs les plus âgés. Se faisant plus rare dans un usage courant, le /ɑ/ est peu à peu devenu, dans la seconde moitié du siècle, un trait de distinction sociale, caractéristique d'une prononciation « mondaine » et « affectée » (Mettas, 1970), montrant la fonction sociale ambivalente de ce phonème.
Alors que la neutralisation de l'opposition /a, ɑ/ semble aujourd'hui incontestable, une documentation empirique manque pour caractériser acoustiquement le processus de fusion et son produit — le nouveau phonème unique /a/ — par rapport aux deux /a, ɑ/ qu'il a supplantés. S'inscrivant dans le cadre plus large du système de voyelles à double timbre — /e, ɛ/, /o, ɔ/ et /ø, œ/ — qui ont toutes connues des évolutions plus ou moins importantes au cours des XIX^e^ et XX^e^ siècles (Baraduc *et al.*, 1989 ; Hansen & Juillard, 2011), la question qui doit également être soulevée est celle du positionnement de la fusion des deux A dans l'évolution globale du système vocalique du français.
À la lumière des descriptions antérieures, nous formulons l'hypothèse d'un rapprochement acoustique et, en conséquence, une fusion aboutissant à la perte de distinctivité entre /a/ et /ɑ/ au cours du XX^e^ siècle. Plus précisément, nous nous attendons à observer un rehaussement progressif des valeurs de F2 des /ɑ/, jusqu'à atteindre celles de F2 de /a/, parallèlement à une éventuelle centralisation de ce dernier dans la période étudiée. Pour tester cette hypothèse, nous examinerons ici les caractéristiques acoustiques des /a, ɑ/ en syllabe finale de mot dans un corpus d'archives de parole couvrant la période de 1925 à 2023.

2 Corpus et méthode

2.1. Corpus

Corpus BnF 1925–1929

Le plus ancien des corpus utilisés est composé d'enregistrements effectués à la Sorbonne de 1924 à 1930, dirigés par Hubert Pernot, alors directeur des Archives de la Parole à l'Institut de Phonétique de l'Université de Paris. Il comprend 23 enregistrements de 3 minutes chacun, réalisés de 1925 à 1929 sur des disques 78 tours et numérisés par la Bibliothèque nationale de France (BnF). Les enregistrements de 11 locuteurs masculins délivrant des discours dans un style déclamatoire,

correspondent à un style d'expression orale hyperarticulé et soutenu. Le contenu des enregistrements inclut, entre autres, des souvenirs des carrières scientifiques ou politiques d'intellectuels de l'époque.

Les enregistrements ont été transcrits en utilisant le logiciel de reconnaissance Google Cloud Speech-to-Text API, puis corrigés manuellement et segmentés en phonèmes à l'aide de WebMAUS (Kisler *et al.*, 2017). Malgré leur âge, la qualité des enregistrements était suffisamment satisfaisante pour garantir des mesures de formants robustes. La Figure 1 montre les exemples de spectres des /a, ɑ/ et le trapèze des voyelles orales réalisées en 1927 par un locuteur masculin né en 1856.

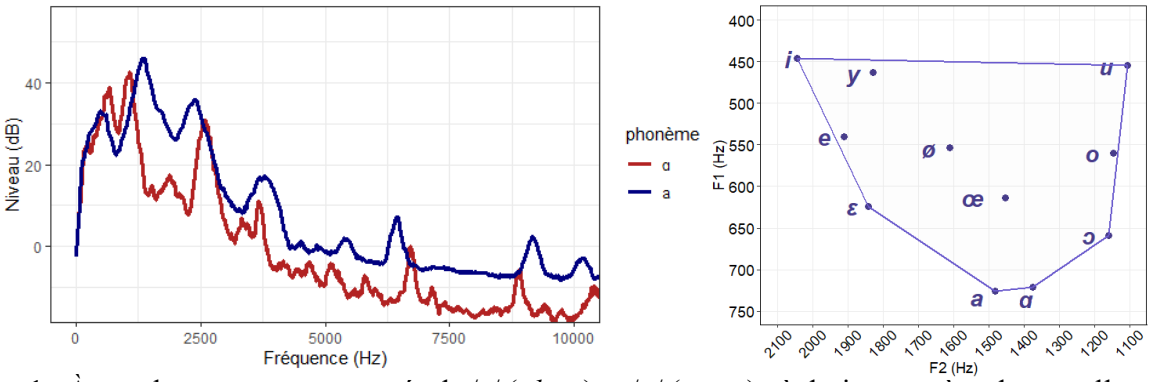


FIGURE 1 : À gauche : spectres comparés de /a/ (*place*) et /ɑ/ (*passe*) ; à droite : trapèze des voyelles orales (F1 et F2 en Hz) produites par un locuteur masculin, archives BnF, 1927.

Corpus INA 1940–1997

Le deuxième corpus, utilisé dans les études antérieures de Boula de Mareüil *et al.* (2011) et Cecelewski *et al.* (2023), contient 10 heures de discours réparties en 160 documents d'archives d'une durée allant de 20 sec. à 20 min., provenant de l'Institut National de l'Audiovisuel (INA) et incluant des émissions d'actualités (Barras *et al.*, 2002). Boula de Mareüil *et al.* (2011) ont trouvé que l'impact potentiel du bruit de fond sur les mesures était faible. Les rares voix féminines présentes dans l'après-guerre ont été exclues de l'analyse. Les données ont été regroupé en quatre périodes : 1940–1959, 1960–1969, 1970–1979 et 1980–1997. Pour l'analyse acoustique, les documents d'archive avaient été segmentés en phonèmes à l'aide d'un système d'alignement automatique développé par le LISN Paris-Saclay (Gauvain *et al.*, 2005), utilisant des modèles acoustiques indépendants du contexte et un dictionnaire de prononciation spécifiquement adapté au corpus.

Corpus ESTER

Le troisième corpus, fréquemment utilisé au sein de la communauté phonétique francophone, ESTER (Galliano *et al.*, 2006), représente près de 50 heures de discours journalistique, composé d'extraits d'émissions radiophoniques diffusées entre 1999 et 2004. Le corpus ESTER inclut des transcriptions orthographiques manuelles qui ont été phonétiquement transcrites et alignées automatiquement (Galliano *et al.*, 2006). Comme les corpus BnF et INA contiennent (presque) exclusivement des enregistrements de locuteurs masculins, notre analyse se concentrera uniquement sur les productions sonores de sujets masculins.

Corpus 2020–2023

Un dernier corpus a été inclus dans cette étude dans le but d'étendre l'analyse diachronique aux données les plus récentes. Les enregistrements, réalisés entre 2020 et 2023, d'une durée allant de 3 à 12 minutes, de six intervenants âgés de 25 à 35 ans, ont été collectés à partir d'une gamme de plateformes, incluant trois locuteurs de la chaîne publique française France Inter, deux de chaînes de communication scientifique sur YouTube et un d'une chaîne d'actualités TikTok. Le but de cette

sélection était de constituer un échantillon représentatif des styles de parole journalistique contemporains au sein du paysage médiatique français.
Ainsi, le choix d'intégrer de nouveaux contenus audiovisuels répond à l'évolution du style de parole des jeunes présentateurs, tendant vers un style plus spontané. Inversement, les contenus récents de radio présentent un style de parole plus proche des conventions médiatiques traditionnelles françaises, comme celles des corpus INA et ESTER avec un discours pré-écrit et plus contrôlé. Le protocole de transcription et d'alignement était identique à celui mis en place pour le corpus BnF.

2.2. Sélection des contextes pour l'analyse

Un premier filtrage a été effectué pour sélectionner parmi 130k contextes d'A (*antérieur* ou *postérieur*), 70 000 contextes d'A (*antérieur* ou *postérieur*) en position finale de mot, considérée comme la plus robuste pour l'opposition /a/ ~ /ɑ/ (Rousselot & Laclotte, 1902). Nous avons inclus les mots lexicaux monosyllabiques (ex. *las*), de même que les mots grammaticaux monosyllabiques postposés (ex. la particule de négation *pas*, contexte fréquent de /ɑ/). En revanche, nous avons exclu les mots grammaticaux antéposés (ex. les déterminants *la*, *ma*, ainsi que les formes de l'auxiliaire *avoir* dans les temps composés, *a*, *as*, en tant qu'apparaissant systématiquement en position non-accentuée à l'intérieur d'une unité prosodique).
Les dictionnaires de prononciation utilisés par les systèmes d'alignement ne distinguant pas entre /a/ et /ɑ/, les occurrences de ce dernier ont été annotées selon un système de règles spécialement conçu pour cette étude, fondé sur les travaux de Grammont (1938), Delattre (1957), Fouché (1959), et Malmberg (1969). L'opposition /a/ ~ /ɑ/ étant décrite comme particulièrement variable et sujette à la variation individuelle, l'enjeu de notre sélection a consisté à d'identifier un nombre suffisamment élevé d'occurrences de /ɑ/ dans chaque période pour obtenir des résultats robustes, tout en ciblant ces contextes où la probabilité de trouver un /ɑ/ postérieur dans les données anciennes était maximale.
Ainsi, nous avons codé /ɑ/ :

- sous l'orthographe <â>, sauf dans les terminaisons verbales *-ât*, *-âtes*, *-âmes* prononcées [a] depuis le XVIII^e^ siècle (Delattre, 1957) ;
- dans les formes en *-as*, *-ase*, *-aze*, *-az*, *-asse* : *hélas*, *extase*, *topaze*, *gaz*, *entasse* ;
- dans les groupes terminés par une liquide *-afle*, *-avre*, *-able*, *-abre*, *-acle*, *-adre* : *il rafle*, *havre*, *sable*, *sabre*, *miracle*, *cadre* ;
- dans certaines formes en *-amne* : *condamne*, en *-oi* : *trois*, *croît* ;
- dans la terminaison verbale *-a* : *donna*.

En revanche, nous avons codé /a/ les contextes faisant exception aux règles ci-dessus pour lesquels au moins deux des auteurs cités (Grammont, 1938 ; Delattre, 1957 ; Fouché, 1959 ; Malmberg, 1969) avaient classé la prononciation [ɑ] comme incertaine, notamment dans des formes en *-as* : *bras*, *cadenas*, *compas*, en *-oi* : *froid*, *mois*, en *-af(f)re* : *balafre*, *affres*, en *-asse* : *débarrasser*, et en *-asse*, *-asses*, *-assent* : *qu'il fasse*.
La Table 1 présente le nombre de contextes de /a/ et /ɑ/ sélectionnés pour l'analyse, séparément dans chaque période :

	1925–1929	**1940–1959**	**1960–1979**	**1980–1997**	**2000–2004**	**2020–2023**	
/a/	476	822	3373	2067	25825	1376	33989
/ɑ/	34	89	394	268	3484	168	4437
Total	510	911	3767	2335	29309	1544	38376

TABLE 1 : Nombre de contextes de /a, ɑ/ dans les corpus 1925–2023.

2.3. Contrôle de l'entourage consonantique des /a/ et /ɑ/

La distribution, essentiellement étymologique, des phonèmes /a/ et /ɑ/ se caractérise, d'un côté, par un pourcentage faible des /ɑ/ par rapport aux /a/ et, de l'autre, par des contraintes contextuelles spécifiques de la présence d'un /ɑ/ dans un ensemble de terminaisons et de suffixes. Le contexte consonantique ayant un impact sur le timbre vocalique (Menzerath & Lacerda, 1933 ; Öhman, 1966 ; Krull, 1989), nous avons contrôlé la distribution de différentes catégories de consonnes en contexte gauche et droit des /a/ ~ /ɑ/ soumis à l'analyse. Nous avons codé séparément les segments coronaux /d, z, ʒ, t, s, ʃ, n, ŋ l/, labiaux /b, v, p, f, m/, vélaires /g, k/, et la fricative uvulaire /ʁ/. La Figure 2 présente la répartition des différentes classes de segments consonantiques en contexte gauche et droit, séparément pour chaque période étudiée.

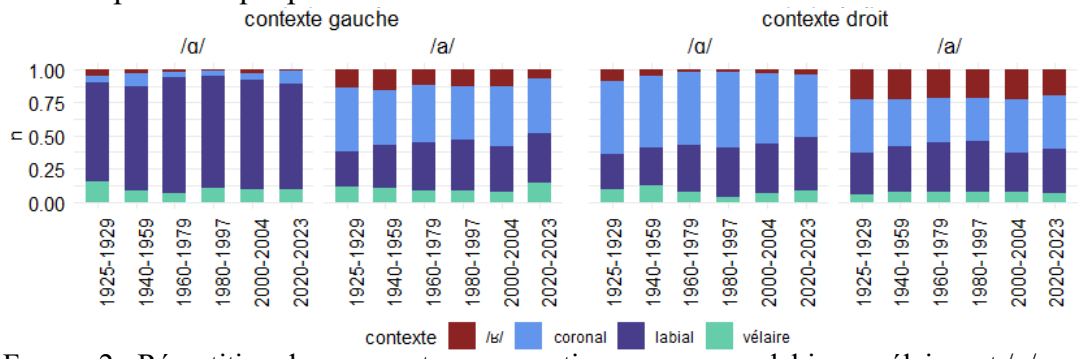


FIGURE 2 : Répartition des segments consonantiques coronaux, labiaux, vélaires et /ʁ/ en contexte gauche et droit, dans les corpus étudiés, par période.

Ainsi, les catégories de consonnes se répartissent de manière différente selon qu'il s'agisse de /a/ ou de /ɑ/ en contexte gauche ou droit. Une proportion importante de segments labiaux en contexte gauche devant /ɑ/ s'explique principalement par la présence de la particule de négation *pas*, qui constitue le contexte le plus fréquent de /ɑ/. En revanche, en raison des contraintes distributionnelles spécifiques (voir § 2.2.), on note une quasi-absence de /ʁ/ dans l'entourage phonétique immédiat de /ɑ/, aussi bien en contexte gauche que droit.

Outre ces particularités qui résultent de la distribution des /a/ et /ɑ/ dans le lexique français en parole continue, les contextes les plus fréquents sont, dans l'ordre décroissant, les segments coronaux > labiaux > /ʁ/ > vélaires. Cependant, ce qui importe pour notre étude n'est pas tant la variation de la répartition des contextes entre les catégories (contexte gauche ou droit, /a/ ou /ɑ/) mais plutôt la variation de leur distribution dans les différentes tranches de temps à l'intérieur de chaque catégorie. Ainsi, la variation de la répartition des contextes entre les différentes périodes au sein de chaque catégorie ne dépasse pas 5 % pour les segments vélaires et le /ʁ/, 10 % pour les segments coronaux et labiaux. Le résultat est une répartition suffisamment équilibrée pour isoler la variation diachronique de l'effet éventuel de la surfréquence d'un type particulier de segments consonantiques.

2.4. Mesures acoustiques

Les valeurs des deux premiers pics formantiques ont été extraites respectivement à ⅓, ½ et ⅔ de la durée de la voyelle à l'aide de l'algorithme Burg implémenté dans Praat (Boersma & Weenink, 2016), puis moyennées pour obtenir une valeur unique. Un script a été utilisé pour automatiser l'extraction des formants, effectuée avec les paramètres ci-après : préemphase à 50 Hz, plage de détection inférieure à 4,9 kHz et fenêtre d'analyse de 25 ms.

Dans l'étape suivante, nous avons procédé à l'élimination des valeurs aberrantes dans chacune des vingt catégories résultant de la combinaison des deux phonèmes (/a, ɑ/) et de six intervalles de temps. Les valeurs extrêmes ont été éliminées dans chaque catégorie en utilisant les seuils des 5^{e} et 95^{e}

percentiles. Des tests de validation statistique ont été conduits dans l'environnement R (R Development Core Team, 2010 ; version `2023.03.0-daily+82.pro2`).

3 Résultats

3.1. Étude en *temps réel* des voyelles /a/ et /ɑ/ entre 1925 et 2023

Les valeurs des deux premiers formants des réalisations des voyelles /a, ɑ/ en position finale de mot ont été représentées dans la Figure 3 pour illustrer l'évolution de ces segments entre 1925 et 2023. En guise d'illustration prolongeant l'empan temporel des corpus analysés, nous avons inclus, dans la Figure 3, les valeurs de F1 et de F2 des /a, ɑ/ extraites de deux enregistrements de récitation littéraire réalisée par un locuteur masculin anonyme en 1912, provenant de la collection BnF. Ainsi ces données permettent-elles d'estimer, ne serait-ce qu'à titre qualitatif, le chemin parcouru par les deux segments depuis le début du siècle, où la différence acoustique entre /ɑ/ et /a/ était encore particulièrement nette.

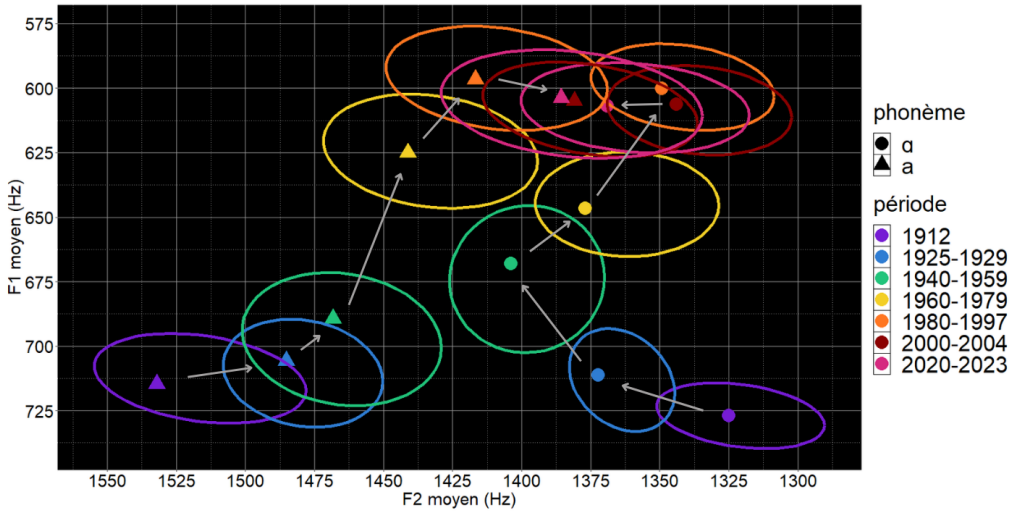


FIGURE 3 : Valeurs moyennes de F1 et F2 (en Hertz) et ellipses (réglées à 25 % des occurrences) des cibles vocaliques (F1, F2) des réalisations des voyelles /a/ et /ɑ/ en position finale de mot, par période.

Pour illustrer le rapprochement acoustique entre /ɑ/ et /a/, la Figure 4 représente les distances euclidiennes moyennes entre ces segments dans l'espace F1/F2. Deux stades du changement y sont repérables, correspondant à deux rapprochements successifs entre 1925–1929 et 1940–1959 et, après une période de stabilité, entre 1980–1997 et 2000–2004.

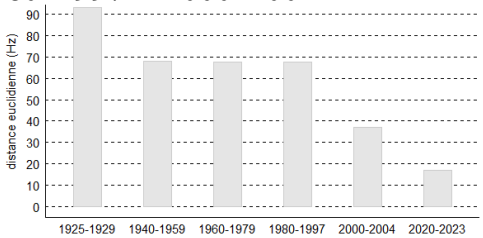


FIGURE 4 : Distances euclidiennes entre moyennes (Hz) entre /a/ et /ɑ/ en position finale de mot, formule : $d = \sqrt{((F2\_a - F2\_\alpha)^2 + (F1\_a - F2\_\alpha)^2)}$.

Afin de valider statistiquement l'évolution de F2 dans les corpus étudiés, nous avons réalisé un modèle de régression linéaire mixte incluant les effets fixes de Phonème (2 niveaux : /a/, /ɑ/, niveau de référence : /a/), Période (6 niveaux : 1925–1929, 1940–1959, 1960–1979, 1980–1997, 2000–2004, 2020–2023, niveau de référence : 1925–1929), un terme d'interaction entre ces deux prédicteurs, ainsi qu'un effet aléatoire pour 'mot'. Nous avons utilisé la fonction `lmer` du package `lme4` dans R. Les coefficients du modèle sont détaillés dans la Table 2.

Contraste	Estimation	Err. St.	*t*	*p*
Intercept	1494.25	6.18	241.67	**<.001**
1940–1959	-25.22	7.37	-3.42	**<.001**
1960–1979	-50.46	6.32	-7.97	**<.001**
1980–1997	-75.95	6.54	-11.60	**<.001**
2000–2004	-108.18	5.97	-18.10	**<.001**
2020–2023	-111.68	6.74	-16.56	**<.001**
Phonème(/ɑ/)	-69.00	18.82	-3.02	**<.01**
1940–1959*Phonème(/ɑ/)	-43.98	16.95	1.76	**<.05**
1960–1979*Phonème(/ɑ/)	-48.00	18.05	1.17	**<.05**
1980–1997*Phonème(/ɑ/)	-45.04	21.47	2.00	**<.05**
2000–2004*Phonème(/ɑ/)	-62.67	21.26	2.94	**<.01**
2020–2023*Phonème(/ɑ/)	-90.62	22.55	4.01	**<.001**

TABLE 2 : Statistiques récapitulatives des effets du modèle de régression linéaire. Formule : `lmer(F2 ~ période + phonème + période*phonème + (1|mot))`.

La Figure 5 représente les prédictions du modèle pour le terme d'interaction `Période*Phonème` illustrant le changement de hauteur de F2 entre 1925 et 2023, pour /a/ et /ɑ/.

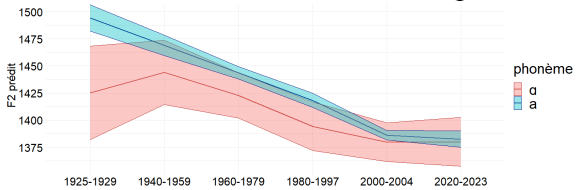


FIGURE 5 : Moyennes marginales prédites de F2 et barres d'erreur (95 %) pour le terme d'interaction `Phonème*Période`, ajustées selon le modèle de régression linéaire mixte.

Nous avons ensuite utilisé la fonction `emmeans()` du package R éponyme pour inspecter toutes les comparaisons (*pairwise contrasts*) entre les périodes consécutives, pour le terme d'interaction `Phonème*Période`. Toutes les comparaisons se sont révélées significatives (p<.001), hormis celles entre 1940–1959 : 1960–1979 (p=.08) et entre 1960–1979 et 1980–1997 (p=.05), ce qui nous amène à conclure qu'un changement significatif en termes de relation réciproque du corrélat acoustique d'antériorité (F2) des réalisations des voyelles /a/ ~ /ɑ/ s'est opéré entre 1925 et 2023, excepté une période de stabilité entre 1940 et 1997 (v. Figure 4). Ainsi, une postériorisation de /a/ conjointement à une antériorisation plus subtile de /ɑ/, ont abouti à une fusion acoustique entre les deux phonèmes, les valeurs de F2 (ni de F1) ne permettant plus guère de distinguer un /ɑ/ postérieur de sa contrepartie antérieure.

4 Discussion

Nous avons, dans cette étude, examiné l'évolution en *temps réel* (Bailey, 2013) du corrélat acoustique d'antériorité (F2) des réalisations des voyelles /a/ et /ɑ/ en position finale de mot, en parole lue (déclamatoire et journalistique) produite par des locuteurs masculins entre 1925 et 2023. Dans l'ensemble des corpus traités, les /ɑ/ ont été, indépendamment de la période, codés selon les règles de distribution établies d'après des traités de prononciation anciens. Une comparaison de valeurs de F1 et F2 des productions des /a, ɑ/ dans différents contextes phonétiques, non présentée ici, par contrainte de place, a permis de conclure à la présence d'un effet contextuel d'envergure similaire dans toutes les périodes étudiées.

Les résultats obtenus nous amènent à confirmer l'hypothèse de la fusion acoustique des /a/ et /ɑ/. Toutefois, contrairement à nos prévisions, plutôt qu'une antériorisation du /ɑ/ au fil du temps, nos résultats suggèrent une postériorisation du /a/ vers une position plus centrale comme ayant le plus contribué au rapprochement acoustique entre les deux segments. Un examen qualitatif des données de 1912 indiquant une distance plus importante (de près de 200 Hz) entre le F2 des réalisations de /a/ et /ɑ/ par rapport à la période 1925–1929, n'a fait que corroborer la variation diachronique mise en évidence.

Ces résultats font écho à ceux de l'enquête phonologique longitudinale de Hansen (2014) qui conclut à un rétrécissement de l'étendue des productions de l'opposition /a/ ~ /ɑ/, entre les années 1970 et 2000 et souligne qu'une perte de variantes très postérieures [ɑ] était accompagnée d'une postériorisation de variantes très antérieures [æ] vers une position plus centrale dans l'espace vocalique.

Quant au F1, corrélat acoustique de l'aperture, celui-ci connaît également une évolution, s'acheminant au fil du temps vers des valeurs plus basses, autant pour le /a/ que pour le /ɑ/. Ce changement n'est pas sans rappeler l'abaissement du F1 des réalisations des /e, ɛ, o, ɔ/ dans la même période, mis en évidence par Cecelewski *et al.* (2023). Corrélé à l'abaissement de la fréquence fondamentale en parole journalistique (Boula de Mareüil *et al.*, 2011), l'abaissement du F1 peut ainsi s'expliquer par l'évolution du style de parole vers moins d'effort vocal (Assmann *et al.*, 2008).

D'autres éléments de l'évolution des /a/ ~ /ɑ/ non présentés ici, par manque de place, méritent également d'être mentionnés. Contrairement aux descriptions traditionnelles (Delattre, 1957), le ratio de durée des deux voyelles ouvertes n'évolue guère dans la période analysée. La perte de la longueur étymologique des /ɑ/ semble donc bien s'être opérée avant 1925.

En outre, les distances euclidiennes (calculées en Hz dans l'espace F1/F2) entre /a, ɑ/ et les voyelles orales les plus proches, plus spécifiquement entre /a/ ↔ /ɛ/ et /ɑ/ ↔ /ɔ/, s'avèrent remarquablement stables dans le temps, oscillant respectivement autour de 375 et 275 Hz tout au long de la période étudiée. Il semble que nous ayons affaire à une évolution tendant vers un rééquilibrage du système des voyelles ouvertes et mi-ouvertes. Ce résultat paraît conforme aux prédictions de nombreuses modélisations postulant des tendances plus ou moins universelles à la symétrie des systèmes vocaliques dans les langues du monde (de Boer, 2001 ; Boë *et al.*, 1994 ; Crothers, 1978), exprimée ici par un déplacement du /a/ proportionnel au réagencement des segments les plus proches.

Ainsi dépeinte, l'évolution des /a, ɑ/ n'est sans doute qu'une facette de l'extrême diversité des usages du français métropolitain. Nous nous sommes limités à examiner une parole soutenue et soignée, produite par des locuteurs masculins cultivés, qui ne peut pas rendre compte des réalisations populaires ou régionales des /a, ɑ/ (Avanzi, 2021). Or, les études impressionnistes soulignent une différence de prononciation importante entre un style plus contrôlé de locuteurs cultivés — dans lequel la distinction entre /a/ et /ɑ/ paraît, somme toute, relativement subtile — et un « accent populaire », « faubourien » (Delattre, 1957) ou aujourd'hui « de banlieue », qui vélariserait davantage le /ɑ/, le rendant plus distinct perceptivement. Dans les archives analysées ici, l'opposition entre /a/ et /ɑ/ demeure en effet assez ténue perceptivement, et ce même dans les périodes les plus anciennes correspondant à la distance maximale entre /a/ et /ɑ/.

L'abaissement de F2 du /a/ au fil du temps suggéré par nos résultats devra être validé par des études supplémentaires. Nous ne pouvons pas exclure l'impact potentiel de facteurs non contrôlés ici, comme les conditions d'enregistrement et d'archivage, ainsi que la variation individuelle. Cette dernière peut inclure des facteurs tels que l'évolution de la taille moyenne des locuteurs dans le temps, ou encore l'ajustement musculaire habituel (*articulatory setting*, Laver, 1978). Une modélisation articulatoire (e.g. VocalTractLab, Birkholz, 2013) pourrait également apporter un éclairage supplémentaire sur les causes de la baisse de F2 des réalisations de /a/ mise en évidence dans notre étude.

Actuellement en création, un corpus diachronique plus grand sera destiné d'une part à modéliser l'évolution globale du système vocalique du français et d'autre part à examiner l'influence des contextes phonémique et lexical sur l'évolution des /a, ɑ/ au cours du XX^{e} siècle.

Remerciements

Ce travail a été soutenu par le projet ANR-21-CE38-0019 DIPVAR.

Références

ASSMANN P. F., NEAREY T. M., BHARADWAJ S. V., HUBBARD D. & JAYARAMAN A. (2008). Developmental study of the relationship between F0 and formant frequencies. *Journal of the Acoustical Society of America*, 122, EL35-EL43. DOI : 10.1121/1.2719045.

AVANZI M. (2021). Géographie des tendances centripètes et centrifuges du français en francophonie : le cas des oppositions phonologiques /a/~/ɑ/ et /ɛ̃/~/œ̃/. In T. CABRÉ & M. GÜELL, Éds., *Norme et diversité linguistique : la gestion normative dans des contextes pluricentriques. Francophonie et catalanophonie*, Barcelona : Servei de Publicacions de l'Institut d'Estudis Catalans, p. 35–74. HAL : 03321768.

BAILEY G. (2013). Real Time and Appartent Time. In J.K. CHAMBERS & N. SCHILLING, Éds., *The Handbook of Language Variation and Change*, John Wiley & Sons, chapter 11, p. 237–262. DOI : 10.1002/9781118335598.ch11.

BARADUC J., BERGOUNIOUX G., CASTELLOTTI V., DUMONT C. & LANSARI M. H. (1989). Le statut linguistique des voyelles moyennes. *Langage et société*, 49, 5–24.

BARRAS C., ALLAUZEN A., LAMEL, L. & GAUVAIN J. L. (2002). Transcribing audio-video archives. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando (USA), 13–16. DOI : 10.1109/ICASSP.2002.5743642.

BIRKHOLZ P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4), e60603. DOI : 10.1371/journal.pone.0060603.

BOË L. J., SCHWARTZ J. L. & VALLÉE N. (1994). The prediction of vowel systems: perceptual contrast and stability. In E. KELLER, Éds., *Fundamentals of speech synthesis and speech recognition*, Chichester : John Wiley, p. 185–213.

BOINDIN N. (1753). *Oeuvres de Monsieur Boindin, de l'Académie Des Inscriptions Et Belles Lettres Tome Second. Contenant des remarques sur les sons de la langue, et sur les noms des Romains, avec des discours sur les tribus romaines et le théâtre des anciens*, Paris : Prault fils.

BOULA DE MAREÜIL P., RILLIARD A. & ALLAUZEN A. (2011). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. *Language and Speech*, 55(2), 263–293. DOI : 10.1177/0023830911417799.

BOULA DE MAREÜIL P., ADDA-DECKER M. & WOEHRLING C. (2010), Antériorisation/aperture des voyelles /ɔ/~/o/ en français du Nord et du Sud, *28es Journées d'Étude sur la Parole*, Mons, 81–84.

CARRÉ R. & MRAYATI M. (1995). Vowel transitions, vowel systems, and the distinctive region model. In C. SORIN, J. MARIANI & H. MELONI, Éds., *Levels in speech communication: relations and interactions*, The Netherlands : Elsevier, p. 73–90.

CECELEWSKI J., GENDROT C., ADDA-DECKER M. & BOULA DE MAREÜIL P. (2023). A diachronic study of vowel harmony in French broadcast speech since 1940. *Proceedings of the 20th International Congress of Phonetic Sciences*, 798–802. HAL : hal-04204781.

CROTHERS J. (1978). Typology and universals in vowel systems. In J. H. GREENBERG, C. A. FERGUSON & E. A. MORAVCSIK, Éds., *Universals of human language*, Stanford : Stanford University Press, p. 93–152.

DE BOER B. (2001). *The Origin of Vowel Systems*. Oxford : University Press. ISBN : 9780198299660.

DE LA TOUCHE P. (1730). *L'art de bien parler françois, Qui comprend tout ce qui regarde la Grammaire & les façons de parler douteuses*, Amsterdam : Wetsteins et Smith.

DELATTRE P. (1957). La Question des deux « a » en français, *The French Review*, 31(2), 141–148.

ESTIENNE R. (1557). *Traicté de la grãmaire Francoise*. Genève.

FOUCHE P. (1959). *Traité de prononciation française*. Paris : C. Klincksieck.

FOUCHÉ, P. (1935) L'évolution phonétique du français du 16e siècle à nos jours. In A. DAUZAT, Éds., *Où en sont les études de français*, Bibliothèque du « français moderne », Paris, p. 35–54.

GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J. F., MOSTEFA, M. & CHOUKRI, K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, p. 139–142.

GAUVAIN J. L., ADDA G., ADDA-DECKER M., ALLAUZEN A., GENDNER V., LAMEL L. & SCHWENK H. (2005). Where are we in transcribing French broadcast news ? *Proceedings of the Ninth European conference on speech communication and technology*, p. 1665–1668. DOI : 10.21437/Interspeech.2005-544.

GRAMMONT M. (1938). *La Prononciation française, traité pratique*. Paris : Delagrave.

HANSEN A. B. & JUILLARD C. (2011). La phonologie parisienne à trente ans d'intervalle – Les voyelles à double timbre. *Journal of French Language Studies*, 21, 313–359. DOI : 10.1017/S0959269510000347.

HANSEN A. B. (2014). Lexique et phonologie. Le cas de la perte de distinction /a/ – /ɑ/ en français parisien. In B. LAKS & J. PEUVERGNE, Éds., *La phonologie du français : Normes, périphéries, modélisation : mélanges pour Chantal Lyche*, Presses Universitaires de Paris Ouest, p. 261–284. ISBN : 978-2-84016-203-2.

KISLER T., REICHEL U. D. & SCHIEL F. (2017). Multilingual processing of speech via web services, *Computer Speech & Language*, 45, 326–347. DOI : 10.1016/j.csl.2017.01.005.

KRULL D. (1989). Consonant–vowel coarticulation in reference words. *Quarterly progress and status report*, 30, 101–105, KTH Department for Speech, Music, and Hearing, STLQPSR.

LAMY B. (1688). *La rhétorique, ou L'art de parler*. Paris : André Pralard.

LAVER J. (1978). The Concept of Articulatory Settings: An Historical Survey. *Historiographia Linguistica*, 5(1-2), 1–14. DOI : 10.1075/hl.5.1-2.02lav.

LÉON P. (1992). *Phonétisme et prononciations du français, avec des travaux pratiques d'application et leurs corrigés*. Paris : Nathan.

LILJENCRANTS J. & LINDBLOM B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–862.

MALMBERG B. (1969). *Phonétique française*. Malmö : Hermods.

MENZERATH P. & LE LACERDA A. (1933). *Koartikulation, Steuerung und Lautabgrenzung*. Berlin : Dümmler.

METTAS O. (1970). Étude sur le A dans deux sociolectes parisiens. *Revue Romane,* 5(1), 94–105.

ÖHMAN S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–168.

OUDIN A. (1640). *Grammaire françoise rapportée au langage du temps*. Paris : Sommaville.

PERNOT H. (1928). Les voyelles parisiennes, *Revue de Phonétique*, t. V, Paris.

ROUSSELOT P. & LACLOTTE F. (1902). *Précis de prononciation française*, Paris, Leipzig : Welter.

SCHWARTZ J. L., BOË L. J., VALLEE N. & ABRY C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25, 255–286.

WALTER H. (1977). *La phonologie du français*. Paris : Presses universitaires de France.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Exploration de la représentation multidimensionnelle de paramètres acoustiques unidimensionnels de la parole extraits par des modèles profonds non supervisés.*

Maxime Jacquelin[1, 2] Maëva Garnier[1] Laurent Girin[1] Rémy Vincent[2]
Olivier Perrotin[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
(2) Vogo, F-38190 Bernin, France
maxime.jacquelin, maeva.garnier, laurent.girin,
olivier.perrotin@grenoble-inp.fr, r.vincent@vogo-group.com

RÉSUMÉ

Cet article propose une méthodologie pour interpréter les dimensions de variation de la parole conversationnelle, extraites de façon non-supervisée, et sur des données multilocuteurs, par un algorithme d'apprentisage profond (Auto-Encodeur Variationnel). Par des analyses de corrélation et de similarité cosinus, nous montrons que la distribution de la fréquence fondamentale et de la fréquence centrale des trois premiers formants de l'ensemble d'apprentissage est encodée par une direction dédiée de l'espace latent. Lorsque la distribution est multimodale, les différents modes du paramètre acoustique sont encodés dans des dimensions distinctes. De plus, nous avons identifié les directions expliquant la variation des paramètres au sein de chaque mode, et entre eux.

ABSTRACT

Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models

This paper proposes a methodology for interpreting the dimensions of variations of conversational speech, extracted in an unsupervised manner, and on multi-speaker data, by a deep learning algorithm (Variational Auto-Encoder). Using correlation and cosine similarity analyses, we show that the distribution of the fundamental frequency and the central frequencies of the first three formants of the training set is encoded by one dedicated latent space direction. When the distribution is multimodal, different modes of the acoustic feature are encoded in separate dimensions. In addition, we also have identified the directions that explain the variation of the feature within and across modes.

MOTS-CLÉS : apprentissage de représentations, codage de la parole, auto-encodeur variationnel, modèle source-filtre.

KEYWORDS: representation learning, speech encoding, variational autoencoder, source-filter model.

*. note aux relecteurs : cette soumission est une traduction d'un article publié à XAI-SA IEEE ICASSP Workshop, Explainable Machine Learning for Speech and Audio, 2024 (Jacquelin *et al.*, 2024).

1 Introduction

Depuis les modèles physiques jusqu'aux approches d'apprentissage profond sur données massives, la modélisation de la parole trouve des applications dans la reconnaissance automatique de la parole, le codage de la parole, la synthèse vocale expressive ou la conversion de voix. Pour chaque cas, l'objectif de la modélisation de la parole est de comprendre comment les signaux sont générés et comment leurs caractéristiques acoustiques peuvent être modulées à partir d'un nombre limité de dimensions de contrôle, aussi indépendantes que possible les unes des autres.

En ce sens, les premiers modèles acoustiques tels que le modèle source-filtre de Fant (Fant, 1971) établissent une distinction claire entre les variations acoustiques liées à la source glottique (variations de la fréquence fondamentale f_0, d'apériodicité, d'inharmonicité ou de la pente spectrale) et celles, supposées indépendantes, liées à l'articulation du conduit vocal (formants ou pics spectraux dans les bruits turbulents). Bien qu'un tel ensemble de paramètres acoustiques présente l'avantage d'être facilement interprétable en termes de physiologie et de contrôle gestuel sous-jacent, ils sont largement interdépendants, avec des contraintes anatomiques et physiques qui sous-tendent les covariations de ces paramètres au sein d'un même individu et d'un individu à l'autre (Coleman, 1971).

Bien que les modèles génératifs profonds non- et auto-supervisés soient des outils puissants pour modéliser toute la complexité des signaux de parole (Kingma & Welling, 2014; Van Der Oord *et al.*, 2017; Baevski *et al.*, 2020; Hsu *et al.*, 2021; Lakhotia *et al.*, 2021), peu de recherches ont été menées jusqu'à présent sur l'interprétation de leurs représentations latentes. Plusieurs études ont déjà utilisé des auto-encodeurs variationnels (VAE) (Kingma & Welling, 2014), VQ-VAE (Van Der Oord *et al.*, 2017), ou des modèles auto-supervisés tels que HuBERT (Hsu *et al.*, 2021), pour trouver des espaces de représentation des variations discrètes de la parole, avec des dimensions qui distinguent les informations phonémiques (de bas niveau) de celles liées à la langue (Williams *et al.*, 2021), à l'identité du locuteur (Chou *et al.*, 2018) et/ou au "style" de parole (Williams & King, 2019) (de haut niveau). D'autres études récentes ont été en mesure de trouver un espace latent de faible dimension pour représenter et contrôler les variations acoustiques continues de la parole expressive (en termes d'intonation, de contenu spectral ou de rythme) (Blaauw & Bonada, 2016; Hsu *et al.*, 2017; Wang *et al.*, 2018; Zhang *et al.*, 2019; Tits *et al.*, 2019; Bous & Roebel, 2022; Lenglet *et al.*, 2022b; Vaidya *et al.*, 2022; Sadok *et al.*, 2023). Elles ont montré que, même avec une approche d'apprentissage entièrement non-supervisée, les paramètres acoustiques étaient encodés selon différents sous-espaces quasi orthogonaux de la représentation apprise. Cela leur a permis de contrôler certains aspects de l'intonation liés à la source glottique, de manière presque indépendante des variations de l'enveloppe spectrale, liées à l'articulation du conduit vocal.

Un phénomène qui reste cependant inexpliqué est que chaque paramètre acoustique est souvent encodé par plusieurs dimensions latentes (Sadok *et al.*, 2023), et la question de savoir quel type d'information est capturé par chacune de ces dimensions reste peu explorée. Parmi les multiples interactions possibles entre les paramètres acoustiques, nous faisons l'hypothèse dans cette étude que *la multiplicité des dimensions latentes observées peut refléter l'encodage des différentes sources de variabilité inter- et intra-individuelle de chaque paramètre acoustique*. À notre connaissance, il s'agit de l'une des premières tentatives visant à étudier l'interaction entre les représentations vocales de bas niveau (paramètres acoustiques) et de haut niveau (liées au locuteur, telles que le genre) dans les espaces latents. Pour cela, le choix d'un VAE offre plusieurs avantages par rapport à d'autres méthodes non- ou auto-supervisées. D'abord, sa nature stochastique et la régularisation de son espace latent fournit une représentation latente désenchevêtrée, ce qui permet de modéliser les variations des

caractéristiques de la parole dépendantes et indépendantes du locuteur dans des directions distinctes de l'espace latent. Ensuite, les VAEs permettent une forte réduction de dimension, ce qui favorise l'interprétabilité de l'espace latent. Enfin, les modèles auto-supervisés tels que wav2vec 2.0 ou HuBERT (Baevski *et al.*, 2020; Hsu *et al.*, 2021) sont entraînés en utilisant une tâche de clustering de phonèmes, donnant ainsi la priorité à l'encodage des informations phonétiques au détriment de la prosodie ou de la paralinguistique, qui sont souvent modélisées séparément (Polyak *et al.*, 2021). Dans notre étude, il est crucial de garantir l'encodage à la fois des informations vocales de bas niveau et de haut niveau dans l'espace latent du modèle.

Nous optons donc pour une approche basée sur les VAEs et introduisons une méthodologie reposant sur l'analyse en composantes principales (PCA), la régression linéaire (LR) et l'analyse discriminante linéaire (LDA), afin d'analyser et d'interpréter l'aspect multidimensionnel de la représentation des paramètres acoustiques individuels, avant de tester notre hypothèse.

2 Méthodologie

2.1 Entraînement du VAE

L'architecture VAE utilisée dans cette étude est similaire à celle utilisée par Sadok *et al.* (2023). Elle prend en entrée des trames de spectrogrammes d'amplitude de la transformée de Fourier à court terme (TFCT) de taille 513. La dimension du vecteur latent $\mathbf{z}$ est fixée à 16. L'encodeur se compose de trois couches cachées entièrement connectées de 256, 64 et 2×16 unités (pour les vecteurs de moyenne et de variance de $\mathbf{z}$), toutes avec une activation tangente hyperbolique. Le décodeur est construit symétriquement à l'encodeur.

Deux modèles VAEs indépendants ont été entraînés, l'un avec la base de données VCTK (Yamagishi *et al.*, 2019), l'autre avec la base de données Att-HACK (Le Moine & Obin, 2020). VCTK comprend 109 locuteurs anglais lisant les mêmes 400 énoncés. Att-HACK comprend 25 locuteurs français qui ont acté les mêmes 100 énoncés dans quatre attitudes sociales (amicale, distante, dominante et séductrice) avec 3 à 5 répétitions. Nous appelons VAE-VCTK le modèle entraîné sur VCTK et VAE-AH le modèle entraîné sur Att-HACK. L'ensemble d'entraînement VAE-VCTK contient 25 heures de parole provenant de 29 locuteurs féminins et 29 locuteurs masculins, sélectionnés de manière aléatoire. L'ensemble de validation contient 3 heures provenant de 10 locuteurs féminins et 10 locuteurs masculins qui n'ont pas été utilisés pour l'entraînement. L'ensemble d'entraînement VAE-AH contient 20 heures de parole de 7 locuteurs féminins et 7 locuteurs masculins, sélectionnés de manière aléatoire. L'ensemble de validation contient 3 heures provenant de 2 locuteurs féminins et 2 locuteurs masculins non utilisés pour l'entraînement. Tous les signaux ont été sous-échantillonnés de 48 pour VCTK et 44.1 pour Att-HACK à $16\,\mathrm{kHz}$. La TFCT a été définie avec une fenêtre de Hanning de $64\,\mathrm{ms}$ et un chevauchement de $50\,\%$.

Les modèles ont été entraînés avec l'optimiseur Adam (Kingma & Ba, 2015) sur 500 époques, avec une taille de batch de 128 et un taux d'apprentissage de 10^{-4}. Nous avons utilisé le VAE d'Itakura-Saito (IS), c'est-à-dire que la fonction de coût est la somme pondérée d'une divergence IS pour le terme de reconstruction et de la divergence de Kullback-Leibler (KL) pour le terme de régularisation (Girin *et al.*, 2019) : $\mathcal{L}_{total} = \mathcal{L}_{\mathsf{IS}} + \beta\mathcal{L}_{\mathsf{KL}}$. Pour éviter le problème de disparition du gradient rencontré dans les VAEs, le critère de régularisation β a été utilisé (Higgins *et al.*, 2017).

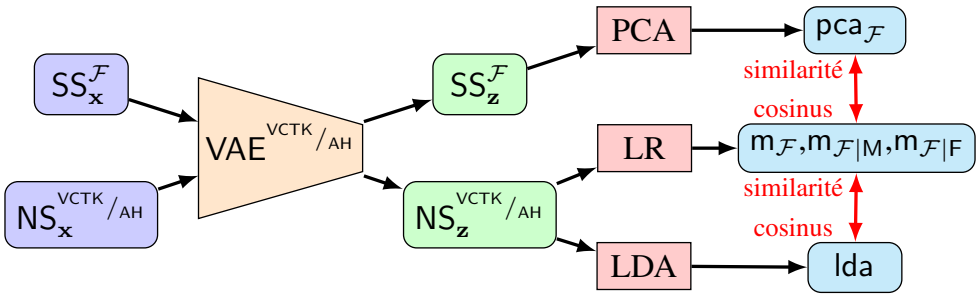


FIGURE 1 – Résumé de la méthodologie de test proposée.

2.2 Ensembles de test

Nous avons conçu plusieurs ensembles de tests pour évaluer les capacités de représentation de nos deux VAE. Tout d'abord, quatre ensembles de données ont été conçus pour démontrer l'aspect multidimensionnel de la représentation latente de chaque paramètre acoustique. Suivant Sadok *et al.* (2023), nous avons utilisé le logiciel Soundgen (Anikin, 2019) pour générer quatre signaux de 5 s avec une variation respective de f_0 et de la fréquence des trois premiers formants F_1, F_2, ou F_3. Quand l'un des paramètres varie, les trois autres paramètres sont laissés constants, fixés à la médiane de leur distribution dans les deux ensembles de données VCTK et Att-HACK (c'est-à-dire 140 Hz, 450 Hz, 1600 Hz et 2800 Hz, pour f_0, F_1, F_2 et F_3, respectivement). Pour chaque signal, la plage de variation du paramètre correspondant correspond à sa distribution dans les ensembles de données VCTK et Att-HACK : 85–310 Hz pour f_0 ; 290–890 Hz pour F_1 ; 960–2360 Hz pour F_2 ; et 2000–3430 Hz pour F_3. Nous avons ensuite calculé le spectrogramme d'amplitude de chaque signal, chacun constituant une base de données de test appelé $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$, $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, avec SS pour "synthesis speech " (parole synthétique). $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ sont les ensembles de test correspondants dans l'espace latent, c'est-à-dire $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$ passés par l'encodeur VAE.

Ensuite, pour analyser la capacité des VAEs à représenter les covariations naturelles entre les paramètres acoustiques, nous avons généré deux ensembles de test supplémentaires appelés $\mathsf{NS}_{\mathbf{x}}^{\mathsf{VCTK}}$ pour VAE-VCTK et $\mathsf{NS}_{\mathbf{x}}^{\mathsf{AH}}$ pour VAE-AH (avec $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ les ensembles de test correspondants dans l'espace latent des VAE), NS signifiant "natural speech" (parole naturelle). Ces ensembles de données sont constitués de 3 heures de signaux de parole naturelle provenant de 9 femmes et 9 hommes pour VCTK, et de 3 femmes et 3 hommes pour Att-HACK, qui ne font partie ni de l'ensemble d'apprentissage ni de l'ensemble de validation. Pour chaque signal de cet ensemble de test, une analyse acoustique par trame a été effectuée avec Praat (Boersma & Weenink, 2001) pour extraire f_0 et $F_{1,2,3}$.

2.3 Analyse PCA, LR et LDA

Notre analyse vise à identifier les directions dans l'espace latent qui capturent la plus grande variabilité expliquée par chaque paramètre acoustique de parole considéré. Notre première étape a été d'*étudier l'encodage de chaque paramètre acoustique séparément*, en utilisant les ensembles de données de test *de parole synthétique* $\mathsf{SS}_{\mathbf{x}}^{\mathcal{F}}$ spécifiquement conçus à cette fin. Indépendamment pour chaque paramètre acoustique $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, nous avons appliqué une PCA sur les bases de données encodées $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$, afin d'extraire par le biais des composantes principales notées $\mathsf{pca}_{\mathcal{F}}$ les multiples directions dans l'espace latent qui expliquent la variation du paramètre acoustique.

Notre deuxième étape consiste à *identifier le rôle de ces dimensions multiples*, grâce à l'analyse de *la parole naturelle*. Pour ce faire, nous avons recherché la direction de variation (DV) de chaque paramètre acoustique dans nos bases de données de test de parole naturelle encodée $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$. Indépendamment pour chaque paramètre $\mathcal{F}$, nous avons calculé une régression linéaire des valeurs $\mathcal{F}$, extraites de $\mathsf{NS}_{\mathbf{x}}$ avec Praat, sur les valeurs correspondantes de $\mathbf{z}$ dans $\mathsf{NS}_{\mathbf{z}}$. La DV de $\mathcal{F}$, notée $\mathsf{m}_{\mathcal{F}} \in \mathbb{R}^{16}$, est le vecteur des coefficients de la LR, c'est-à-dire :

$$\hat{\mathcal{F}} = \mathsf{m}_{\mathcal{F}}^{\top}\mathbf{z} + b_{\mathcal{F}} \approx \mathcal{F}, \tag{1}$$

au sens des moindres carrés ($b_{\mathcal{F}}$ est l'ordonnée à l'origine et $^{\top}$ désigne l'opérateur de transposition). Pour identifier le rôle de chaque dimension de la PCA ($\mathsf{pca}_{\mathcal{F}}$), nous avons ensuite analysé leur colinéarité avec les DV des paramètres acoustiques extraits dans des conditions spécifiquement choisies. En particulier, pour vérifier notre hypothèse, à savoir si les différentes dimensions latentes reflètent des sources de variabilité inter- et intra-individuelle de chaque paramètre acoustique, nous avons mesuré la DV sur l'ensemble du jeu de test, ainsi que la DV pour chaque genre de locuteurs. Nous désignons la DV résultante par $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$ et $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$ pour les locuteurs masculins et féminins, respectivement.

La représentation possible des paramètres acoustiques liés au genre dans des directions distinctes nous amène à faire un pas de plus pour *identifier une représentation indépendante de la variabilité inter- et intra-genre* dans l'espace latent. Le genre étant l'une des caractéristiques les plus discriminantes entre individus dans la parole, nous émettons l'hypothèse qu'une LDA calculée sur les locuteurs d'une base de données encodées $\mathsf{NS}_{\mathbf{z}}$ (noté lda), qui trouve la combinaison linéaire des dimensions latentes qui discrimine le mieux les locuteurs, devrait afficher une direction inter-genre sur sa première composante, et donc une direction intra-genre sur les composantes restantes. Pour vérifier que cette LDA met en évidence une représentation démêlée de la variabilité inter- et intra-genre, nous avons analysé la colinéarité entre les composantes de lda et les DV calculées sur des valeurs de f_0 propres à chaque genre et indépendantes du genre. La figure 1 résume notre analyse, réalisée indépendamment pour chaque paramètre acoustique $\mathcal{F}$. La colinéarité des directions extraites par PCA, LR ou LDA a été évaluée à l'aide de la similarité cosinus (CS) entre $\mathsf{pca}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$, and lda.

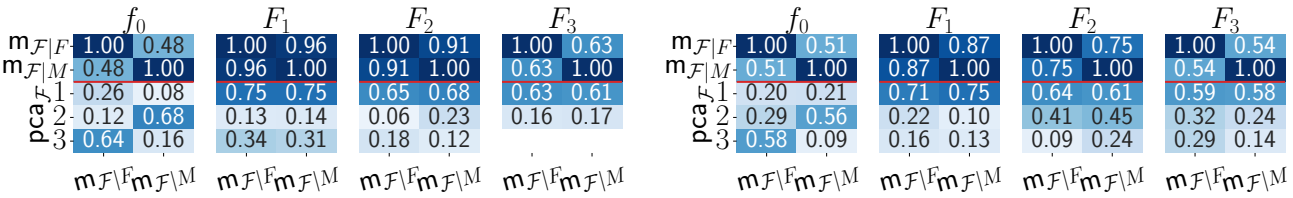
3 Résultats

3.1 Représentation multidimensionnelle des paramètres acoustiques

La première étape de notre analyse consiste à étudier séparément la représentation de chaque paramètre acoustique par les VAEs. Pour VAE-VCTK, les quatre PCA distinctes appliquées à $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ ont montré que trois composantes principales (PC) sont nécessaires pour expliquer au moins $80\,\%$ de la variance de chaque base de données encodée, à l'exception de F_3 (deux PC). Dans le cas de VAE-AH, chaque $\mathcal{F}$ a besoin de cinq PCs pour expliquer $80\,\%$ de la variance. En particulier, les variances minimales expliquées par les premières PCs sont d'environ $43\,\%$ pour VAE-VCTK sur f_0 et $36\,\%$ pour VAE-AH sur f_0. Nous avons également observé pour les deux VAEs que toutes les premières PCs de paramètres acoustiques différents sont relativement orthogonaux entre eux, avec une valeur maximale de CS de 0.33 entre pca_{f_0} et pca_{F_1}. Bien que ces résultats soient similaires à ceux de Sadok *et al.* (2023), nous n'avons pas obtenu le même nombre de PCs pour expliquer $80\,\%$ de la variance et une plus grande orthogonalité entre les PCs était rapportée. Ces différences suggèrent une dépendance possible de ces analyses aux données de parole utilisées pour l'entraînement et le test.

	m_{f_0}	$\mathsf{m}_{f_0\|\mathsf{F}}$	$\mathsf{m}_{f_0\|\mathsf{M}}$	m_{F_1}	$\mathsf{m}_{F_1\|\mathsf{F}}$	$\mathsf{m}_{F_1\|\mathsf{M}}$	m_{F_2}	$\mathsf{m}_{F_2\|\mathsf{F}}$	$\mathsf{m}_{F_2\|\mathsf{M}}$	m_{F_3}	$\mathsf{m}_{F_3\|\mathsf{F}}$	$\mathsf{m}_{F_3\|\mathsf{M}}$
VAE-VCTK	**0.65**	**0.64**	**0.58**	0.37	**0.73**	**0.75**	0.40	**0.71**	**0.74**	0.32	**0.64**	**0.58**
VAE-AH	**0.61**	**0.58**	**0.53**	0.31	**0.58**	**0.61**	0.36	**0.65**	**0.67**	0.27	**0.56**	**0.52**

TABLE 1 – Score de régression R^2 pour toutes les LRs testées



(a) VAE-VCTK (b) VAE-AH

FIGURE 2 – Similarité cosinus entre $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$ et $\mathsf{pca}_{\mathcal{F}}$. Pour plus de clarté, seules les trois premières composantes principales sont présentées.

Cependant, l'analyse des colinéarités entre les PCs extraites pour chaque paramètre a montré que notre VAE crée un équilibre entre les représentations latentes. D'une part, une pseudo-indépendance des paramètres de source (f_0) et de filtre ($F_{1,2,3}$) sur la première PC qui permet une modélisation séparée des variations d'intonation et d'articulation, et d'autre part une colinéarité plus forte sur les autres PCs qui laisse envisager la modélisation des covariations bien connues entre les paramètres acoustiques (Titze, 2004; Sundberg & Nordenberg, 2006). Ces résultats sont observés pour nos deux VAEs entraînés sur des données différentes, et ont également été rapportés par Sadok *et al.* (2023). Cela montre une propriété générale de la représentation multidimensionnelle des paramètres acoustiques unidimensionnelles considérées dans l'espace latent du VAE, qui sera examinée plus en détail dans la section suivante.

3.2 Interprétation des dimensions apprises

Nous avons observé que les variations de chaque paramètre acoustique, lorsqu'elles sont isolées dans l'ensemble de test de la parole synthétique, sont encodées par au moins deux directions dans l'espace latent du VAE. Notre hypothèse est que ces multiples dimensions sont nécessaires pour modéliser les variations acoustiques inter- et intra-individuelles de la parole naturelle. Ainsi, pour vérifier la fonction de chaque dimension, nous étudions maintenant les DV de chaque paramètre en contexte, c'est-à-dire sur les ensembles de tests de parole naturelle encodée $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ et données par $\mathsf{m}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$, et $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$. Nous essayons alors de corréler ces directions avec celles observées sur chaque ensemble de tests de parole synthétique $\mathsf{SS}_{\mathbf{z}}^{\mathcal{F}}$ ($\mathsf{pca}_{\mathcal{F}}$).

Le score de régression R^2 pour chaque DV est donné dans le Tab. 1, et la Fig. 2 affiche les CS entre $\mathsf{m}_{\mathcal{F}|\mathsf{M}}$ et $\mathsf{m}_{\mathcal{F}|\mathsf{F}}$ (obtenues sur la parole naturelle) et $\mathsf{pca}_{\mathcal{F}}$ (obtenues sur la parole synthétique). Pour les deux modèles et les trois formants, nous pouvons observer un contraste entre les petites valeurs de R^2 obtenues sur les DV globales (m_{F_i}, $i \in \{1,2,3\}$) et les valeurs élevées de R^2 obtenues sur les DV en fonction du genre. Parallèlement, nous observons une CS élevée entre $\mathsf{m}_{F_i|\mathsf{M}}$ et $\mathsf{m}_{F_i|\mathsf{F}}$, pour $i \in \{1,2,3\}$ (Fig. 2). Tout cela montre que, pour les deux modèles, les valeurs de fréquence des formants sont encodées linéairement dans l'espace latent lorsque l'on considère les deux genres séparément, avec des DV assez similaires, mais des ordonnées à l'origine différentes. En outre, la première PC de pca_{F_i} est la plus corrélée avec les DV des deux genres pour les deux modèles.

En ce qui concerne f_0, le Tab. 1 montre des scores de régression élevés pour les deux modèles dans toutes les conditions (par genre et globalement). En outre, la Figure 2 montre que $\mathsf{m}_{f_0|\mathsf{M}}$ et $\mathsf{m}_{f_0|\mathsf{F}}$ sont les DVs par genre les moins corrélées entre elles, mais qu'elles sont les plus corrélées respectivement avec la deuxième et la troisième PC de pca_{f_0}. Rappelons que pca_{f_0} est calculé sur l'ensemble de test de parole synthétique $\mathsf{SS}_{\mathbf{z}}^{f_0}$, pour lequel aucun paramètre autre que f_0 ne varie dans le signal d'entrée, c'est-à-dire qu'aucune autre information sur le genre n'est disponible. Pourtant, sur cette base de données, les deux VAEs sont capables de distinguer les valeurs de f_0 qui sont plus susceptibles d'appartenir à des locuteurs masculins ou féminins. Nous supposons que les modèles ont appris la distribution bimodale des valeurs f_0 rencontrées dans leurs ensembles d'apprentissage respectifs et qu'ils sont capables de distinguer les trames synthétiques sur la base de ces distributions.

Pour tester cette hypothèse, nous avons calculé les corrélations entre la distribution de f_0, mesurée sur chaque base de données utilisée pour entraîner les VAEs, et la projection des vecteurs latents $\mathbf{z}$ dans $\mathsf{SS}_{\mathbf{z}}^{f_0}$ sur les PCs de pca_{f_0} (en résumé, les coefficients de la PCA pour f_0). La corrélation la plus élevée a été obtenue avec la première PC de pca_{f_0} pour les deux VAEs (VAE-VCTK : 0.48, VAE-AH : 0.53). Comme on peut le voir sur la Fig. 3, les deux principaux pics du profil du premier coefficient pca_{f_0} sont proches des médianes des deux modes la distribution f_0 pour les deux modèles. De plus, les valeurs de PC sont élevées pour les deux modes de la distribution de f_0, alors qu'elles sont proches de 0 ou négatives lorsque les deux modes se confondent, modélisant ainsi l'incertitude de la classification entre les trames de parole masculines ou féminines. Nous avons mené la même expérience et observé un comportement similaire sur les trois formants, pour nos deux modèles VAE. Pour chaque formant, la deuxième PC est la plus corrélée avec la distribution de fréquence des formants (valeur absolue de corrélation supérieure à 0.8 pour les deux VAEs). Dans chaque cas, la distribution est unimodale, ce qui explique de manière cohérente la corrélation d'une seule autre PC avec le DV de la valeur du formant, comme démontré précédemment (Fig. 2).

Dans l'ensemble, nous avons montré que la représentation multidimensionnelle d'un paramètre acoustique unique est étroitement liée à la multimodalité de la distribution du paramètre. Pour chaque paramètre, nous avons constaté qu'une PC encode la distribution du paramètre qui est apprise à partir de l'ensemble d'apprentissage, et que les variations de paramètres acoustiques spécifiques à chaque mode sont encodées par quelques autres PCs distinctes.

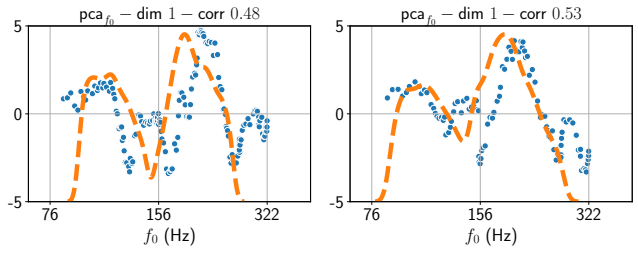


FIGURE 3 – Distribution des valeurs f_0 sur les données d'entraînement (en orange) et projection de $\mathsf{SS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (gauche) et $\mathsf{SS}_{\mathbf{z}}^{\mathsf{AH}}$ (droite) sur la composante PCA la plus corrélée (en bleu).

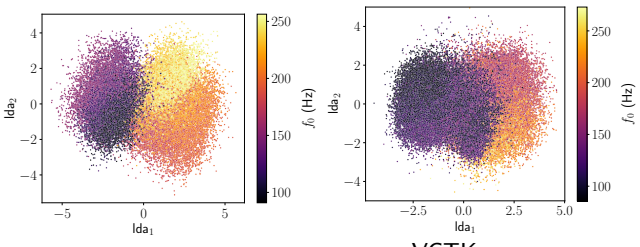


FIGURE 4 – Projection de $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (gauche) et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ (droite) sur les deux premières composantes de leurs lda respectifs, colorés en fonction de leurs valeurs f_0.

3.3 Modélisation des variations inter- vs. intra-genre

Nous avons observé que, pour nos deux modèles VAE, les valeurs f_0 de la parole des hommes et des femmes sont encodées dans des directions distinctes, mais non orthogonales de l'espace latent (voir Fig. 2). En revanche, pouvons-nous identifier des directions orthogonales qui modélisent

les variations inter- et intra-genre, afin de fournir une représentation latente mieux démêlée de la variabilité liée au genre ? Comme indiqué dans la Section 2.3, les candidats appropriés pour de telles directions sont la première et deuxième composante de lda, une LDA sur les locuteurs calculée sur l'espace latent de chaque VAE ($\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ et $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ pour VAE-VCTK et VAE-AH, respectivement). Pour valider cette hypothèse, nous avons mesuré la colinéarité des composantes de lda avec les DV des représentations de f_0 globale (m_{f_0}) et selon le genre ($\mathsf{m}_{f_0|\mathsf{M}}$ et $\mathsf{m}_{f_0|\mathsf{F}}$). Les CS mettent en évidence une forte corrélation (supérieure à 0.85 pour les deux VAE) entre la première composante lda et m_{f_0}, qui comprend des informations sur le genre. Par ailleurs, la deuxième composante lda est bien corrélée avec les DV calculées par genre, respectivement 0.68 avec $\mathsf{m}_{f_0|\mathsf{M}}$ et 0.61 avec $\mathsf{m}_{f_0|\mathsf{F}}$.

Ces résultats sont cohérents avec notre hypothèse selon laquelle les informations intra- et inter-genres sont encodées suivant des directions LDA distinctes. Pour illustrer cette analyse, la Fig. 4 représente la position des trames des signaux encodés $\mathsf{NS}_{\mathbf{z}}^{\mathsf{VCTK}}$ (à gauche) et de $\mathsf{NS}_{\mathbf{z}}^{\mathsf{AH}}$ (à droite) selon leurs deux premières composantes lda respectives. Nous observons que les trames sont regroupées en deux groupes le long de la première composante, les trames des locuteurs masculins (violet, f_0 faible) et féminins (orange-jaune, f_0 élevé) étant associées à des valeurs négatives et positives, respectivement. La seconde composante lda modélise la variation intra-genre de f_0. Dans l'ensemble, ces résultats mettent en évidence la capacité du VAE à démêler les variations inter- et intra-genre le long de deux directions distinctes de l'espace latent que nous avons identifiées grâce à l'analyse LDA, et ceci est observé pour deux ensembles de données de parole différents. Les faibles CS entre ces deux directions (< 0.35) sont prometteuses pour imaginer un contrôle indépendant de f_0 entre les classes de genre et à l'intérieur de celles-ci.

4 Conclusion

Nous avons introduit une méthodologie pour analyser l'espace latent du VAE entraîné sur une base de données multilocuteurs en combinant l'utilisation d'ensembles de données de test synthétiques et naturelles, et l'extraction et la comparaison des directions qui expliquent le mieux la variation des paramètres acoustiques sélectionnés. Après avoir montré que la variation de chaque paramètre est encodée par de multiples dimensions dans l'espace latent, nous avons démontré que l'une de ces dimensions encode la forme globale de la distribution des paramètres sur l'ensemble d'apprentissage. Dans le cas de f_0, la distribution est bimodale et les valeurs de f_0 appartenant à différents modes sont encodés sur des dimensions distinctes supplémentaires. Dans ce cas, nous avons identifié des directions dans l'espace latent du VAE qui expliquent les variations inter- et intra-genre de f_0. Pour valider notre approche, nous avons mené nos expériences sur deux ensembles de données (VCTK et Att-HACK) qui diffèrent en termes de langue (anglais vs. français) et de style (lecture/narration vs. jeu d'acteur/expression).

Alors que plusieurs études ont utilisé la réduction de la dimension de l'espace latent (Tits *et al.*, 2019; Dieck *et al.*, 2022), ont abordé l'orthogonalité des différentes directions qui expliquent un paramètre donné (Hsu *et al.*, 2017; Sadok *et al.*, 2023), ou identifié la variation de paramètres acoustiques dans l'espace latent par régression linéaire (Sadok *et al.*, 2023; Vaidya *et al.*, 2022; Lenglet *et al.*, 2022a), ce travail est l'un des rares à tenter d'interpréter la représentation multidimensionnelle de chaque paramètre acoustique unidimensionnel. Dans de futurs travaux, nous visons à augmenter le nombre de paramètres d'intérêt et à utiliser notre méthode pour contrôler la variation des paramètres acoustiques dans l'espace latent des modèles non- ou auto-supervisés.

Références

ANIKIN A. (2019). Soundgen : an open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, **51**, 778–792.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, p. 12449–12460, virtual conf.

BLAAUW M. & BONADA J. (2016). Modeling and transforming speech using variational autoencoders. In *Proc. of Interspeech*, p. 1770–1774, San Francisco, CA, USA.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5**(9-10), 341–347.

BOUS F. & ROEBEL A. (2022). A bottleneck auto-encoder for F0 transformations on speech and singing voice. *Information*, **13**(3), 102–121.

CHOU J.-C., YEH C.-C., LEE H.-Y. & LEE L.-S. (2018). Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In *Proc. of Interspeech*, p. 501–505, Hyderabad, India.

COLEMAN R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *J. of speech and hearing research*, **14**(3), 565–577.

DIECK T. T., PÉREZ-TORO P. A., ARIAS T., NOETH E. & KLUMPP P. (2022). Wav2vec behind the scenes : How end2end models learn phonetics. In *Proc. of Interspeech*, p. 5130–5134, Incheon, Korea.

FANT G. (1971). *Acoustic theory of speech production*. Mouton.

GIRIN L., ROCHE F., HUEBER T. & LEGLAIVE S. (2019). Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In *Int. Conf. on Digital Audio Effects*, p. 1–8, Birmingham, UK.

HIGGINS I., MATTHEY L., PAL A., BURGESS C., GLOROT X., BOTVINICK M., MOHAMED S. & LERCHNER A. (2017). β-VAE : Learning basic visual concepts with a constrained variational framework. In *Int. Conf. on Learning Representations*, Toulon, France.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *Trans. on Audio, Speech, and Language Processing*, **29**, 3451–3460.

HSU W.-N., ZHANG Y. & GLASS J. (2017). Learning latent representations for speech generation and transformation. In *Proc. of Interspeech*, p. 1273–1277, Stockholm, Sweden.

JACQUELIN M., GARNIER M., GIRIN L., VINCENT R. & PERROTIN O. (2024). Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models. In *ICASSP Workshop XAI-SA*, Seoul, Korea.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In *Int. Conf. on Learning Representations*, San Diego, USA.

KINGMA D. P. & WELLING M. (2014). Auto-encoding variational bayes. In *Int. Conf. on Learning Representations*, Banff, Canada.

LAKHOTIA K., KHARITONOV E., HSU W.-N., ADI Y., POLYAK A., BOLTE B., NGUYEN T.-A., COPET J., BAEVSKI A., MOHAMED A. & DUPOUX E. (2021). On generative spoken language modeling from raw audio. *Trans. of the Association for Computational Linguistics*, **9**, 1336–1354.

LE MOINE C. & OBIN N. (2020). Att-hack : An expressive speech database with social attitudes. In *Speech Prosody*.

LENGLET M., PERROTIN O. & BAILLY G. (2022a). Modélisation de la parole avec tacotron2 : Analyse acoustique et phonétique des plongements de caractere. In *Journées d'Études sur la Parole*, p. 788–796, Noirmoutier, France.

LENGLET M., PERROTIN O. & BAILLY G. (2022b). Speaking rate control of end-to-end TTS models by direct manipulation of the encoder's output embeddings. In *Proc. of Interspeech*, p. 11–15, Incheon, Korea.

POLYAK A., ADI Y., COPET J., KHARITONOV E., LAKHOTIA K., HSU W.-N., MOHAMED A. & DUPOUX E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. In *Proc. of Interspeech*, p. 3615–3619, Brno, Czechia.

SADOK S., LEGLAIVE S., GIRIN L., ALAMEDA-PINEDA X. & SÉGUIER R. (2023). Learning and controlling the source-filter representation of speech with a variational autoencoder. *Speech Comm.*, **148**, 53–65.

SUNDBERG J. & NORDENBERG M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The J. of the Acoust. Soc. of Am.*, **120**(1), 453–457.

TITS N., WANG F., EL HADDAD K., PAGEL V. & DUTOIT T. (2019). Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. In *Proc. of Interspeech*, p. 4475–4479, Graz, Austria.

TITZE I. R. (2004). A theoretical study of F0-F1 interaction with application to resonant speaking and singing voice. *J. of Voice*, **18**(3), 292–298.

VAIDYA A. R., JAIN S. & HUTH A. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. In *Int. Conf. on Machine Learning*, p. 21927–21944, Baltimore, USA.

VAN DER OORD A., VINYALS O. *et al.* (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, p. 6306–6315, Long Beach, USA.

WANG Y., STANTON D., ZHANG Y., RYAN R.-S., BATTENBERG E., SHOR J., XIAO Y., JIA Y., REN F. & SAUROUS R. A. (2018). Style tokens : Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Int. Conf. on Machine Learning*, p. 5180–5189, Stockholm, Sweden.

WILLIAMS J., FONG J., COOPER E. & YAMAGISHI J. (2021). Exploring disentanglement with multilingual and monolingual VQ-VAE. In *ISCA Speech Synthesis Workshop*, p. 124–129, Budapest, Hungary.

WILLIAMS J. & KING S. (2019). Disentangling style factors from speaker representations. In *Proc. of Interspeech*, p. 3945–3949, Graz, Austria.

YAMAGISHI J., VEAUX C. & MACDONALD K. (2019). CSTR VCTK corpus : English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

ZHANG Y.-J., PAN S., HE L. & LING Z.-H. (2019). Learning latent representations for style control and transfer in end-to-end speech synthesis. In *Int. Conf. on Acoustics, Speech and Signal Processing*, p. 6945–6949, Brighton, UK.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Identification du locuteur : ouvrir la boîte noire

Carole Millot[1, 2*] Cédric Gendrot[2] Jean-François Bonastre[1, 3]
(1) Inria, Domaine de Voluceau, 78153 Le Chesnay, France
(2) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 rue des Irlandais, 75005 Paris, France
(3) Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, Avignon, France
{carole.millot, cedric.gendrot}@sorbonne-nouvelle.fr,
jean-francois.bonastre@inria.fr

RÉSUMÉ

L'explicabilité des systèmes relevant du *deep learning* est devenue un enjeu central ces dernières années, dans le droit européen comme le domaine criminalistique. L'approche BA-LR introduit en identification du locuteur un nouveau paradigme de modélisation : elle fait émerger automatiquement les attributs partagés par un groupe de locuteurs et qui sous-entendent la discrimination de ceux-ci. Le score produit est décomposable au niveau des attributs, ce qui augmente significativement l'explicabilité de la méthode. Cette étude propose de compléter la caractérisation des attributs obtenus par le BA-LR, à l'aide de paramètres de qualité de voix. L'analyse suggère que plusieurs attributs utilisent les types de phonation pour regrouper les locuteurs, ceux-ci encodant des informations humainement perceptibles. Cet article pose ainsi des bases pour l'analyse acoustique des attributs, qui permettra à terme d'utiliser le BA-LR dans le cadre du profilage vocal.

ABSTRACT

Opening the black box in speaker recognition

Explaining how deep-learning-based systems work has recently become a central issue, as seen in European legislation and forensic research. The BA-LR approach introduces a new paradigm for speaker recognition, discriminating speakers by bringing out binary attributes shared between them. The obtained score can be broken down at the attribute level, augmenting significantly the BA-LR explainability. This study aims to characterize the attributes proposed by BA-LR, with the help of voice quality parameters. Analysis suggests that multiple attributes use phonation types to group speakers : this shows attributes can be phonetically characterized, and encode humanly perceptible informations. This paper lays foundations for acoustic analysis of binary attributes, which may eventually permit using BA-LR for voice profiling.

MOTS-CLÉS : qualité de voix, traitement automatique de la parole, explicabilité, perception de la parole, reconnaissance du locuteur.

KEYWORDS: voice quality, speech processing, explainability, speech perception, speaker recognition.

∗. Thèse co-financée par la Direction générale de l'armement à travers l'Agence de l'innovation de défense.

1 Introduction

La reconnaissance automatique du locuteur consiste à reconnaître ou vérifier l'identité d'une personne à partir d'un échantillon de sa voix. La comparaison de voix s'inscrit dans ce champ et détermine si deux enregistrements de parole ont été produits par le même locuteur, ou deux locuteurs différents. Les systèmes *state of the art* (état de l'art) de reconnaissance du locuteur sont basés sur l'apprentissage d'un modèle *deep learning* (apprentissage profond), appris sur de grandes bases données de locuteurs (Bai & Zhang (2021), Kwon *et al.* (2021)). Leurs performances sont excellentes (Sarni *et al.*, 2023), mais ils ne fournissent aucun élément d'information permettant d'expliquer leur score (Campbell *et al.*, 2009). L'explicabilité est pourtant un enjeu central pour la vérification du locuteur, par exemple dans une optique criminalistique pour la vérification du locuteur (Ben Amor & Bonastre (2022b), Ben Amor *et al.* (2023)) ou de manière générale, pour toutes les activités dîtes « *High Risk* » dans le cadre de l'*AI Act* (Sovrano *et al.*, 2022). En réponse à cette limitation, l'approche BA-LR, a été récemment proposée (Ben Amor & Bonastre, 2022a). Elle représente un enregistrement audio par la présence ou l'absence d'attributs de voix dans celui-ci. Les attributs sont issus d'un ensemble fermé déterminé automatiquement (*bottom-up*) à partir d'une approche de *deep learning* appliquée sur une base de données de plus d'un million d'enregistrements. BA-LR propose comme score un Rapport de Vraisemblance (LR) entre la probabilité pour qu'un seul locuteur ait prononcé les deux enregistrements, versus l'hypothèse inverse. Ce score n'est basé que sur la présence (activation) ou l'absence des attributs dans les deux fichiers et sur les caractéristiques des attributs. Ce paradigme favorise l'explicabilité intrinsèque car la participation de chaque attribut à la décision est connue et est issue des caractéristiques de celui-ci, apprises durant l'entraînement (rareté et fiabilité d'extraction).

Caractériser la nature des informations encodées par ces attributs découverts par un système automatique est important dans le cadre de cette démarche. Nous conjecturons que la qualité de voix fait partie des paramètres pris en compte par le système pour définir les attributs. En effet, Kreiman (Kreiman *et al.* (2003), Lee & Kreiman (2019)) définit la qualité de voix comme la façon « dont les locuteurs projettent leur identité — leurs caractéristiques physiques, psychologiques, et sociales — au monde ». La qualité de voix peut être décomposée en différents corrélats acoustiques et perceptuels (Barsties & De Bodt, 2015), et est liée à des paramètres linguistiques tels que la nasalité ou encore le type de phonation (Lee & Kreiman, 2022).
Les types de phonation qualifient les différents positionnements possibles de la glotte pendant la phonation. On y compte la voix modale, mais aussi les voix craquée (présence de vibrations voisées irrégulières) et soufflée (présence importante de bruit dans le signal), ainsi que tendue et relâchée (Gordon & Ladefoged, 2001). Les types de phonation peuvent être liés à des variations d'ordres divers : le sexe est un facteur important — on retrouve souvent plus de souffle dans la voix des femmes du fait de la fermeture incomplète (*glottal chink*) de leur plis vocaux (Hanson & Chuang, 1999). La langue parlée par un locuteur (Benoist-Lucy & Pillot-Loiseau, 2013) et son appartenance à une communauté sociale ou géographique sont d'autres influences impactant le type de phonation, comme le cas de jeunes femmes étasuniennes utilisant la voix craquée (Greer & Winters, 2015).

Cette étude s'axe autour de deux enjeux. Le premier est la caractérisation d'attributs de la voix discriminants au sens du locuteur, par des paramètres de la qualité de la voix, ici les types de phonation. Le second est le développement d'une méthodologie cohérente pour l'étude des attributs découverts par un processus automatique.
La corrélation entre les différents attributs extraits par le BA-LR et les types de phonation est étudiée, suivie d'une analyse révélant les paramètres acoustiques pris en compte par le système automatique. Les liens avec d'autres attributs et le sexe des locuteurs sont également analysés.

2 Méthode

2.1 Annotation du corpus

Afin d'extraire les attributs de chaque extrait de parole, le système automatique BA-LR « standard » (Ben Amor & Bonastre, 2022a) est utilisé. Le modèle de celui-ci a été appris à partir de données du corpus anglophone VoxCeleb2 (Nagrani *et al.*, 2017), une base de plus d'un million d'enregistrements produits par plus de 6000 locuteurs.
L'étude est réalisée à partir du corpus francophone PTSVOX (Chanclu *et al.*, 2020), sélectionné en raison de son nombre de locuteurs, 369, permettant d'avoir un vaste éventail de profils vocaux. Chaque locuteur est enregistré pendant deux à quatre minutes, avec une à quatre sessions par locuteur (minimum deux pour les 24 premiers). Comme évoqué précédemment, le modèle BA-LR a été appris sur de l'anglais et est appliqué sur des données francophones. Bien que constituant une limitation potentielle, cette utilisation est justifiée par plusieurs expériences précédentes.
Dans le cadre de ce travail, le corpus a été annoté selon les types de phonation présents dans les enregistrements pour établir les profils vocaux des locuteurs. Pour chaque enregistrement, l'annotateur [1] a attribué une étiquette de profil (craqué, soufflé, modal) perceptuellement pour chaque locuteur. Pour attribuer une étiquette craquée ou soufflée, le type de phonation doit être présent pendant environ les deux tiers des données du locuteur. Un groupe témoin composé d'extraits de 100 locuteurs sélectionnés aléatoirement est également composé.
Dans un deuxième temps, les enregistrements sont décomposés en extraits de trois secondes à l'aide d'un script Praat (Boersma, 2001) qui supprime les pauses dans l'enregistrement. Chaque extrait est ré-annoté selon le type de phonation présent dans celui-ci. Un accord inter-annotateurs est calculé à l'aide d'un autre évaluateur sur 10% des données à l'aide du *package* `psych` dans R (R Core Team, 2023; Revelle, 2024), dont le Kappa de Cohen résultant est de $\kappa = 0,79$. Les extraits présentant le type de phonation renseigné dans le profil du locuteur sont alors sélectionnés afin de réduire le bruit dans les données et la variation intra-locuteur.
La Table 1 montre la répartition des locuteurs en groupes de types de phonation, le nombre d'extraits vocaux par groupe, ainsi que le pourcentage de femmes pour chaque groupe.

Profil vocal du locuteur	Nombre de locuteurs	Nombre d'extraits	% femmes
Profil craqué	38 locuteurs	4227 extraits	16%
Profil soufflé	31 locuteurs	3763 extraits	58%
Profil modal	32 locuteurs	4510 extraits	40%
Groupe témoin	100 locuteurs	8332 extraits	40%
Total	201 locuteurs	20832 extraits	

TABLE 1 – Tableau montrant la répartition des locuteurs en groupes de type de phonation. Le nombre d'extraits correspondants et le pourcentage de femmes dans chaque groupe sont également présentés.

2.2 Extraction des attributs

Le système BA-LR, dérivé des x-vecteurs, représente un extrait de parole (de trois secondes ici) par un *embedding* neuronal de 256 coefficients, ensuite binarisés. Seuls 206 coefficients sont conservés après suppression des coefficients inactifs. Un 1 indique la présence d'un attribut, un 0 son absence.

1. L'annotation a été réalisée par Carole Millot.

3 Résultats

Afin de comparer le comportement des attributs en fonction des groupes de types de phonation, des moyennes d'activation pour chaque attribut sont calculées par groupe de locuteurs. Le groupe témoin permet d'établir des taux d'activation étalons pour chaque attribut. Un aperçu de l'ensemble des taux d'activation des attributs pour chaque groupe de locuteurs est visible Figure 1.

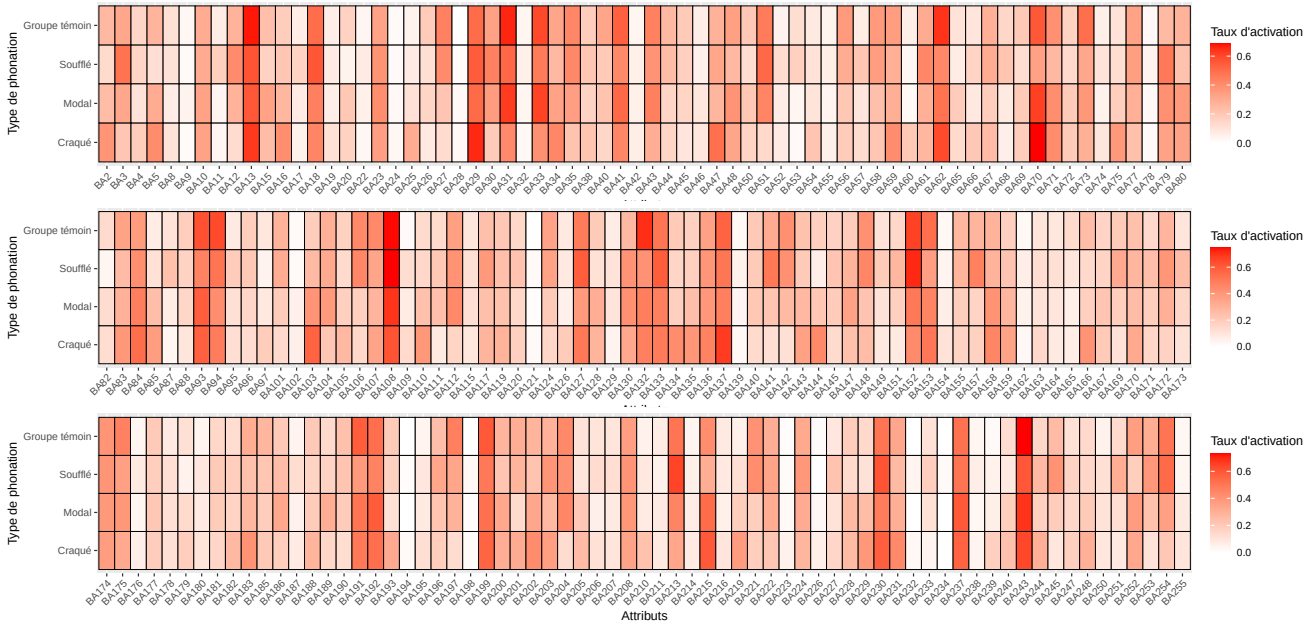


FIGURE 1 – Taux d'activation moyen des 206 attributs pour chaque groupe de locuteurs.

Les moyennes d'activation de chaque attribut pour les trois groupes de type de phonation sont ensuite divisées par la moyenne d'activation du groupe témoin. Les attributs avec les plus forts écarts entre les ratios pour le groupe craqué/témoin et le groupe soufflé/témoin sont sélectionnés pour la suite de cette étude. Afin d'établir un seuil, un critère perceptuel est appliqué : pour les attributs avec les différences de ratio les plus importantes, une écoute de chaque extrait est réalisée pour vérifier si le type de phonation (voix craquée ou de la voix soufflée) est perçu dans l'ensemble des extraits du groupe, afin d'établir un lien perceptuel entre attribut et type de phonation. Ainsi, seuls les huit attributs comportant les différences de ratio les plus importantes (au moins neuf dixièmes de points) sont retenus pour la suite de l'analyse : BA3, 5, 27, 51, 85, 141, 144 et 157. Ils sont renseignés dans la Table 2.

Attribut	Ratio craqué/témoin	Ratio modal/témoin	Ratio soufflé/témoin
BA3	$0,65$	1	$1,55$
BA5	$1,53$	$1,12$	$0,42$
BA27	$0,48$	$0,78$	$1,48$
BA51	$0,33$	1	$1,21$
BA85	$4,63$	$3,13$	$1,50$
BA141	$0,41$	$0,72$	$1,56$
BA144	$2,50$	$1,50$	$0,56$
BA157	$0,62$	$0,59$	$1,66$

TABLE 2 – Tableau représentant les huit attributs sélectionnés pour l'analyse, et le ratio entre les taux d'activation moyens pour les groupes de type de phonation et le groupe témoin.

3.1 Interactions par attribut

Pour chacun des extraits de trois secondes, des mesures acoustiques sont extraites par openSMILE via l'ensemble de paramètres eGeMAPS (Eyben *et al.*, 2010) qui contient 88 paramètres acoustiques. La mesure de la fréquence fondamentale f_0 en demi-tons, le Ratio Harmoniques/Bruit (HNR), ou encore la différence $h_1 - h_2$ font partie de ces paramètres. $h_1 - h_2$ est utilisée pour évaluer le type de phonation d'un extrait et calcule la différence entre la première et la seconde harmonique d'un spectre (Keating *et al.*, 2010). Le HNR permet d'estimer le taux de bruit dans l'extrait, élevé pour les voix craquées et soufflées, par rapport au taux d'harmoniques (Davidson, 2019). Ces mesures sont exploitées à l'aide du *package* `lme4` (Bates *et al.*, 2015) afin de construire les modèles mixtes des interactions entre taux d'activation des attributs et mesures acoustiques. Les modèles mixtes permettent de contrôler la significativité des interactions calculées dans de grands corpus, grâce à l'inclusion d'effets aléatoires et de prédicteurs [2]. Les résultats sont regroupés dans la Table 3 et décrits dans les paragraphes ci-après. L'interaction est considérée significative à partir de $p < 0.005$ **.

	p-value			
Attribut	**Sexe**	f_0	$h_1 - h_2$	**HNR**
BA3	0.043 *	0.101	0.153	0.064
BA5	0.413	0.102	0.000259 ***	0.0015 **
BA27	0.031 *	0.025 *	0.882	0.000483 ***
BA51	0.362	0.164	0.000207 ***	2e-16 ***
BA85	0.039 *	0.561	0.00017 ***	1.05e-07 ***
BA141	6.31e-05 ***	6.31e-05 ***	0.319	0.129
BA144	6.31e-05 ***	6.31e-05 ***	0.055	0.052
BA157	6.31e-05 ***	6.31e-05 ***	0.056	0.021 *
206 attributs	0.810	0.085	0.975	0.020 *

TABLE 3 – Tableau présentant les p-values obtenues à partir des modèles mixtes calculés pour chacun des huit attributs étudiés. La significativité des interactions est évaluée avec le *package* `lmerTest` (Kuznetsova *et al.*, 2017) à l'aide de l'approximation de Satterthwaite.

Le sexe est un paramètre pris en compte de manière explicite ou implicite par les systèmes de reconnaissance de locuteurs du fait de son haut potentiel de discrimination (Jacquelin *et al.*, 2023), ce qui peut introduire un biais dans l'analyse présentée. De plus, les groupes de locuteurs étudiés ne contiennent pas le même ratio d'hommes et de femmes (Table 1). Afin de vérifier que l'information retenue par les huit attributs ne dépend pas du sexe, un modèle mixte est utilisé pour calculer l'interaction entre les attributs et le sexe, avec le type de phonation en prédicteur fixe et le locuteur en variable aléatoire. L'interaction entre les huit attributs et le sexe est significative avec $p < 0.001$ pour trois attributs (Table 3) : BA141, BA144 et BA157.

La fréquence fondamentale est le paramètre acoustique le plus proéminent pour la prédiction du sexe d'un locuteur (Jacquelin *et al.*, 2023). L'interaction entre la f_0 (en demi-tons) et les trois attributs corrélés au sexe du locuteur est calculée, avec le type de phonation et le sexe en prédicteurs fixes et le locuteur en variable aléatoire. Pour les trois attributs ayant une interaction significative avec le sexe du locuteur, les interactions avec la f_0 sont également significatives ($p < 0.001$), voir Table 3.

Pour vérifier les informations utilisées par les cinq autres attributs, un modèle mixte est utilisé pour calculer l'interaction entre les attributs et $h_1 - h_2$, avec le sexe et le type de phonation des extraits

2. Une interaction significative entre deux variables indique l'influence d'une variable sur l'effet de la deuxième.

en prédicteurs fixes, et le locuteur en variable aléatoire. L'interaction entre BA5 et $h_1 - h_2$ est significative ($p < 0.001$), ainsi que pour BA51 et BA85 (voir Table 3). Cependant, pour BA3 et BA27 il n'y a pas d'interaction significative ($p > 0.05$).

Enfin, un modèle mixte similaire est utilisé pour les attributs et le Ratio Harmoniques/Bruit en variables. L'interaction entre les attributs et la mesure est significative ($p < 0.001$) pour BA27, BA51 et BA85, ainsi que pour BA5 ($p < 0.005$), voir Table 3.

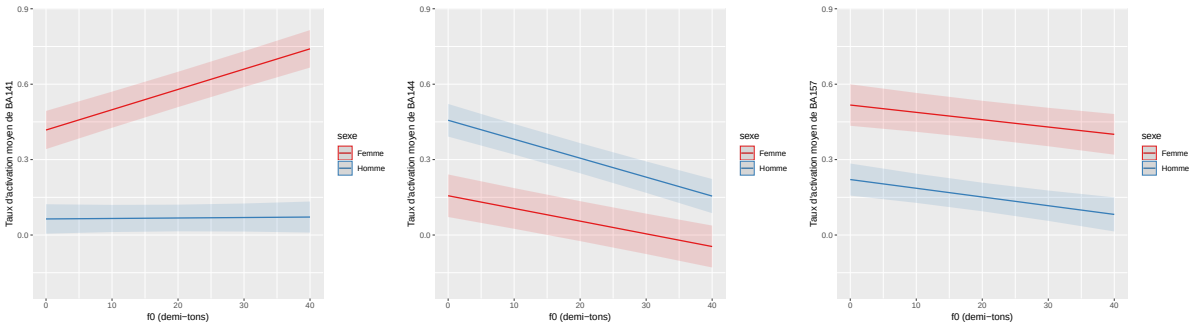
Les interactions calculées précédemment sont comparées avec celles obtenues entre les moyennes combinées des activations de l'entièreté des attributs (206) et les différentes mesures (sexe, f_0, $h_1 - h_2$, HNR), en conservant les paramètres des modèles mixtes. Cette comparaison permet de vérifier si les sept attributs analysés ci-avant contribuent significativement à l'encodage du sexe ou du type de phonation par le système. Les interactions utilisant les 206 attributs ne sont pas significatives pour le sexe, la f_0 et $h_1 - h_2$ ($p > 0.05$). L'interaction avec le HNR est significative ($p = 0.020$).

La vérification des informations retenues par les attributs a apporté les informations suivantes : le type de phonation est encodé par quatre attributs, pour lesquels des interactions significatives ont été relevées pour les mesures $h_1 - h_2$ et le Ratio Harmoniques/Bruit. Trois autres attributs encodent le sexe du locuteur, présentant une interaction significative avec la f_0.
BA3 a un comportement atypique : bien qu'il ait été sélectionné pour sa différence de taux d'activation moyens entre les groupes craqué et soufflé, il ne présente pas d'interaction significative avec la fréquence fondamentale ou le type de phonation. La présence d'autres biais tels que l'âge, des pathologies de la voix ou l'accent régional peuvent expliquer son comportement.

3.2 Interactions entre attributs

D'après les résultats précédents, le type de phonation et le sexe font partie de l'information retenue par sept attributs analysés. Certains attributs ont leurs taux d'activation moyens en interaction significative avec une mesure acoustique commune, comme BA141, 144 et 157 avec f_0. Cette sous-section détaille les différentes conditions d'activation des attributs en interaction avec une mesure acoustique commune.

La Figure 2 représente la régression logistique entre les trois attributs et la f_0 : on observe qu'ils n'encodent pas tous la même information concernant la f_0 . BA141 est activé pour les femmes à la voix aiguë, BA144 est activé pour les hommes à la voix grave, et BA157 est activé pour les femmes à la voix grave. Ces attributs permettent au système d'identifier les principaux contrastes de hauteur de voix selon le sexe : femme à la voix aiguë (BA141 activé), homme à la voix grave (BA144 activé), femme à la voix grave (BA157 activé) et homme à la voix aiguë (aucun des trois attributs activés).



(a) (b) (c)

FIGURE 2 – Modélisation de l'interaction significative entre BA141 (a), BA144 (b) et BA157 (c), et la f_0 en demi-tons et le sexe du locuteur dans chaque extrait, avec le type de phonation en prédicteur fixe et le locuteur en variable aléatoire.

L'écoute des extraits qui activent les quatre attributs liés au type de phonation permet d'obtenir des renseignements supplémentaires sur leurs conditions d'activation.
Par exemple, 80% des extraits activés pour BA85 contiennent des voyelles allongées qui peuvent provoquer le craquement, comme les disfluences verbales. La durée d'allongement de la voyelle requise pour l'activation de l'attribut est de minimum 500ms, c'est-à-dire un sixième de l'extrait. Un exemple avec l'hésitation « euh » est visible sur la Figure 3, où l'on constate l'irrégularité du signal et du spectrogramme lors de sa production. Les extraits activés pour BA5 ont en commun la présence de voix craquée pendant au moins 700ms, et ce peu importe les phonèmes présents.
BA51 est présent lors de la présence de souffle ou d'écho dans l'enregistrement, notamment lorsque des locuteurs parlent près du micro. Enfin, BA27 peut être activé lors de la production de voix dîte « peu efficace », pendant laquelle les locuteurs sont souvent à court de souffle.

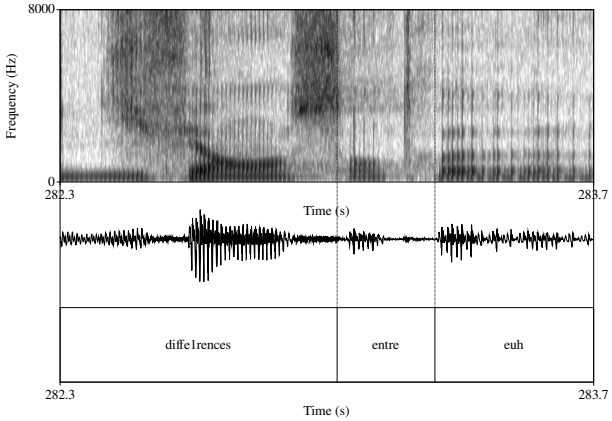


FIGURE 3 – Exemple de la disfluence craquée « euh » présente dans un extrait activant BA85 chez le locuteur LL054.

3.3 Encodage de la prototypicalité de sexe

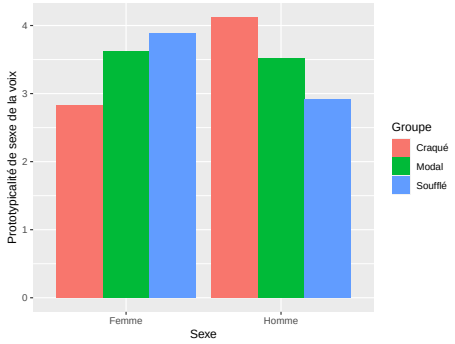
Le type de phonation produit par un locuteur est influencé par son sexe (voir Section 1). Il a été montré en Section 3 que les interactions avec le sexe ne sont pas significatives pour les attributs BA5, BA27, BA51 et BA85 ($p > 0.005$). On étudie dans cette sous-section l'hypothèse suivante : l'activation des attributs BA5, BA27, BA51 et BA85 est corrélée aux caractéristiques perceptuelles des voix prototypiques d'un sexe.
Une voix perceptuellement prototypique d'un sexe possède des caractéristiques associées stéréotypiquement à ce sexe. Cela est dû à des différences biologiques récurrentes entre les deux sexes, qui sont intériorisées par les auditeurs et sont utilisées lors de la perception d'une voix (Latinus & Belin, 2011). Par exemple, des voix prototypiques féminines peuvent comporter un type de phonation soufflé et des contours intonatifs importants (Avery & Liss, 1996), tandis que des voix prototypiques masculines peuvent comporter un type de phonation tendu et une f_0 basse (Baumann & Belin, 2010).
Si cette hypothèse est vérifiée, alors une voix prototypiquement féminine active plus fréquemment les attributs liés à la voix soufflée, et inversement pour une voix prototypiquement masculine.

Une annotation perceptuelle en prototypicalité de voix selon le sexe est effectuée pour les locuteurs des différents groupes de types de phonation. L'évaluation se fait sur chaque enregistrement des différents groupes, à l'aide une échelle de 1 à 5 avec 1 la note la plus basse (locuteur à la voix non prototypique) et 5 la note la plus haute (locuteur à la voix prototypique). L'accord inter-annotateur est calculé selon le Kappa de Cohen, dont le résultat est $\kappa = 0,60$. Ainsi, pour prendre en compte les avis de chaque annotateur, l'annotation finale est la moyenne des annotations précédentes [3]. La Figure 4a

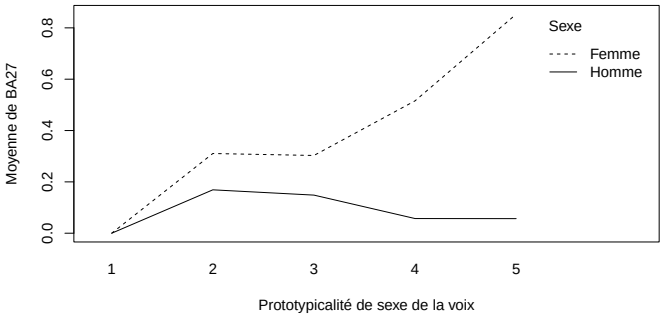
3. Les annotations ont été réalisées par les deux premiers auteurs, en se basant sur l'ampleur des contours intonatifs, l'intensité de la voix et la hauteur de la f_0.

montre la répartition des voix prototypiquement féminines et masculines d'après l'annotation finale, selon le groupe de locuteurs. Un calcul de l'interaction entre les scores de prototypicalité et la f_0 est effectué à l'aide d'un modèle mixte. L'interaction est significative ($p < 0,001$).

L'interaction entre les quatre attributs et l'annotation en prototypicalité de sexe est calculée avec un modèle mixte, dont le prédicteur fixe est le sexe et la variable aléatoire est le type de phonation. Le modèle montre que BA27 est le seul attribut possédant une interaction significative ($p < 0,001$). La Figure 4b représente l'interaction entre BA27 et la prototypicalité de sexe des voix avec le sexe comme prédicteur fixe. Le taux d'activation moyen de BA27 croît selon le score de prototypicalité féminine (80% d'activations pour une note de 5, aucune activation pour une note de 1). BA27 encode donc le principe de prototypicalité de la voix féminine. Les trois autres attributs ne retiennent pas cette information ($p > 0,05$).



(a) Diagramme en barres de l'annotation des locuteurs selon la prototypicalité de leur voix en termes de sexe.



(b) Interaction entre la prototypicalité de sexe des voix des locuteurs et le taux d'activation moyen de l'attribut BA27, avec le sexe comme facteur.

FIGURE 4 – Résultats de l'analyse des attributs selon la prototypicalité de sexe des voix des locuteurs.

4 Conclusion

Les résultats de cette étude montrent que plusieurs attributs du BA-LR sont corrélés à des paramètres de qualité de voix, ici les types de phonation craqué et soufflé (Section 3). Le sexe est également une information discriminante pour trois des huit attributs étudiés, ainsi que la prototypicalité homme/femme des voix pour un d'entre eux (Sous-section 3.3).
Les attributs interagissent entre eux, et de nombreux paramètres sont à prendre en compte afin de comprendre leurs conditions d'activation. Ces résultats encouragent à procéder à l'annotation d'autres caractéristiques perceptibles dans les enregistrements, notamment au niveau de la qualité de voix, afin de caractériser d'autres attributs. Utiliser le BA-LR sur un autre corpus, multi-sessions, afin d'en comparer les résultats avec PTSVOX, est la prochaine piste d'étude.

La méthodologie suivie dans cet article a permis d'établir une interaction entre des attributs extraits à partir d'un système automatique et des paramètres de qualité de voix. Cette démarche confère une plus grande explicabilité au système, utile dans un cadre judiciaire.
La meilleure compréhension des attributs sous un angle perceptuel permet à la fois d'évaluer la proximité entre la perception humaine et le réseau de neurones derrière le BA-LR, mais aussi de documenter les paramètres de qualité de voix utilisés par cet outil : cela le rend utilisable dans des tâches de profilage du locuteur, la qualité de voix pouvant apporter des renseignements physiques et culturelles sur le locuteur étudié. Cela permettrait à terme l'extraction automatique d'un profil de locuteur à partir d'un simple enregistrement sonore.

Références

AVERY J. D. & LISS J. M. (1996). Acoustic characteristics of less-masculine-sounding male speech. *The Journal of the Acoustical Society of America*, **99**(6), 3738–3748. DOI : 10.1121/1.414970.

BAI Z. & ZHANG X.-L. (2021). Speaker recognition based on deep learning : An overview. *Neural Networks*, **140**, 65–99.

BARSTIES B. & DE BODT M. (2015). Assessment of voice quality : current state-of-the-art. *Auris Nasus Larynx*, **42**(3), 183–188. DOI : 10.1016/j.anl.2014.11.001.

BATES D., M
ÄCHLER M., BOLKER B. & WALKER S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. DOI : 10.18637/jss.v067.i01.

BAUMANN O. & BELIN P. (2010). Perceptual scaling of voice identity : common dimensions for different vowels and speakers. *Psychological Research PRPF*, **74**(1), 110–120. DOI : 10.1007/s00426-008-0185-z.

BEN AMOR I. & BONASTRE J.-F. (2022a). Ba-lr : Binary-attribute-based likelihood ratio estimation for forensic voice comparison. In *2022 International workshop on biometrics and forensics (IWBF)*, p. 1–6 : IEEE. DOI : 10.1109/IWBF55382.2022.9794542.

BEN AMOR I. & BONASTRE J.-F. (2022b). Ba-lr : une approche transparente de comparaison de voix en criminalistique. In *Actes de la 7e conférence conjointe Journées d'Études sur la Parole (JEP, 34e édition), Traitement Automatique des Langues Naturelles (TALN, 29e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 24e édition). Volume 1 : Journées d'Études sur la Parole*, p. 646–654. DOI : 10.21437/JEP.2022-68.

BEN AMOR I., BONASTRE J.-F., O'BRIEN B. & BOUSQUET P.-M. (2023). Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition. In *Proceedings of Interspeech 2023*. HAL : hal-04155146.

BENOIST-LUCY A. & PILLOT-LOISEAU C. (2013). The influence of language and speech task upon creaky voice use among six young american women learning french. In *Proceedings of Interspeech 2013*, p. 2395–2399. HAL : hal-00862349.

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

CAMPBELL J. P., SHEN W., CAMPBELL W. M., SCHWARTZ R., BONASTRE J.-F. & MATROUF D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, **26**(2), 95–103.

CHANCLU A., GEORGETON L., FREDOUILLE C. & BONASTRE J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire. In *Actes de la 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*, p. 73–81. HAL : hal-02798519.

DAVIDSON L. (2019). Perceptual coherence of creaky voice qualities. In *Proceedings of the 19th International Congress of Phonetic Sciences. Canberra, Australia : Australasian Speech Science and Technology Association Inc*, p. 147–151.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). opensmile : the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, p. 1459–1462.

GORDON M. & LADEFOGED P. (2001). Phonation types : a cross-linguistic overview. *Journal of phonetics*, **29**(4), 383–406. DOI : 10.006/jpho.2001.0147.

GREER S. D. & WINTERS S. J. (2015). The perception of coolness : Differences in evaluating voice quality in male and female speakers. In *Proceedings of ICPhS 2015*.

HANSON H. & CHUANG E. (1999). Glottal characteristics of male speakers : Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, **106**, 1064–77. DOI : 10.1121/1.427116.

JACQUELIN M., GARNIER M., GIRIN L., VINCENT R. & PERROTIN O. (2023). Exploring the multidimensional representation of individual speech acoustic parameters extracted by deep unsupervised models. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, p. 240–241. HAL : hal-04274170.

KEATING P., ESPOSITO C., GARELLEK M., KHAN S. & KUANG J. (2010). Phonation contrasts across languages. In *Poster presented at the 12th Conference on Laboratory Phonology*.

KREIMAN J., VANLANCKER-SIDTIS D. & GERRATT B. R. (2003). Defining and measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality : Functions, Analysis and Synthesis*.

KUZNETSOVA A., BROCKHOFF P. B. & CHRISTENSEN R. H. B. (2017). lmerTest package : Tests in linear mixed effects models. *Journal of Statistical Software*, **82**(13), 1–26. DOI : 10.18637/jss.v082.i13.

KWON Y., HEO H.-S., LEE B.-J. & CHUNG J. S. (2021). The ins and outs of speaker recognition : lessons from voxsrc 2020. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5809–5813 : IEEE.

LATINUS M. & BELIN P. (2011). Human voice perception. *Current Biology*, **21**(4), R143–R145.

LEE Y. & KREIMAN J. (2019). Within and between speaker variation in voices. In *International Congress of Phonetic Sciences*, p. 1460–1464.

LEE Y. & KREIMAN J. (2022). Linguistic versus biological factors governing acoustic voice variation. In *Proceedings of Interspeech 2022*, p. 640–643.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : a large-scale speaker identification dataset. In *Proceedings of Interspeech 2017*, p. 2616–2620.

R CORE TEAM (2023). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 4.3.2 (2023-10-31).

REVELLE W. (2024). *psych : Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.4.1.

SARNI S., CUMANI S., SINISCALCHI S. M., BOTTINO A. *et al.* (2023). Description and analysis of the kpt system for nist language recognition evaluation 2022. In *Proceedings of 24th INTERSPEECH Conference*, p. 1–5 : ISCA.

SOVRANO F., SAPIENZA S., PALMIRANI M. & VITALI F. (2022). Metrics, explainability and the european ai act proposal. *J*, **5**(1), 126–138.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Les représentations de locuteurs pour prédire l'intelligibilité de la parole lors de conversations médicales

Sebastião Quintas[1] Mathieu Balaguer[1, 2] Julie Mauclair[1] Virginie Woisard[2, 3]
Julien Pinquier[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) Hôpital Larrey, Toulouse, France
(3) Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France
sebastiao.quintas@irit.fr

RÉSUMÉ

Dans le contexte des troubles de la parole, l'une des tâches du thérapeute est de définir l'intelligibilité de la parole du patient. Les systèmes automatiques peuvent aider dans cette tâche, mais dans la plupart des cas, ils sont entraînés dans des environnements spécifiques et contrôlés, avec des conditions propres qui ne reflètent pas un environnement médical. Dans cet article, nous développons un système automatique qui prédit l'intelligibilité de la parole à partir de données provenant de patients ayant un cancer de la tête et du cou obtenues dans des conditions cliniques. Ce système repose sur des représentations de locuteurs entraînées selon une méthodologie multi-tâches pour prédire simultanément l'intelligibilité de la parole et la sévérité des troubles de la parole. Il atteint une corrélation allant jusqu'à 0,891 pour une tâche de lecture. De plus, il affiche des résultats prometteurs sur de la parole spontanée, qui est une tâche plus écologique mais sous-étudiée et pourtant essentielle pour un déploiement direct d'un système automatique dans un environnement hospitalier.

ABSTRACT

Speaker Embeddings to Predict Speech Intelligibility in Medical Conversations

In the context of speech disorders, one of the therapist task is to asses the speech intelligibility of a patient. Automatic systems can help in that task but in most cases, they are trained in specific controlled environments with clean conditions that do not reflect a healthcare environment. In this paper, we develop an automatic system that predict speech intelligibility on head and neck cancer data obtained in clinical conditions. This system relies on speaker embeddings trained using a multi-task methodology to simultaneous predict speech intelligibility and speech disorder severity. It achieves a correlation up to 0.891 on a reading task. Morever, it display promosing results on spontaneous speech, which is a more ecologic task yet understudied but nevertheless essential for a direct deployment in a hospital setting.

MOTS-CLÉS : Intelligibilité de la parole, traitement automatique de la parole, représentations de locuteur, cancer de la tête et du cou, parole spontanée.

KEYWORDS: speech intelligibility, automatic speech processing, speaker embeddings, head and neck cancer, spontaneous speech.

Actes des 35èmes Journées d'Études sur la Parole, pages 102–111.

1 Introduction

Une altération fonctionnelle au niveau de la communication est généralement présente dès lors qu'un traitement intervient pour des maladies qui affectent les voies aérodigestives supérieures (VADS), telles que le cancer de la tête et du cou (HNC) et les maladies neurodégénératives responsables de dysarthries. Étant donné que des répercussions fonctionnelles majeures sur les VADS sont susceptibles de survenir, une perte d'intelligibilité de la parole est souvent observée, impactant la qualité de vie du patient (de Graeff *et al.*, 2000). En raison du temps nécessaire à la mise en œuvre progressive du post-traitement et de sa durée, un diagnostic précoce est pertinent. Ce diagnostic ainsi que le suivi des troubles sont couramment basés sur une évaluation perceptive de l'intelligibilité de la parole.

Dans les mesures cliniques perceptives, il existe en plus de l'intelligibilité, la sévérité des troubles de la parole qui peut être considérée comme une mesure plus globale incluant la première. Malgré le fait qu'elles servent à deux objectifs différents, ces deux mesures partagent des corrélations élevées : l'une évalue la qualité de parole à un bas niveau, acoustico-phonétique (intelligibilité) et l'autre évalue le degré d'impact du trouble de la parole de façon plus globale sur la communication fonctionnelle (sévérité). De plus, ces mesures sont connues pour être hautement variables, biaisées et subjectives, car leurs évaluations peuvent être conditionnées par la connaissance préalable de la tâche à réaliser (par exemple, la lecture de textes), des évaluations antérieures ou encore une connaissance *a priori* des patients (Fex, 1992). Une approche automatique est alors une alternative pouvant favoriser des prédictions plus fiables et plus objectives.

Les approches pour la prédiction automatique de l'intelligibilité vont de scores basés sur les performances de reconnaissance automatique de la parole (Christensen *et al.*, 2012; Fontan *et al.*, 2017) à des techniques de traitement du signal plus traditionnelles ou à des méthodologies d'apprentissage automatique (Quintas *et al.*, 2022; Bin *et al.*, 2019). Le paradigme de l'embedding de locuteurs, où les énoncés de parole sont représentés par des vecteurs de dimension fixe ayant des propriétés discriminantes entre les locuteurs, a montré des apports intéressants sur des tâches distinctes telles que l'intelligibilité de la parole (Laaridh *et al.*, 2018; Quintas *et al.*, 2020), mais aussi sur l'évaluation générale de la parole pathologique (Zargarbashi & Babaali, 2019; Codosero *et al.*, 2019).

Alors que les travaux récents sur la prédiction automatique de l'intelligibilité de la parole affichent des résultats prometteurs, la majorité des systèmes sont testés sur des données qui ne reproduisent que difficilement les conditions réelles d'un hôpital. Les salles d'enregistrement sont traitées acoustiquement, en utilisant toujours le même microphone, une distance de microphone prédéfinie et des tâches de parole prédéfinies(Clapham *et al.*, 2012; Woisard *et al.*, 2020). Étant donné que l'objectif final de ces systèmes est de fournir des estimations d'intelligibilité plus robustes et écologiques, il devient essentiel de les évaluer sur différents ensembles de données et scénarii cliniques lorsqu'on envisage leur mise en œuvre directe.

De plus, les tâches de parole généralement utilisées pour les évaluations perceptives ou automatiques (Fredouille *et al.*, 2019; Quintas *et al.*, 2020) de la parole sont habituellement la lecture de textes et les pseudo-mots. Les tâches impliquant la parole spontanée, elles, n'ont pas encore été suffisamment étudiées dans le domaine des évaluations automatiques (Balaguer *et al.*, 2019b). Une évaluation automatique sur de la parole spontanée dans des conditions cliniques réelles se rapproche pourtant grandement de l'environnement dans lequel ce type de systèmes serait déployé, tout en utilisant des données représentant étroitement la capacité de communication réelle d'un locuteur.

Par conséquent, dans ce travail, nous menons des expériences sur un système de prédiction d'in-

telligibilité adapté à des données hospitalières plus écologiques. Ainsi, nous avons l'intention de : (i) Analyser la fiabilité d'un système de prédiction de l'intelligibilité/sévérité basé sur des données enregistrées dans des conditions cliniques réelles, et (ii) Évaluer la fiabilité du même système lors de la prédiction de l'intelligibilité/sévérité basée sur des segments de parole spontanée, obtenus à partir d'entretiens patient-soignant.

Le reste de cet article est organisé comme suit. La section 2 décrit notre système et notre méthodologie globale. La section 3 présente notre corpus, nos expériences et nos résultats. Enfin, les sections 4 et 5 proposent une discussion et nos conclusions et perspectives, respectivement.

2 Méthodologie

De manière similaire à (Quintas *et al.*, 2020), le système automatique de prédiction de l'intelligibilité utilise le paradigme de représentation de locuteurs et un réseau de neurones. Dans cet article, le système [1] est adapté pour prédire deux mesures perceptives dans un cadre multi-tâches : l'intelligibilité de la parole (INT), définie comme le degré selon lequel le message du locuteur peut être compris par un auditeur, et la sévérité des troubles de la parole (SEV) définie comme le degré d'altération de l'intelligibilité associé à d'autres variables du signal de parole telles que la qualité d'émission du code acoustico-phonétique, la vitesse de parole et d'autres paramètres temporels ou prosodiques pertinents (Balaguer *et al.*, 2019a). Ces deux mesures, bien que partageant une corrélation élevée et un certain degré de similarité, servent à des fins distinctes. Alors que l'intelligibilité évalue directement la qualité de parole d'un patient donné au niveau acoustico-phonétique, la sévérité des troubles de la parole sert de score global de la maladie qui encapsule différents aspects de la communication verbale.

2.1 Représentations de locuteurs

Les représentations de locuteurs sont des représentations de longueur fixe généralement utilisées dans la vérification des locuteurs, la segmentation en locuteurs et la reconnaissance automatique de la parole. Récemment, ils ont montré une capacité à transmettre des attributs du locuteur permettant la détection des troubles affectant la parole (Codosero *et al.*, 2019). Depuis, nous observons une utilisation croissante de ces représentations pour l'évaluation automatique de la parole pathologique. De plus, étant donné la bonne performance de ce paradigme sur des tâches qui traitent généralement d'une parole plus spontanée (par exemple, la segmentation en locuteurs dans des contextes conversationnels) (Larcher *et al.*, 2021), nous émettons l'hypothèse qu'une approche basée sur les représentations pourrait mieux aider à prédire l'intelligibilité de la parole dans ce même contexte, par opposition à la parole lue, généralement utilisée dans les évaluations cliniques. Étant donné que les représentations de locuteurs à base de *x-vecteurs* ont surpassé les *i-vecteurs* dans la prédiction d'intelligibilité (Quintas *et al.*, 2020), deux classes de représentations de locuteurs ont été testées dans la présente étude, toutes deux extraites à l'aide du toolkit Speechbrain (Ravanelli *et al.*, 2021).

Les premières classes sont des *x-vecteurs*. Elles sont utilisées dans (Quintas *et al.*, 2020) pour prédire l'intelligibilité de la parole en extrayant les caractéristiques discriminantes entre les locuteurs (Snyder *et al.*, 2018). L'extraction des représentations [2] fonctionne en faisant passer le signal de parole à travers

1. https://gitlab.irit.fr/samova/embedding_intelligibility
2. https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

un bloc de réseaux neuronaux à retard temporel (TDNN) qui opère sur des trames de parole avec un contexte temporel réduit centré sur la trame actuelle. Les couches TDNN ultérieures s'appuient sur le contexte temporel des couches précédentes. Une couche de regroupement statistique agrège toutes les sorties au niveau des trames en une dimension de longueur fixe, qui est ensuite alimentée dans un bloc entièrement connecté. Les *x-vecteurs* sont extraits à partir de la composante affine de la dernière couche entièrement connectée. Le système a été pré-entraîné avec les données voxceleb1 (Nagrani *et al.*, 2017) et voxceleb2 (Chung *et al.*, 2018), puis testé sur l'ensemble de test voxceleb1, atteignant une erreur (EER, (Cheng & Wang, 2004)) de 3,2%.

Ensuite, les représentations de locuteurs Ecapa TDNN [3], plus récents que les *x-vecteurs*, ont été expérimentés. Ces représentations de longueur fixe s'appuient sur le concept des *x-vecteurs*, avec cependant plusieurs améliorations qui suggèrent une meilleure performance en vérification des locuteurs par rapport à d'autres représentations (Desplanques *et al.*, 2020). Nous supposons que ces améliorations permettent au réseau de se concentrer davantage sur les caractéristiques du locuteur qui ne s'activent pas aux mêmes instants, par exemple les propriétés spécifiques du locuteur sur les voyelles par rapport aux propriétés spécifiques du locuteur sur les consonnes. Nous émettons l'hypothèse que ces améliorations pourraient fournir un embedding de locuteur plus robuste pour l'évaluation de la parole pathologique, et par conséquent surpasser les *x-vecteurs* précédemment utilisés. De manière similaire à l'extracteur précédemment introduit, le système Ecapa TDNN a été pré-entraîné en utilisant les données voxceleb1 et voxceleb2, puis testé sur l'ensemble de test voxceleb1, atteignant une EER de seulement 0,8%.

2.2 Réseau neuronal

La figure 1 présente un diagramme de notre réseau neuronal. Le réseau reçoit en entrée les représentations de locuteurs qui, selon le type, ont des dimensions fixes distinctes (512 pour les *x-vecteurs* et 192 pour l'Ecapa TDNN). De plus, le signal passe à travers deux couches de dimensions fixes, puis enfin les deux couches multi-tâches qui prédisent les deux mesures perceptives différentes. Le système est optimisé à l'aide d'une fonction de perte de l'erreur quadratique moyenne (MSE) et d'un algorithme d'optimisation Adam. Afin de prédire deux mesures distinctes, la fonction de perte prend en compte ces deux mesures avec des contributions égales, ce qui signifie un poids de 50 % pour l'intelligibilité et 50 % pour la sévérité. En raison des corrélations élevées généralement observées entre ces deux mesures, nous émettons l'hypothèse que leur apprentissage conjoint conduira à une estimation de l'intelligibilité de la parole meilleure et plus robuste.

2.3 Entraînement et validation

Un système pour chaque type de représentations a ainsi été entraîné sur le corpus C2SI (Woisard *et al.*, 2020). Le corpus comprend une variété de patients souffrant de cancer de la tête et du cou avec des localisations tumorales initiales différentes, ainsi que des locuteurs sains. Les deux systèmes ont été entraînés et validés en utilisant la tâche de lecture de texte segmentée. Un schéma d'augmentation des données, similaire à celui de (Quintas *et al.*, 2020), basé sur une distorsion temporelle (Vachhani *et al.*, 2018; Ko *et al.*, 2015) qui préserve le timbre et l'enveloppe spectrale, a été implémenté pendant l'entraînement. Un total de 98 locuteurs a été utilisé pour l'entraînement et un sous-ensemble de 10

3. https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

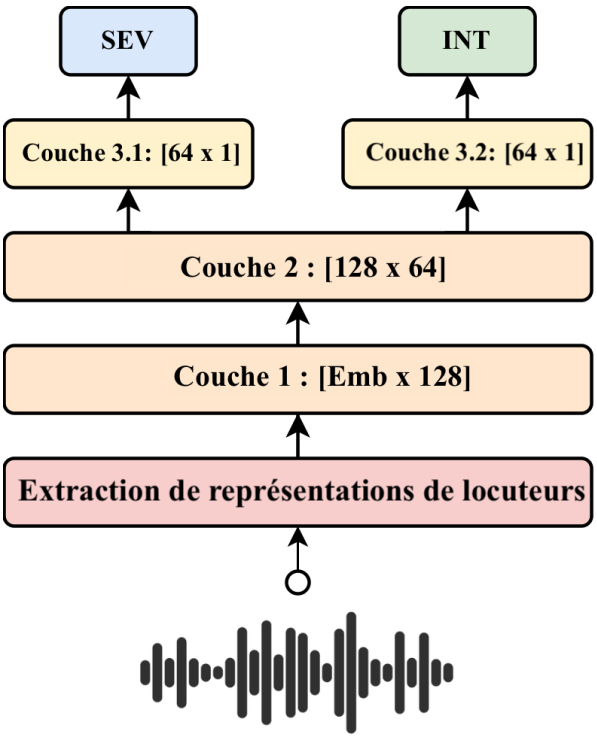


FIGURE 1 – Diagramme schématique du réseau neuronal proposé et de l'approche d'apprentissage multi-tâches correspondante. La mention "Emb" indique la taille de la représentation de locuteurs.

locuteurs avec des degrés d'intelligibilité variables a été utilisé comme ensemble de validation. Un batchsize de 8, un taux d'apprentissage de 0,001 et une fonction de perte (loss) de 0,2 ont été utilisés sur 20 epochs.

3 Expériences et Résultats

3.1 Corpus SpeeCOmco

Notre corpus de parole et de communication en oncologie (SpeeCOmco) est un ensemble de 27 patients présentant des degrés d'intelligibilité variables ayant enregistré différentes tâches dans des conditions cliniques réelles (Balaguer, 2021). Dans la population du corpus, l'âge moyen est de 66,3 ans (min. 38 ans, max. 83 ans) avec une représentation féminine de 37%. Les enregistrements ont été réalisés dans des salles de consultation, non traitées acoustiquement, avec l'utilisation d'un micro-casque couramment utilisé en pratique clinique et la présence d'un certain niveau de bruit de fond. Les enregistrements ayant eu lieu dans un environnement hospitalier, plus précisément lors de rendez-vous cliniques en orthophonie, les conditions d'enregistrement imitent exactement les conditions dans lesquelles le présent système serait déployé. Tous les patients sont des locuteurs natifs du français.

Pour cet ensemble de patients, l'intelligibilité moyenne et la sévérité des troubles de la parole ont été calculées sur la base d'une évaluation perceptive indépendante de six professionnels de la santé. Chaque locuteur a reçu un score entre 0 et 10, plus la valeur est petite, moins la parole est intelligible. La même échelle est utilisée pour la sévérité. Le coefficient de corrélation intraclasse (CCI) a été

calculé pour évaluer la fiabilité inter-juges. Un CCI de 0,816 a été obtenu pour les six juges lors de l'évaluation de l'intelligibilité de la parole parmi tous les patients et un CCI de 0,852 a été obtenu pour la sévérité, montrant un bon niveau d'accord entre les experts.

Plusieurs tâches, classiques dans l'évaluation de la parole pathologique, ont été utilisées :

1. **Lecture de texte (LEC)**. Les locuteurs ont été invités à lire le premier paragraphe de « La chèvre de M. Seguin », un conte d'Alphonse Daudet choisi car il est assez long pour inclure presque tous les phonèmes français. Ce passage est également bien connu et largement utilisé en phonétique clinique française (Ghio *et al.*, 2012).
2. **Phrases avec semi-voyelles (PHR)**. Les locuteurs lisent deux phrases contenant les semi-voyelles françaises [w] et [U], absentes du texte LEC.
3. **Inflexion consonantique (CSN)**. Les locuteurs ont été invités à lire 17 phrases sous la forme de *« Le sac eu**C**eu convient »*, où le ***C*** est remplacé à chaque phrase par une consonne différente.
4. **Pseudo-mots (DAP)**. Chaque locuteur enregistre un ensemble de 52 pseudo-mots, inexistants dans la langue française (Lalain *et al.*, 2020). Chaque pseudo-mot a été généré automatiquement de manière à respecter les règles phonotactiques et orthographiques du français.
5. **Parole spontanée (SPO)**. Dans cette tâche, l'échantillon audio provient d'un entretien entre un orthophoniste et le locuteur. La conversation porte sur la communication quotidienne et les limitations perçues par le locuteur. Les segments de parole spontanée sont obtenus grâce à un détecteur d'activité vocale (VAD). Les segments de moins de 3s et de plus de 10s ont été écartés afin de minimiser le nombre d'artefacts capturés. Les segments avec une trop forte présence de la voix du thérapeute ont également été supprimés. La durée de l'entretien peut ici beaucoup différer entre les locuteurs (ainsi que le nombre de fichiers associés : de 8 à 56 dans ce corpus).

3.2 Tests et résultats

Les 27 patients et les cinq tâches de parole enregistrées ont été évalués à l'aide des deux systèmes décrits précédemment, l'un utilisant les *x-vecteurs* et l'autre les représentations de locuteurs Ecapa TDNN. À l'exception de la tâche de parole spontanée, dont la prédiction d'intelligibilité correspond à la moyenne des fichiers segmentés en parole (via une VAD mentionnée précédemment), les tâches ont été analysées sur un seul fichier audio par locuteur. Le tableau 1 illustre la corrélation de Spearman (ρ) pour les différentes tâches évaluées et le type de représentations, et les valeurs de l'erreur quadratique moyenne (RMSE) sur l'intelligibilité de la parole (ainsi que sur la prédiction de la sévérité des troubles de la parole). Les corrélations sont élevées ($\rho > 0{,}82$) sur quatre des cinq tâches lors de l'utilisation des *x-vecteurs*. De même, les erreurs sont faibles (RMSE < 1,5) sur trois des tâches. La figure 2 affiche un graphique des prédictions associées à la tâche de parole spontanée.

4 Discussion

Les résultats ont montré des valeurs de corrélation et d'erreur encourageantes. De plus, étant donné que l'objectif final de cet article est de valider la mise en œuvre d'un système de prédiction de l'intelligibilité et de sévérité sur le plan clinique, une combinaison de corrélation élevée et d'erreurs

TABLE 1 – Valeurs de corrélation et d'erreur obtenues lors du test du système sur les différentes tâches du corpus SpeeCOmco. Les valeurs en gras marquent les tâches avec les meilleurs résultats. Toutes les corrélations ont atteint une valeur $p < 0{,}05$, les rendant statistiquement significatives.

Mesures perceptives		**Sévérité**				**Intelligibilité**			
Représentations		**Ecapa TDNN**		***X-vecteurs***		**Ecapa TDNN**		***X-vecteurs***	
Métriques d'évaluation		ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
Tâches de parole	**LEC**	0.783	1.873	**0.866**	**1.384**	0.826	2.070	**0.891**	**1.322**
	PHR	0.784	1.782	0.805	1.772	0.807	1.854	**0.842**	**1.460**
	CSN	0.643	2.060	0.861	2.124	0.673	2.147	0.859	1.971
	DAP	0.296	2.673	0.724	2.219	0.371	2.805	0.731	1.881
	SPO	0.657	2.124	0.818	1.820	0.695	2.252	**0.828**	**1.468**

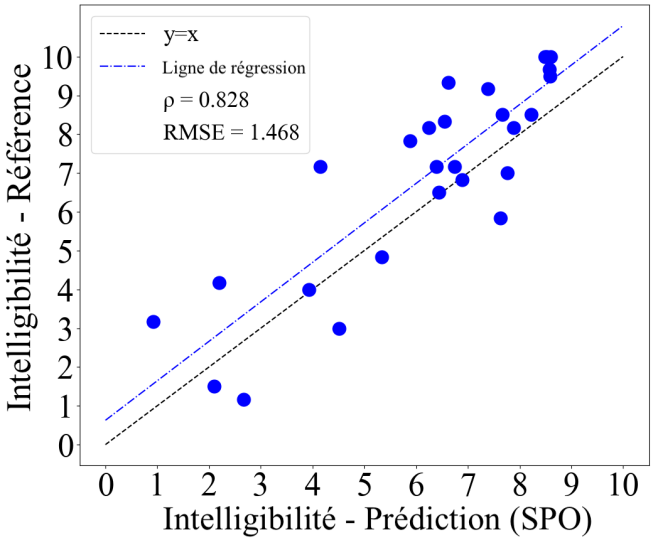


FIGURE 2 – Résultats de la prédiction de l'intelligibilité sur le corpus SpeeCOmco en utilisant la tâche SPO (*x-vecteurs*).

faibles doit être envisagée. Malgré les résultats prometteurs des représentations ECAPA pour la vérification des locuteurs (Desplanques *et al.*, 2020), dans notre contexte spécifique, les *x-vecteurs* les surpassent dans toutes les tâches de parole, tant pour l'intelligibilité que pour la sévérité. Cet aspect valide non seulement l'utilisation des *x-vecteurs* pour la parole pathologique, mais montre également que tous les types de représentations de locuteurs ne conviennent pas à ce type d'analyse. Une étude comparative approfondie sur les différentes représentations de locuteurs pour ce type d'évaluation ainsi que la recherche d'une meilleure métrique pour analyser leur performance sur la parole pathologique est une piste intéressante pour les travaux futurs.

Les *x-vecteurs* ont obtenus des résultats intéressants et fiables, à l'exception de la tâche DAP. D'une part, ceci peut être expliqué par la qualité de ces enregistrements : artefacts de bruit entre les pseudo-mots. D'autre part, aucune prosodie ni coarticulation entre les pseudo-mots ne sont présentes. Le système ayant été entraîné sur la tâche de lecture de texte, bien évidemment, il se comporte mieux

sur cette même tâche LEC. Cependant, la tâche PHR a également obtenue des résultats fiables, bien que les fichiers audio soient beaucoup plus courts que ceux de la tâche de lecture. Enfin, les résultats sur la parole spontanée ont présenté une corrélation forte et une faible erreur (comparable aux autres tâches). L'évaluation de ce type de parole devient très pertinente en raison du fait qu'il s'agit d'un médium peu exploré (Balaguer *et al.*, 2019b), offrant une vision écologique de la capacité réelle du patient à communiquer. Cette évaluation automatique sur la parole spontanée comble un vide dans la littérature concernant le test des approches automatiques sur ce type de parole enregistrée, et peut être considérée comme une pierre angulaire vers des prédictions d'intelligibilité plus pertinentes, plus écologiques et plus fiables.

5 Conclusions et perspectives

Cet article a examiné la fiabilité d'un prédicteur automatique de l'intelligibilité de la parole basé sur des représentations de locuteurs dans des conditions écologiques (consultations cliniques à l'hôpital). Différentes évaluations ont été réalisées dans une méthodologie d'apprentissage multi-tâches. Les résultats ont suggéré une bonne capacité de généralisation, illustrée par des corrélations allant jusqu'à 0,891 et des erreurs de 1,322. Les métriques obtenues sur la tâche de parole spontanée sont non seulement comparables à celles des autres tâches, mais ouvrent également la possibilité d'une utilisation plus large des évaluations automatiques, un sujet actuellement sous-exploré.

Même si l'intelligibilité de la parole est la principale mesure subjective à analyser et à prédire ici, les résultats sont similaires sur la sévérité des troubles de la parole. Cet aspect montre que le paradigme multi-tâches peut être efficace pour ce type de mesure, et permet d'apprendre d'autres mesures perceptives, telles que la prosodie, la résonance et les distorsions phonémiques. De plus, une mesure d'intelligibilité qui peut être considérée comme une combinaison de ces autres paramètres (de Bodt *et al.*, 2002) peut être plus interprétable, avec une valeur ajoutée dans un environnement clinique.

Le système développé a été entraîné sur une tâche de lecture. Malgré la bonne généralisation du système sur une variété de nouveaux patients et de tâches de parole, un entraînement supplémentaire sur d'autres tâches (comme de la parole spontanée), ainsi que sur d'autres langues et maladies (comme la maladie de Parkinson, la sclérose latérale amyotrophique, etc.) pourrait non seulement augmenter les performances, mais aussi rendre le système encore plus robuste. Le développement d'un modèle d'intelligibilité automatique multi-pathologies est une perspective intéressante pour les travaux futurs. Cependant, il convient de le concevoir avec soin, car une solution fonctionnelle pour l'intelligibilité de la parole dans les cancers de la tête et du cou peut ne pas nécessairement correspondre à la meilleure approche pour les maladies neurologiques. Cela est principalement dû au type de problèmes affectant la parole qui diffèrent grandement entre les deux ensembles de maladies, rendant les systèmes conçus difficilement transposables à toutes pathologies.

Au vu de la capacité de généralisation du système proposé sur différents types de données hospitalières dans le contexte des cancers de la tête et du cou, les travaux futurs prévoient la mise en œuvre directe du présent système en clinique, grâce à une application mobile qui sera utilisée par les thérapeutes.

6 Remerciements

Ce projet a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention Marie Skłodowska-Curie No 766287.

Références

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*. Thèse de doctorat, Université Paul Sabatier - Toulouse III.

BALAGUER M., BOISGUÉRIN A., GALTIER A., GAILLARD N., PUECH M. & WOISARD V. (2019a). 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.

BALAGUER M., POMMÉE T., FARINAS J., PINQUIER J., WOISARD V. & SPEYER R. (2019b). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Journal of the Sciences and Specialities of Head and Neck*, **42(1)**, 111–130.

BIN L., KELLEY M. C., AALTO D. & TUCKER B. V. (2019). Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks. *International Congress of Phonetic Sciences (ICPHs')*.

CHENG J.-M. & WANG H.-C. (2004). A method of estimating the equal error rate for automatic speaker verification. *Proceedings of ISCSLP*.

CHRISTENSEN H., CUNNINGHAM S., FOX C., GREEN P. & HAIN T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of Interspeech*.

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. *Proceedings of Interspeech*.

CLAPHAM R., VAN DER MOLEN L., VAN SON R., VAN DEN BREKEL M. & HILGERS F. (2012). Nki-ccrt corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

CODOSERO J. M. P., ESPINOZA-CUADROS F., ANTÓN-MARTÍN J., BARBERO-ALVAREZ M. A. & GÓMEZ L. A. H. (2019). Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training. *IEEE Journal of Selected Topics in Signal Processing*, **14(2)**, 240–250.

DE BODT M., HUICI M. E. & HEYNING P. V. D. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, **35(3)**, 283–292.

DE GRAEFF A., DE LEEUW R. J., ROS W. J., HORDIJK G.-J., BLIJHAM G. H. & WINNUBST J. A. (2000). Long-term quality of life of patients with head and neck cancer. *The Laryngoscope, Volume 110, Issue 1*, p. 98–106.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). ECAPA-TDNN : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Proceedings of Interspeech*.

FEX S. (1992). Perceptual evaluation. *Journal of Voice*, **6**(2), 155–158.

FONTAN L., FERRANÉ I., FARINAS J., PINQUIER J., TARDIEU J. & MAGNEN 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, Volume 50(1)*, **60(9)**, 2394–2405.

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. *International Congress of Phonetic Sciences (ICPhS)*, p. 3051–3055.

GHIO A., POUCHOULIN G., TESTON B., PINTO S., FREDOUILLE C. & ET AL (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers ? *Speech Communication*, **54**, 664–679.

KO T., PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). Audio augmentation for speech recognition. *Proceedings of Interspeech*.

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. *Proceedings of Interspeech*.

LALAIN M., GHIO A., GIUSTI L., ROBERT D., FREDOUILLE C. & WOISARD V. (2020). Design and development of a speech intelligibility test based on pseudowords in french : Why and how ? *Journal of Speech, Language and Hearing Research*, **63(7)**, 2070–2083.

LARCHER A., MEHRISH A., TAHON M., MEIGNIER S., CARRIVE J., DOUKHAN D., GALIBERT O. & EVANS N. (2021). Speaker embeddings for diarization of broadcast data in the allies challenge. *Proceedings of ICASSP*.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : A largescale speaker identification dataset. *Proceedings of Interspeech*.

QUINTAS S., MAUCLAIR J., WOISARD V. & PINQUIER J. (2020). Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer. *Proceedings of Interspeech*.

QUINTAS S., MAUCLAIR J., WOISARD V. & PINQUIER J. (2022). Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. *Proceedings of Interspeech*.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. *arXiv :2106.04624*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embedings for speaker recognition. *Proceedings of ICASSP*.

VACHHANI B., BHAT C. & KOPPARAPU S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. *Proceedings of Interspeech*.

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. & ROGER V. (2020). 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.

ZARGARBASHI S. & BABAALI B. (2019). A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language. *arXiv :1910.00330*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Mesure du niveau de proximité entre enregistrements audio et évaluation indirecte du niveau d'abstraction des représentations issues d'un grand modèle de langage

Maxime Fily[1, 2] Guillaume Wisniewski[1] Séverine Guillaume[2] Gilles Adda[3]
Alexis Michaud[2]
(1) LLF, CNRS, Université Paris-Cité, F-75013, Paris, France
(2) LACITO, CNRS, Université Sorbonne Nouvelle, F-94800, Villejuif, France
(3) LISN, CNRS, Université Paris-Saclay, F-91405, Orsay, France
maxime.fily@gmail.com, guillaume.wisniewski@u-paris.fr,
{severine.guillaume, alexis.michaud}@cnrs.fr, gilles.adda@limsi.fr

RÉSUMÉ

Nous explorons les représentations vectorielles de la parole à partir d'un modèle pré-entraîné pour déterminer leur niveau d'abstraction par rapport au signal audio. Nous proposons une nouvelle méthode non-supervisée exploitant des données audio ayant des métadonnées soigneusement organisées pour apporter un éclairage sur les informations présentes dans les représentations. Des tests ABX déterminent si les représentations obtenues via un modèle de parole multilingue encodent une caractéristique donnée. Trois expériences sont présentées, portant sur la qualité acoustique de la pièce, le type de discours, ou le contenu phonétique. Les résultats confirment que les différences au niveau de caractéristiques linguistiques/extra-linguistiques d'enregistrements audio sont reflétées dans les représentations de ceux-ci. Plus la quantité d'audio par vecteur est importante, mieux elle permet de distinguer les caractéristiques extra-linguistiques. Plus elle est faible, et mieux nous pouvons distinguer les informations d'ordre phonétique/segmental. La méthode proposée ouvre de nouvelles pistes pour la recherche et les travaux comparatifs sur les langues peu dotées.

ABSTRACT

Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models

In the highly constrained context of low-resource language studies, we examine vector representations of speech from a pretrained model to determine their level of abstraction from the audio signal. We propose a new unsupervised method using ABX tests on audio recordings with carefully curated metadata to shed light on the type of information present in the representations. ABX tests determine whether the representations computed by a multilingual speech model encode a given characteristic. Three experiments are devised : one on room acoustics aspects, one on linguistic genre, and one on segmental phonetic aspects. The results confirm that the representations extracted from recordings with different linguistic/extra-linguistic characteristics differ along the same lines. Embedding more audio into a vector better discriminates extra-linguistic characteristics, whereas shorter snippets are better for distinguishing segmental information. The method is fully unsupervised, potentially opening up new avenues of research for comparative work on under-documented languages.

MOTS-CLÉS : TAL, langues peu dotées, méthodes non-supervisées.

KEYWORDS: NLP, under-documented languages, unsupervised methods.

1 Introduction

La parole, lorsqu'elle se présente sous la forme d'enregistrements audio, est multi-factorielle : une voix enregistrée transmet un message, une intention, une émotion. L'enregistrement contient également des informations au delà du signal de parole, sur l'environnement, l'identité du locuteur ou de la locutrice par exemple. Ce travail aborde la question de la nature de l'information encodée dans les représentations vectorielles produites par un réseau de neurones appris de manière auto-supervisée comme `wav2vec2` (Baevski *et al.*, 2020). Notre objectif est d'étudier le niveau d'abstraction présent dans les représentations construites par ces modèles et plus précisément de vérifier si celles-ci ne contiennent que des informations liées au contenu linguistique ou si elles contiennent également des informations *indexicales* (Foulkes, 2010, 7).

Notre dispositif expérimental s'appuie sur des jeux de données « sur mesure », sélectionnées à partir de corpus de linguistique documentaire afin d'évaluer comment une différence donnée dans le signal d'entrée se reflète dans les vecteurs construits par le réseau de neurones. Considérer des données issues de ce type de corpus nous permet d'accéder à des métadonnées riches sur les conditions d'enregistrement. Nous utilisons des tests `ABX` pour mesurer l'impact de certains facteurs (locuteur·rice, microphone, ...) sur les distances entre représentations à travers trois jeux de données décrits en section 2.

Les résultats fournissent un aperçu sur la nature des informations encodées dans les représentations d'un modèle tel que XLSR-53 (Baevski *et al.*, 2020; Babu *et al.*, 2021), et suggèrent que les tests `ABX` peuvent être exploités pour faire ressortir des différences dans la configuration acoustique (salle, microphone), certaines caractéristiques de la voix, ou dans le contenu linguistique. Une étude paramétrique montre que le traitement de l'audio par extraits de 10 s est suffisant pour faire ressortir les différences dans la configuration acoustique et dans les propriétés de la voix, tandis que des extraits de 1 s sont meilleurs pour explorer les caractéristiques segmentales.

Cette étude fournit une méthode innovante pour détecter des facteurs de confusion dans des corpus destinés à de l'apprentissage automatique, ainsi qu'un moyen pour accélérer la classification d'enregistrements (p.ex. par niveau de bruit, par type) selon des informations fines, qui ne sont pas toujours présentes dans les métadonnées.

2 Méthode expérimentale

Notre méthode repose sur : (i) des tests de similarité pour déterminer si une caractéristique d'un enregistrement audio est encodée dans la représentation vectorielle d'un enregistrement ou non, et (ii) des corpus audio avec des métadonnées riches, qui nous permettent de construire des ensembles de données partageant ou non **une caractéristique à la fois** : langue, locuteur·rice, acoustique de la pièce, type de microphone, caractéristiques de la voix, contenu segmental, etc.

Tests ABX Pour déterminer, de manière non supervisée, si un modèle de parole encode une caractéristique $\mathcal{C}$ du signal de parole, nous utilisons un test `ABX` (Carlin *et al.*, 2011; Schatz *et al.*, 2013). Ce test s'appuie sur les représentations vectorielles construites par un modèle pré-entraîné de

trois audios : deux extraits, notés A et X, partagent une caractéristique $\mathcal{C}$, et un, noté B, ne la possède pas. Le test ABX consiste simplement à vérifier si la distance [1] $d(A, X)$ est plus petite que $d(A, B)$.

Le score ABX correspond à la proportion de triplets pour lesquels la condition $d(A, X) < d(A, B)$ est vraie. Un score ABX proche de 50 % indique qu'en moyenne, la distance entre A et X est supérieure ou égale à la distance entre A et B, ce qui suggère que $\mathcal{C}$ n'est pas encodé dans la représentation vectorielle de l'enregistrement. Inversement, plus le score est proche de 100 %, plus la représentation capture la caractéristique $\mathcal{C}$.

Les tests ABX sont intéressants pour les scénarios à faibles ressources (comme le nôtre) car ils ne nécessitent pas de données pour entraîner un classifieur (contrairement aux sondes linguistiques (Belinkov & Glass, 2019) souvent utilisées pour analyser les représentations vectorielles) : toutes les données disponibles sont utilisées pour tester si la propriété est encodée dans la représentation ou non.

Corpus Notre étude s'appuie sur des enregistrements en na et en naxi, deux langues parlées dans le sud-ouest de la Chine. Le na est la langue maternelle d'environ 50 000 personnes. Le naxi est plus répandu, puisqu'il est la langue maternelle d'environ 200 000 personnes. Typologiquement similaires (langues SOV, de structure (C)(G)V+T [2], isolantes) le naxi et le na sont des langues apparentées, sans être pour autant mutuellement intelligibles. Les systèmes tonals diffèrent au niveau de la complexité des phénomènes morpho-tonologiques qui est très importante en na (Michaud, 2017, 425). Au plan phonologique, le na a une nasalité de type vocalique derrière les consonnes /h/ et plus marginalement derrière une attaque vide (Michaud *et al.*, 2012). Ces deux langues sont aujourd'hui en déclin, car progressivement remplacées par le mandarin, langue officielle utilisée dans les écoles, les administrations et les médias (Michaud & Latami, 2011; Zhao, 2022).

Tous les enregistrements proviennent de la collection Pangloss, une archive en libre accès de « langues peu documentées », et sont consultables via les DOI fournis en annexe (voir https://hal.science/hal-04583516). Trois séries d'enregistrements sélectionnées pour leurs caractéristiques sont examinées :

(i) La ***série du conte populaire*** consiste en sept sessions d'enregistrement d'un même conte en na, raconté par une même locutrice. La i^{e} session d'enregistrement sera désignée par V_i. Cette série se concentre sur l'effet des conditions d'enregistrement, qui sont légèrement différentes d'une version à l'autre, et que nous chercherons à distinguer à l'aide de tests ABX sont effectués. Plus précisément, nous nous concentrons sur trois lots :

- Le premier lot étudié comprend trois versions : V_1, V_2 et V_3. V_1 a été enregistré dans une salle avec une réverbération perceptible, tandis que V_2 et V_3 ont été enregistrés dans une salle mieux isolée phoniquement.
- Le deuxième lot est composé de V_6 et V_7. Ces deux versions ont été enregistrées dans les mêmes conditions acoustiques. L'audio a été capté simultanément par deux microphones : un micro-casque et un micro à main inséré dans un petit support.
- Le troisième lot compare V_4 et V_5 à tous les autres enregistrements de la *série du conte populaire*. Les enregistrements V_4 et V_5 ont pour récepteur [3] un auditeur natif alors que dans les autres enregistrements le récepteur était le linguiste collectant les données.

1. Nous avons utilisé une distance cosinus dans toutes nos expériences.
2. G = *Glide*, T = Ton.
3. Ici, récepteur est le destinataire d'un message, en cohérence avec la terminologie de Shannon (1948) qui propose l'idée très *traitement du signal* qu'il existe un canal de communication *(communication channel)* entre un émetteur et un récepteur. Ce canal est doublé d'un canal de rétroaction en sens inverse *(communication backchannel)*, qui influence la façon dont l'émetteur livre son message.

Ces enregistrements sont particulièrement intéressants car certaines variables (le sujet de conversation et le·la locuteur·rice) sont contrôlées, ce qui permet de se concentrer sur l'influence d'autres facteurs spécifiques (par exemple, l'acoustique de la pièce).

(ii) La ***série des répertoires de chant*** consiste en cinq enregistrements d'une même chanteuse professionnelle Naxi. Trois enregistrements sont de la chanson seule, un enregistrement est un récit, et un enregistrement comporte les deux genres (« Alili », qui est 50% texte et 50% chanson). Notre objectif est de comparer ces enregistrements. Un chanteur entraîné présente des propriétés vocales très différentes selon qu'il chante ou qu'il parle : en particulier, le timbre des voyelles et la tessiture sont affectées (Castellengo, 2016, 458). Ces différences sont perçues par les auditeurs (Castellengo, 2016, 187). Cette expérience vise à vérifier si ces différences sont reflétées dans les représentations.

(iii) La ***série phonétique*** est composée de cinq enregistrements d'élicitations phonétiques basés sur un même corpus et d'un enregistrement de mots dans une phrase porteuse basé sur un corpus différent, en langue na. Trois locutrices identifiées comme `AS`, `RS` et `TLT` sont prises en compte. Nous avons inclus deux sessions d'enregistrement, ce qui permet une comparaison inter et intra-locuteurs.

En utilisant ces enregistrements, nous adoptons une approche à l'opposé de ce qui se fait dans de nombreux travaux portant sur l'analyse des représentations neuronales : plutôt que de considérer de « grands » corpus avec de fortes variabilités (approche *big data*), nous privilégions un corpus de petite taille mais dans lequel de nombreux facteurs sont contrôlés (approche *beautiful data*).

Modèle Dans toutes nos expériences, nous utilisons le modèle `XLSR-53`[4], une architecture `wav2vec2` entraînée sur 56 kh de données audio (brutes) dans 53 langues (Conneau *et al.*, 2020), y compris des langues à tons (p.ex. mandarin, cantonais, lao, ...). Ni le na ni le naxi ne sont présents dans les données de pré-entraînement de ce modèle, mais Macaire *et al.* (2021) ont montré que ce modèle pouvait être affiné pour faire de la reconnaissance de la parole sur du na (après affinage, le CER en transcription est de 8 %), et qu'il était donc capable de gérer la diversité des réalisations de surface de cette langue, et notamment les tons.

Pour les comparaisons, nous considérons des extraits audio d'une durée de 1, 5, 10 et 20 secondes afin d'étudier l'effet de la longueur de l'extrait sur notre test `ABX`. Nous utilisons une stratégie de *max-pooling* pour construire un unique vecteur représentant l'extrait. Nous créons ensuite des cartes thermiques des scores `ABX`, pour comparer un à un les enregistrements.

Nous utilisons les représentations de la couche 21. Ce choix est basé sur les conclusions de Pasad *et al.* (2021, 2023) et Li *et al.* (2022, 2023) ainsi que sur notre pré-étude de sensibilité (méthode des sondes linguistiques), qui montrent toutes que la capacité de `wav2vec2` à capturer l'information linguistique diminue fortement dans les trois dernières couches (voir annexes : `https://hal.science/hal-04583516`).

3 Résultats

Avec les tests `ABX`, nous étudions si les représentations audio calculées par `wav2vec2` capturent ou non des informations spécifiques dans des signaux audio soigneusement sélectionnés. Notre approche est non-supervisée ce qui nous permet de l'appliquer à de petits jeux de données. Nos expériences

4. L'API HuggingFace a été utilisée (signature du modèle : `facebook/wav2vec2-large-xlsr-53`).

préliminaires (voir les annexes : https://hal.science/hal-04583516) montrent que si le choix de la couche (21 dans nos expérience) a un impact fort sur les résultat d'une sonde linguistique, ce n'est plus le cas avec notre approche.

3.1 Étude des différentes versions d'un même conte populaire

Notre première expérience a pour objectif de déterminer si des variables extra-linguistiques telles que l'acoustique de la pièce et le type de microphone sont présentes dans les représentations neuronales. Pour ce faire, nous utilisons un test ABX pour distinguer les différents enregistrements de la *série de contes populaires* : ces scores sont calculés à partir de triplets composés de deux extraits de 10 s d'un même enregistrement et d'un extrait de 10 s d'un enregistrement différent. [5]

La figure 1 montre que, dans la plupart des cas, avec une longueur d'extrait de 10 s, il est possible de distinguer les différents enregistrements, bien qu'il s'agisse toujours du même locuteur racontant la même histoire : hormis quelques rares exceptions, qui seront abordées plus loin, les scores rapportés sont largement supérieurs à 50 %. De plus, les scores sur la diagonale, correspondant à des tests où tous les extraits proviennent du même enregistrement, sont tous proches de 50 %. Cela indique clairement que les différences constatées dans les autres tests ABX ne sont pas dues au contenu linguistique (les mots prononcés), mais plutôt à la configuration acoustique et suggère que les représentations neuronales capturent bien plus que les informations linguistiques nécessaires à la compréhension de la parole : elles encodent également des informations liées aux conditions d'enregistrement.

FIGURE 1 – Scores ABX pour l'étude de la *série du conte populaire*. Taille des extraits : 10 s.

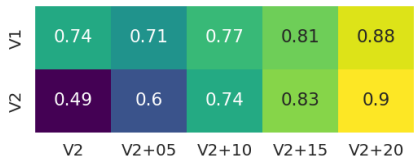


FIGURE 2 – Reproduction de la réverbération de la pièce V_1 avec une réverbération artificielle appliquée sur V_2. Taille des extraits : 5 s.

Une mise en perspective de ces observations avec les conditions d'enregistrement (rendue possible par une connaissance fine des enregistrements) permet de mieux comprendre les informations capturées (ou non) par les représentations. Ainsi, la comparaison de V_1, V_2 et V_3 montre que les représentations de V_2 et V_3 ne sont que peu discernables entre elles (score ABX de 0,67) mais peuvent être distinguées de celles de V_1 (scores de 0,79 et 0,81). La principale différence entre ces trois enregistrements est

5. Les résultats pour d'autres durées d'extraits sont rapportés en annexe (voir https://hal.science/hal-04583516).

liée au lieu d'enregistrement : V_2 et V_3 ont été enregistrés dans un lieu moins réverbérant que le lieu où V_1 a été enregistré. Pour confirmer l'influence de ce paramètre, nous avons réalisé une expérience de contrôle en ajoutant artificiellement de la réverbération [6] aux enregistrements V_2 et en mesurant le score ABX entre les enregistrements V_1 et V_2 modifiés. La figure 2 montre l'évolution du score ABX en fonction de la quantité de réverbération ajoutée. Lorsque la quantité de réverbération dans V_2 augmente, le score ABX diminue d'abord avant d'augmenter à nouveau. Cela signifie que V_1 est plus proche de V_2 avec 5 % de réverbération, ce qui suggère une relation de causalité entre la quantité de réverbération et le degré de proximité entre les enregistrements de ce lot.

Une seconde comparaison intéressante est celle entre les différentes versions des enregistrements V_6 et V_7. Ces enregistrements ont été réalisés avec un micro casque (enregistrements étiquetés h pour *headset*) et un micro à main inséré dans un support placé sur une table (enregistrements étiquetés t pour *table*). La figure 1 montre que les représentations peuvent être distinguées avec précision sur le type de microphone. Par exemple, les scores ABX entre $V_{6,h}$ et $V_{6,t}$ sont parmi les plus élevés de notre expérience, alors que pour deux enregistrements différents réalisés avec le même microphone (c'est-à-dire $V_{6,h}$-$V_{7,h}$ et $V_{6,t}$-$V_{7,t}$), les scores ABX ne sont que légèrement meilleurs que les scores pour le même enregistrement. Cela suggère que les représentations issues d'extraits de 10 s dépendent fortement du microphone utilisé : deux vecteurs représentant le même signal audio mais issus de microphones différents sont plus dissemblables que ceux représentant deux signaux audio différents enregistrés par le même microphone.

La figure 1 fait également ressortir une similitude inattendue entre les enregistrements V_4 et V_5. Le score ABX entre ces deux enregistrements n'est que de 54 %, alors qu'il n'est jamais inférieur à 71 % entre V_4, V_5 et toutes les autres paires. Or, V_4 et V_5 sont les seuls enregistrements dans lesquels le récepteur était un locuteur de la communauté linguistique. Cela ressemble à un cas d'adaptation linguistique (Piazza *et al.*, 2022) et suggère qu'il serait possible de générer automatiquement des hypothèses sur le contexte d'énonciation, par exemple, sur la base des méta-données.

Dans cette série d'expériences, nos observations sont plus visibles avec des extraits de 10 s. Cela semble être le réglage approprié pour mettre en évidence des différences au niveau de l'acoustique globale. Cette taille d'extrait semble aussi convenir pour révéler les différences au niveau du rythme de parole, de la prosodie. D'autres expériences sont néanmoins nécessaires pour confirmer cela.

3.2 Étude de différents répertoires de chansons

Le but de cette expérience est d'explorer si les paramètres d'extraction qui fonctionnent le mieux dans l'expérience précédente permettent ou non d'explorer les représentations en fonction des caractéristiques de voix du locuteur ou de la locutrice. Plusieurs enregistrements d'une chanteuse professionnelle Naxi sont comparés les uns aux autres : une chanson dans le style « Alili », deux dans le style « Guqi », une dans le style « Wo Menda », et un récit. Les chansons contenaient à l'origine une introduction non chantée qui a été supprimée pour les comparaisons, sauf pour la chanson de style « Alili », qui est pour moitié du texte et pour moitié une chanson.

La figure 3 montre que toutes les chansons se distinguent fortement du récit, à l'exception de l'enregistrement « Alili », qui est mi-texte mi-chanson. Il est intéressant de noter que celui-ci ne correspond ni aux chansons ni au récit : il se situe à mi-chemin entre les deux. Quant aux deux chansons dans le style « Guqi », elles présentent le score ABX le plus bas (0,57) même si leur contenu

6. Nous utilisons Audacity pour ajouter 5, 10, 15 ou 20 % de réverbération.

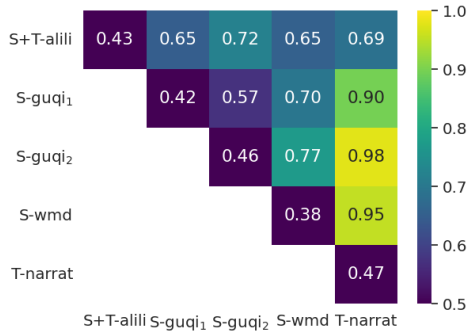


FIGURE 3 – Scores ABX pour les comparaisons entre les différents genres (T=texte, S=chanson). Des chansons dans trois styles différents sont interprétées par une chanteuse professionnelle Naxi. Taille des extraits : 10 s.

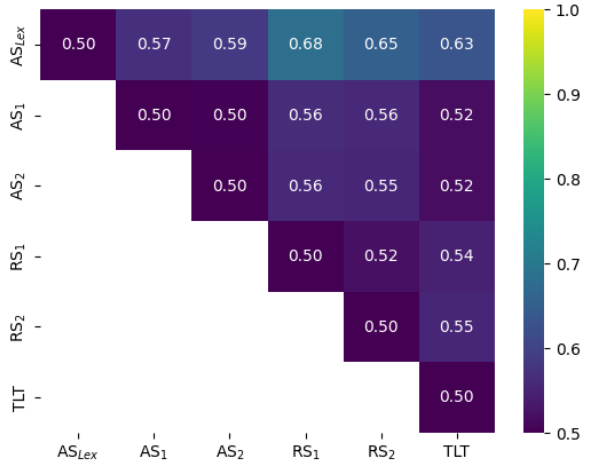


FIGURE 4 – Scores ABX pour les comparaisons entre les éléments de la *série phonétique*. La locutrice AS a trois enregistrements (AS_1, AS_2, AS_{Lex}), RS en a deux (RS_1, RS_2) et TLT en a un. Taille des extraits : 1 s.

linguistique est différent, ce qui suggère que le style de la chanson peut être détecté.

Ces résultats suggèrent que les propriétés de la voix sont présentes dans les représentations, puisque nous pouvons distinguer entre une narration et différents styles de chansons pour un même locuteur, et même regrouper par style de chanson. Ces résultats fournissent des pistes pour de futures études visant à utiliser des modèles neuronaux pour réaliser des études prosodiques.

3.3 Étude d'un corpus de phonétique

S'il est raisonnable de penser que deux phrases au contenu linguistique différent dans des conditions parfaitement contrôlées apparaîtront comme différentes lorsqu'elles seront soumises à un test ABX, la réponse n'est pas immédiate lorsqu'il s'agit d'un enregistrement entier. Il n'est pas non plus évident que deux phrases différentes prononcées par deux locuteurs·rices différent·e·s soient distinguées uniquement en raison d'une différence de contenu linguistique : l'identité du locuteur ou de la locutrice agit comme un facteur de confusion.

L'objectif de cette expérience est de comparer des données présentant des différences d'ordre « phonétique », mais où le paramètre de l'identité de la locutrice varie. Pour ce faire, nous nous appuyons sur un corpus phonétique enregistré de manière contrôlée, où chaque locutrice a reçu les mêmes instructions. Ces enregistrements ont le même contenu ($AS_{1,2}$, $RS_{1,2}$, TLT). Un enregistrement a un contenu différent (AS_{lex}). Les scores sont calculés à partir de triplets composés de deux extraits de 1 s issus du même enregistrement et d'un extrait issu d'un enregistrement différent. [7]

La figure 4 montre qu'avec une longueur de 1 s, il est difficile de distinguer les différents enregistrements des mêmes phrases (c.-à-d. issus du même corpus, mais prononcé par des locutrices différentes, ou simplement différentes répétitions d'un même enregistrement), et ce même lorsque les locutrices

7. Les résultats obtenus pour d'autres tailles d'extraits sont présentés en annexe (voir https://hal.science/hal-04583516).

diffèrent. l'écart à locutrice fixe varie de 0,0 à 0,02, tandis que, tandis que l'écart « cross-locutrices » varie de 0,02 à 0,06. Cela suggère que même avec des extraits d'1 s, l'identité de la locutrice est toujours reflétée, mais d'une manière difficilement repérable pour cette taille d'extrait. En revanche, l'information locuteur·rice ressort extrêmement clairement pour une taille d'extrait de 10 s (voir https://hal.science/hal-04583516), ce qui suggère que les représentations neuronales, pour de petits extraits, « centrifugent » mieux les informations extra-linguistiques. Cette observation n'est pas surprenante étant donné la façon dont les modèles sont pré-entraînés (Baevski *et al.*, 2020), et elle constitue une amorce intéressante pour la deuxième partie de l'analyse, qui consiste à comparer ces enregistrements de phrases identiques à un autre enregistrement avec des phrases différentes.

Les résultats de la première ligne de la Figure 4, par les différences observées relativement aux autres lignes de la figure, suggèrent que les tests ABX révèlent des différences lorsque le contenu linguistique diffère. Ces différences viennent vraisemblablement s'additionner à des différences d'identité de locuteur·rice, mais celles-ci sont suffisamment réduites pour laisser apparaître les différences au niveau du contenu.

Dans cette étude, les scores ABX sont calculés en moyenne sur l'ensemble d'un enregistrement. Pour les différences phonétiques, il serait intéressant de pouvoir effectuer des comparaisons par phrase, mais cela reviendrait à s'écarter d'une approche entièrement non supervisée.

4 Discussions et conclusions

Entreprendre de comparer des représentations vectorielles de parole, en mode non-supervisé, s'apparente à une gageure tant la parole est multi-factorielle et sujette à variation. Nous avons adopté une méthode expérimentale pour soumettre un modèle donné à différentes expériences avec des variables de test, en contrôlant par ailleurs un certain nombre de paramètres. Il faut tout d'abord mentionner que la connaissance des données étudiées et de leurs métadonnées a été déterminante car elle nous a permis d'obtenir des résultats dans un contexte où les corpus sont *de facto* de petite taille puisqu'ils concernent des langues peu documentées, et donc difficilement exploitables en TAL. Cette étude nous a permis dans un premier temps de mieux comprendre le degré d'abstraction des représentations de la parole issues de grands modèles de langage.

Ainsi, nous avons montré que le contenu des représentations n'était pas exclusivement centré sur le contenu linguistique dans la mesure où, en particulier lorsque les représentations encodent plusieurs secondes d'audio, il nous a été possible de corréler les résultats des tests ABX avec des informations relatives à l'acoustique de la pièce, le type de microphone, et le type de discours. Nous voyons donc que la méthode employée pourrait servir pour distinguer des enregistrements en fonction d'un grand nombre de critères extra-linguistiques.

Enfin, lors de l'étude de représentations d'extraits audio de petite taille, nous avons constaté que le poids des facteurs extra-linguistiques diminuait dans les tests ABX, et que des critères phonétiques/segmentaux prenaient au contraire en importance, ce qui est encourageant dans la perspective de comparaison typologiques de langues voisines entre elles, à l'aide d'approches non-supervisées.

L'utilisation de méthodes non-supervisées, notamment faisant intervenir des mesures telles que la distance cosinus, trouvent des applications pour l'amélioration des méthodes de reconnaissance automatique pour les langues rares (San *et al.*, 2024). Notre communication permet de mieux comprendre le lien entre l'audio et ses représentations vectorielles.

Références

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J. *et al.* (2021). XLS-R : Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv :2111.09296*.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BELINKOV Y. & GLASS J. (2019). Analysis methods in neural language processing : A survey. *Transactions of the Association for Computational Linguistics*, **7**, 49–72.

CARLIN M. A., THOMAS S., JANSEN A. & HERMANSKY H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

CASTELLENGO M. (2016). *Ecoute musicale et acoustique*. Paris : Eyrolles.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, **abs/2006.13979**.

FOULKES P. (2010). Exploring social-indexical knowledge : A long past but a short history. *Laboratory Phonology*, **1**(1), 5–39.

LI Y., BELL P. & LAI C. (2022). Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7362–7366 : IEEE.

LI Y., MOHAMIED Y., BELL P. & LAI C. (2023). Exploration of a self-supervised speech model : A study on emotional corpora. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 868–875 : IEEE.

MACAIRE C., WISNIEWSKI G., GUILLAUME S., GALLIOT B., JACQUES G., MICHAUD A., ROSSATO S., NGUYÊN M.-C. & FILY M. (2021). Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Journées GDR LIFT 2021, Grenoble, France. HAL : halshs-03475443.

MICHAUD A. (2017). *Tone in Yongning Na : lexical tones and morphotonology*. Volume 13 de Studies in Diversity Linguistics. Berlin : Language Science Press. 10.5281/zenodo.439004.

MICHAUD A., JACQUES G. & RANKIN R. L. (2012). Historical transfer of nasality between consonantal onset and vowel : from c to v or from v to c ? *Diachronica*, **29**(2), 201–230.

MICHAUD A. & LATAMI D. (2011). A description of endangered phonemic oppositions in Mosuo (Yongning Na). In T. DE GRAAF, X. SHIXUAN & C. BRASSETT, Éds., *Issues of language endangerment*, p. 55–71. Beijing : Intellectual Property Publishing House.

PASAD A., CHOU J.-C. & LIVESCU K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 914–921 : IEEE.

PASAD A., SHI B. & LIVESCU K. (2023). Comparative layer-wise analysis of self-supervised speech models.

PIAZZA G., MARTIN C. D. & KALASHNIKOVA M. (2022). The acoustic features and didactic function of foreigner-directed speech : A scoping review. *Journal of Speech, Language, and Hearing Research*, **65**(8), 2896–2918.

SAN N., PARASKEVOPOULOS G., ARORA A., HE X., KAUR P., ADAMS O. & JURAFSKY D. (2024). Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association*, p. 1–5.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3), 379–423. DOI : 10.1002/j.1538-7305.1948.tb01338.x.

ZHAO S. (2022). *Looking for a disappearing voice : place making, place-belongingness, and Naxi language vitality in Lijiang Ancient Town*. Thèse de doctorat, Massey University, Wellington, New Zealand.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Perception et production des clusters en position initiale par des sinophones : le rôle du Principe de Sonorité Séquentielle

Xuejing Chen[1] Pierre Hallé[1] Rachid Ridouane[1]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 Rue des Irlandais, 75005 Paris, France
xuejing.chen@sorbonne-nouvelle.fr, pierre.halle@sorbonne-nouvelle.fr, ridouane.rachid@sorbonne-nouvelle.fr

RESUME

Dans deux expériences avec des sujets sinophones, nous avons examiné le rôle du Principe de Sonorité Séquentielle (PSS) dans la perception et la production des clusters en position initiale. Dans la première expérience, nous avons évalué la discrimination de contrastes C1C2-C1əC2 avec 3 types de profil de sonorité C1C2 : montant, plateau, descendant. Nos résultats montrent que les C1C2 moins marqués selon le PSS induisent une meilleure discrimination, attribuable à une réparation perceptive moindre pour ce type de séquences. Ces résultats sont en accord avec les résultats de l'expérience d'imitation où la production d'éléments vocaliques est moins fréquente pour les C1C2 moins marqués. L'effet induit par le PSS est plus important en production qu'en perception, suggérant un effet indépendant du PSS en production. Par ailleurs, les propriétés acoustiques des éléments vocaliques produits suggèrent qu'ils sont d'autant plus ciblés que les clusters à imiter sont marqués.

ABSTRACT

Perception and production of word-initial clusters by Chinese speakers: The role of the Sonority Sequencing Principle.

In two experiments with Chinese speakers, we examined the role of the Sonority Sequencing Principle (SSP) in the perception and production of clusters in word-initial position. In the first experiment, we assessed discrimination of C1C2-C1əC2 contrasts with 3 types of sonority profiles for C1C2: rising, plateau, and falling. Our results show that less marked C1C2 sequences in terms of SSP induce better discrimination, attributable to less perceptual repair for this type of sequence. These results are congruent with those of the imitation experiment in which we observed that, for less marked C1C2 sequences, production of vocalic elements is less frequent. The effect induced by SSP was greater in production than in perception, suggesting an independent effect of SSP in production. In addition, the acoustic properties of the inserted vocalic elements suggest that they are all the more targeted that the clusters to be imitated are marked.

MOTS-CLES : Principe de Sonorité Séquentielle, production et perception des clusters, locuteurs sinophones

KEYWORDS : Sonority Sequencing Principle, cluster production and perception, Chinese speakers

1 Introduction

Le Principe de Sonorité Séquentielle (PSS), énoncé entre autres par Clements (1990), propose que les syllabes bien-formées ont un profil de sonorité en cloche (“arch-shaped”). Plus précisément, le profil idéal d’une syllabe a un profil de sonorité séquentiel croissant de façon maximum jusqu’au sommet et décroissant de façon minimum jusqu’à la fin de la syllabe. En ce qui concerne les séquences de type #C1C2V, ceci signifie que la séquence la mieux formée a un profil de sonorité montant de C1 à C2 ; tandis que la séquence la moins bien formée a un profil descendant. Ainsi, il existe une hiérarchie d’acceptabilité des profils de sonorité en début de mot : montant > plateau > descendant. Cette hiérarchie est basée sur une échelle de sonorité des segments, la plus courante étant : voyelle > glide > liquide > nasale > obstruante (Clements, 1990). La hiérarchie d’acceptabilité des profils est largement observée dans les langues du monde (Greenberg, 1978). Cependant, il convient de noter que de nombreuses langues admettent des profils qui ne respectent pas le PSS, comme le russe qui admet des profils descendants comme /lp/, ou l’hébreu qui admet des profils “plateau” comme /kp/, etc. (Yin et al., 2023). Ces configurations marquées ont fourni le champ d’investigation qui a permis de démontrer le rôle du PSS en perception.

Les clusters phonotactiquement illégaux ne sont pas perçus fidèlement. Comme de nombreuses études l’ont démontré, ces clusters ont tendance à être “réparés” perceptivement. La réparation la plus courante d’un CC illégal est l’insertion d’une voyelle : CC > CvC (e.g., Dupoux et al., 1999 : *ebzo* > *ebuzo*). Autrement dit, les sujets entendent une voyelle épenthétique à l’intérieur des clusters. Des travaux antérieurs, notamment ceux de Berent et ses collègues, ont montré que la réparation par épenthèse vocalique était modulée par le PSS : la réparation #CC > #CəC est d’autant plus fréquente que CC est mal formé du point de vue du PSS. Par exemple l’incidence de l’épenthèse perceptive, chez les auditeurs anglophones, augmente pour les items suivants : *bnif* > *bdif* > *lbif* (Berent et al., 2007), bien qu’ils soient tous *également* illégaux en anglais. Cette modulation par le PSS semble être universelle (Berent et al. 2007, 2008, 2012, 2016 ; Maïonchi-Pino et al., 2015) et opère dès la naissance (Gomez et al., 2014). Elle est également observée chez les rats (Santolin et al., 2023) : les auteurs postulent l’existence d’un mécanisme biologique universel, un biais attentionnel pour les productions sonores présentant un profil d’intensité “arch-shaped”, fréquentes dans l’environnement naturel, qui serait à l’origine des effets PSS trouvés dans la perception humaine.

Une autre hypothèse suggère que les séquences de consonnes mal formées selon le PSS sont plus difficiles à produire (Redford, 2008). Cependant, les données sur les effets PSS en production de parole sont limitées. Le PSS joue un rôle dans l’acquisition des clusters, tant chez les adultes (Redford, 2008 ; Broselow & Finer, 1991) que chez les enfants (Sprenger-Charolles & Siegel, 1997) : les séquences CC qui ont un profil de sonorité bien formé (i.e., moins marqué) seraient plus faciles à acquérir. Cependant, Davidson (2000) n’a pas trouvé d’effet PSS dans une tâche de production de clusters non natifs chez des sujets anglophones. Cette incohérence dans les résultats sur le rôle du PSS en production est une des motivations de notre étude. Nous proposons d’examiner les possibles effets du PSS dans une tâche d’imitation de séquences #CCV, toutes non-natives, avec des profils de sonorité montant, plateau, et descendant. Les séquences CC seront-elles d’autant plus “réparées” qu’elles sont plus marquées pour le PSS ? Crucialement, la tâche d’imitation implique d’abord la perception des modèles à imiter. Nous nous attendons donc à une première réparation *perceptive*. Si des effets PSS propres à la production existent, ils devraient s'additionner à ceux propres à la perception. D’où la nécessité de réaliser d’abord une expérience de perception pour estimer les taux de réparation perceptive des séquences en fonction de leur profil de sonorité. Ces données de perception seront nécessaires pour interpréter les données d’imitation, et éventuellement conclure à des effets PSS propres à la production.

L'effet du PSS en perception a été observé chez les locuteurs natifs du mandarin, une langue qui bannit tout cluster en début de mot (Zhao & Berent, 2016 ; Chen et al., 2022). Ces sujets réparent #CC en #CəC (ibid.). Les réparations que nous examinerons seront donc ici les insertions vocaliques (probablement une majorité de schwas). Nous avons choisi comme langue des stimuli le tachlhit car cette langue permet des séquences #CC avec des profils montant, plateau, ou descendant. Nous prédisons que les réparations par épenthèse vocalique en perception, mesurées par un test de discrimination AX (cf. Zhao & Berent, 2016) seront conformes au PSS. Si la production induit des effets PSS indépendamment de la perception, le test d'imitation devrait montrer des effets PSS plus importants que ceux observés dans le test de discrimination. En outre, nous examinerons la qualité acoustique des voyelles épenthétiques produites lors des imitations, ce qui pourrait éclairer sur leur caractère intentionnel (épenthèse) ou non (transitionnel).

2 Expérience 1 : discrimination AX

2.1 Méthode

Vingt sinophones natifs ont participé à un test de discrimination AX en ligne, implémenté sur PsyToolkit (Stoet, 2010, 2017). Les contrastes à discriminer comprenaient 6 paires de non-mots C1C2a-C1əC2a, comme indiqué dans le Tableau 1, et les sujets étaient chargés de déterminer si ces stimuli étaient 'identiques' ou 'différents'. Les attaques C1C2 présentaient des profils de sonorité montant, plateau ou descendant (désignés k- ou t-pivot pour les attaques avec /k/ ou /t/ en position C1 pour les profils montant et plateau, ou en position C2 pour les profils descendants). Les items C1əC2a et C1C2a différaient uniquement par la présence ou l'absence d'un schwa entre C1 et C2. Chaque item a été enregistré huit fois par le troisième auteur, phonéticien et locuteur natif du tachlhit. Trois répétitions de chaque item ont été soigneusement sélectionnées comme stimuli utilisés dans l'expérience, garantissant l'homogénéité des durées de [ə] ($\bar{x}$=58.7 ms, σ=16.9), des durées du [a] final ($\bar{x}$=290.6 ms, σ=49.2) et de sa f0 moyenne ($\bar{x}$=168.5 Hz, σ=4.8), tant au niveau inter-contraste qu'intra-contraste.

	Montant		**Plateau**		**Descendant**	
	k-pivot	**t-pivot**	**k-pivot**	**t-pivot**	**k-pivot**	**t-pivot**
C1C2	kla	tla	kpa	tka	lka	lta
C1əC2	kəla	təla	kəpa	təka	ləka	ləta

TABLE 1 : Les items C1C2 et C1əC2 utilisés dans l'expérience AX

L'expérience a été divisée en 2 parties, où les stimuli étaient présentés dans 4 ordres différents : AA, AB, BB, BA (A = C1əC2a, B = C1C2a), totalisant ainsi 48 essais de discrimination (6 contrastes × 4 ordres × 2 parties). Les deux parties se distinguaient par les stimuli utilisés pour chacune des 4 combinaisons : A1A2, A2B3, B1B2, B2A3 pour la première partie et A2A3, A1B2, B2B3, B1A2 pour la deuxième partie (les indices font référence aux numéros de répétition choisis pour un même item). Les sujets ont été invités à répondre aussi rapidement et correctement que possible. Ils disposaient de 2 secondes pour répondre, et le temps de réponse a été mesuré à partir de l'onset du deuxième stimulus.

2.2 Résultats

Les résultats concernant les discriminations correctes et les temps de réponse sont présentés dans la Figure 1. Nous avons analysé ces données avec des modèles linéaires mixtes (régression logistique pour les proportions de réponses correctes). Pour les discriminations correctes, le meilleur modèle inclut les effets fixes Partie (Partie 1, Partie 2), Profil (montant, plateau, descendant) × Pattern (identique : AA, BB, différent : AB, BA), Profil × Pivot (t-pivot, k-pivot), et les effets aléatoires intercepts et pentes sur le Pattern (par sujets). L'interaction Profil × Pattern est significative ($\chi^2(2)=27.91$, $p<.001$). Pour le Pattern 'identique' (i.e., AA ou BB), les réponses correctes ne varient pas en fonction des profils (montant-plateau : $z=-0.78$, $p=0.74$, plateau-descendant : $z=0.45$, $p=0.74$; montant-descendant : $z=-0.34$, $p=0.74$). Elles varient en revanche significativement pour le Pattern 'différent', indiquant des performances décroissantes du profil montant au plateau ($z=2.91$, $p<.001$) et du profil plateau au descendant ($z=5.52$, $p<.001$). L'interaction Profil × Pivot est également significative ($\chi^2(2)=9.24$, $p<.01$). En détail, pour le pivot t, le profil montant et plateau induisent une meilleure performance que le profil descendant (montant-descendant : $z=5.18$, $p<.001$; plateau-descendant : $z=4.27$, $p<.001$). Il n'y a pas de différence significative motivée par le profil de sonorité pour le pivot k.

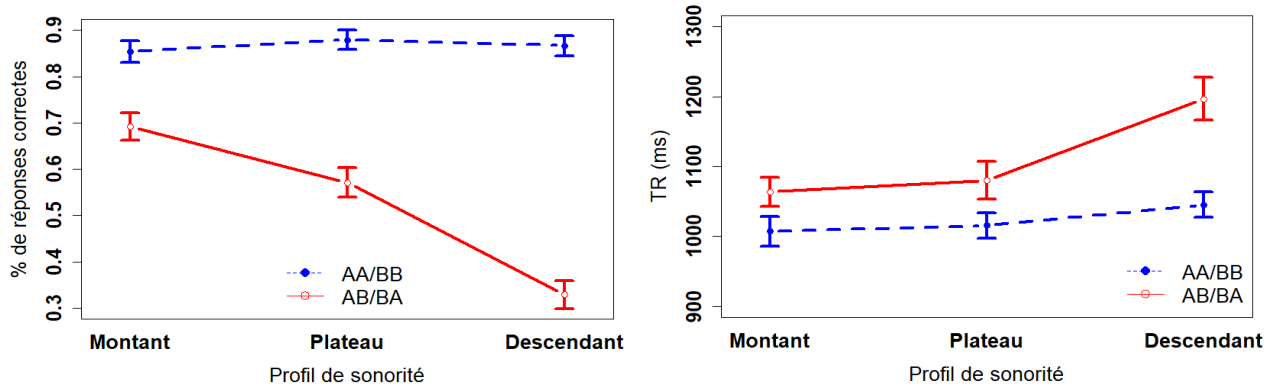


FIGURE 1 : Proportion de réponses correctes (à gauche) et temps de réponse (à droite) selon le profil de sonorité du cluster

Nous avons analysé les temps de réponses correctes (TRs) avec des modèles linéaires mixtes. Le meilleur modèle inclut les facteurs fixes Profil, Pattern, et les effets aléatoires intercepts et pentes sur Pattern (par sujets). Profil est significatif ($F(2, 972.95)=7.11$, $p<.001$). Les sujets sont plus lents à répondre pour descendant comparés à montant ($t(975)=3.65$, $p<.001$) et plateau ($t(976)=2.82$, $p<.01$). Pattern est significatif ($F(1, 17.72)=9.28$, $p<.001$), avec des TRs plus longs pour les patterns AB/BA que pour les AA/BB ($t(19.1)=3.04$, $p<.01$). Pour le Pattern 'différent', les TRs sont plus longs pour descendant que montant ($t(980)=-3.71$, $p<.001$) et plateau ($t(980)=-2.84$, $p<.01$). Quant au Pattern 'identique', il n'y a pas de variation significative.

Pour déterminer si les clusters plus marqués sont plus sujets à la réparation phonologique, il convient de n'examiner que les essais où les C1C2 et C1əC2 sont effectivement contrastés (i.e., avec le Pattern 'différent' : AB ou BA). Les résultats pour le Pattern 'différent' indiquent en effet qu'il y a d'autant plus de réparations perceptives que le profil de sonorité est moins acceptable du point de vue du PSS. A la lumière de ces résultats qui montrent un effet clair du PSS en perception, nous nous posons la question suivante : un effet PSS se manifeste-t-il également en production ? Pour répondre à cette question, nous avons conduit une expérience d'imitation utilisant les mêmes stimuli que de l'expérience 1, dans le but de déterminer si les sinophones produisent des éléments vocaliques entre C1 et C2 et si cela est modulé par le profil de sonorité de C1C2.

3 Expérience 2 : imitation

3.1 Méthode

Les vingt locuteurs natifs du mandarin de l'expérience 1 ont pris part à une expérience d'imitation, où ils étaient invités à reproduire fidèlement les stimuli de l'expérience 1, que nous appellerons "modèles", sans contrainte de temps de réponse. Les participants ont ainsi reçu les mêmes non-mots C1C2a, ou C1əC2a que ceux utilisés dans l'expérience 1, caractérisés donc par un profil de sonorité montant, plateau ou descendant (voir Tableau 1). Deux modèles à imiter ont été sélectionnés pour chaque item à partir des enregistrements effectués par le troisième auteur, comme décrit dans la section 2.1. Les modèles pour les séquences C1C2a et C1əC2a différaient uniquement par la présence d'un schwa d'une durée allant de 42 à 100 ms ($\bar{x}$=71 ms, σ=18.4) dans les séquences C1əC2a. Chaque modèle a été présenté deux fois dans chacune des deux parties constituant l'expérience, totalisant ainsi 96 essais (12 items × 2 modèles × 2 essais × 2 parties). L'ordre de présentation des items a été randomisé de manière différente pour chaque participant.

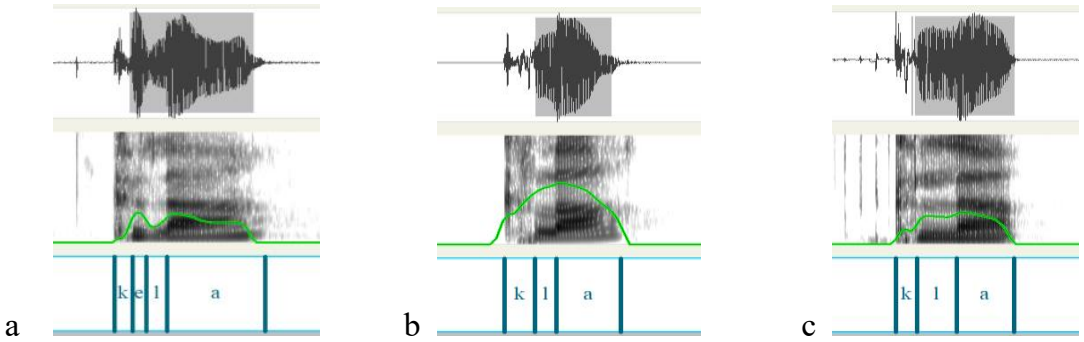


FIGURE 2 : Spectrogrammes et signal de parole de trois imitations d'un modèle /kla/ pour lesquelles la classification est claire (présence de schwa : a ; absence de schwa : b) ou ambiguë (c)

L'expérience a été menée sur la plate-forme SpeechRecorder (Draxler & Jänsch, 2004) dans une chambre calme avec une carte son externe (Komplete Audio 6 MK2) et un micro-casque (AKG Pro Audio C544 L). À chaque essai, les participants, assis devant un ordinateur, recevaient un stimulus modèle une seule fois via le casque et devaient imiter le modèle entendu sans pression de temps. Aucune information orthographique n'était affichée sur l'écran de l'ordinateur pendant les sessions. Les données d'imitation ont été étiquetées et annotées à l'aide de Praat (Boersma & Weenink, 2023). Pour décider de la présence ou de l'absence des schwas entre C1 et C2, nous avons suivi la méthodologie utilisée par Ridouane et Fougeron (2011). Trois critères devaient être remplis pour qu'un schwa soit étiqueté entre le relâchement de C1 et l'onset de C2 : (i) présence d'une période de voisement, (ii) augmentation de l'intensité du signal au relâchement de C1, et (iii) présence d'une structure formantique ou concentration d'énergie dans les régions F2/F3. Nous avons cherché à appliquer cette classification de manière rigoureuse, même si ces critères stricts risquaient d'éliminer à tort certains schwas dans des situations ambiguës (notamment au contact de /l/). La Figure 2 présente des exemples d'imitations de /kla/ réalisées par trois participants, illustrant des cas où l'étiquetage d'un schwa est non-problématique versus douteux.

3.2 Résultats

Pour examiner l'effet du PSS, nous avons calculé la fréquence des éléments vocaliques insérés dans les différents types de clusters. Les résultats sont présentés dans la Figure 3 (à gauche), et ont

été analysés avec des modèles linéaires mixtes de régression logistique. Le meilleur modèle inclut les facteurs fixes Profil (montant, plateau, descendant), Condition (C1C2a, C1əC2a), Partie (partie 1, partie 2) et leurs interactions, et les effets aléatoires intercepts et pentes sur Profil et Condition (par sujets). Comme attendu, les modèles C1əC2a induisent plus d'éléments vocaliques que les C1C2a (z=-6.14, p<.001). Profil est significatif pour les C1C2a ($\chi^2(2)$=102.97, p<.001) et C1əC2a ($\chi^2(2)$=29.23, p<.001) : pour les deux types de modèle, les insertions vocaliques sont plus fréquentes pour descente que plateau (z=3.91 ou 3.96, ps<.001), elles-mêmes plus fréquentes que pour le profil montant (z=2.98 ou 2.76, ps<.01). L'interaction Profil × Condition est significative ($\chi^2(2)$=14.25, p<.001), indiquant un effet PSS plus fort pour les C1C2a que les C1əC2a.

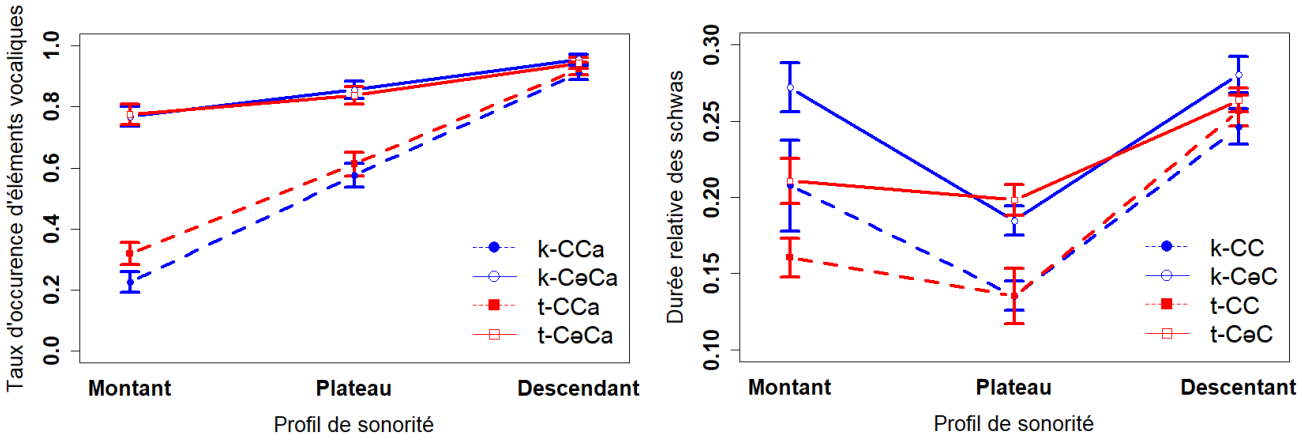


FIGURE 3 : Taux d'insertion de schwas (à gauche) et durée relative des schwas (à droite) en fonction du profil de sonorité pour les items C1C2a et C1əC2a

Les éléments vocaliques produits dans les clusters C1C2a pourraient correspondre à des voyelles intentionnellement insérées par les sujets, reflétant la réparation perceptive pour les clusters mal formés, comme cela a été observé dans l'expérience 1. Une autre possibilité est liée à la difficulté de produire ces séquences, ce qui pourrait se traduire par l'émergence de schwas transitionnels. Selon la Phonologie Articulatoire (Browman & Goldstein, 1992), ces vocoïdes transitionnels résulteraient d'un moindre overlap gestuel entre les consonnes C1 et C2 en raison de certaines contraintes articulatoires. Malgré un petit nombre d'insertions de [i, u], la tendance majoritaire est la production d'un schwa entre C1 et C2 (Figure 4). Afin de mieux comprendre la nature de ces schwas, nous avons mesuré leur durée relative (Figure 3 à droite), définie comme la proportion de leur durée absolue par rapport à celle de la voyelle finale [a], ainsi que les valeurs de F1 et F2 au point médian des schwas pour les modèles C1C2a et C1əC2a (Figure 5).

La durée relative des schwas est plus longue dans les imitations de C1əC2a que de C1C2a (t(1206)=-4.04, p <.001). En détail, pour les profils montant ou plateau, les schwas pour les C1əC2a sont significativement plus longs que pour les C1C2a (montée : t(305)=-3.11, p<.01, plateau : t(362)=-4,84, p<.001). Cette différence s'avère non significative pour le profil descendant (t(535)=-1.96, p=.05). Pour les formes C1C2a, les schwas dans les clusters descendants sont plus longs que dans les montants (t(340)=-4.53, p<.001), qui ont une durée relative supérieure à celle des plateaux (t(218)=2.72, p<.01). La même variation de durée relative des schwas a été observée pour les modèles C1əC2a : ləka/ləta > kəla/təla > kəpa/təka (ləka/ləta - kəla/təla : t(500)=-2.33, p<.05; kəla/təla - kəpa/təka : t(449)=3.95, p<.001). Concernant les formants, le F1 des schwas dans les formes C1C2a est significativement inférieur à celui des schwas dans les formes C1əC2a, tant pour le profil montant (t(304)=-3.55, p <.001) que le profil plateau (t(362)=-3.46, p <.001). En revanche, aucune différence significative n'est observée pour le profil descendant (t(535)=-0.15, p=.88). En termes de F2, les valeurs sont significativement plus élevées

pour les schwas dans les séquences C1C2a comparées aux formes C1əC2a pour le profil descendant ($t(535)=3.46$, $p<.001$). Cet effet n'est cependant pas observé pour les deux autres types de clusters (montant : $t(304)=-1,55$, $p=.12$; plateau : $t(362)=1.37$, $p=.17$).

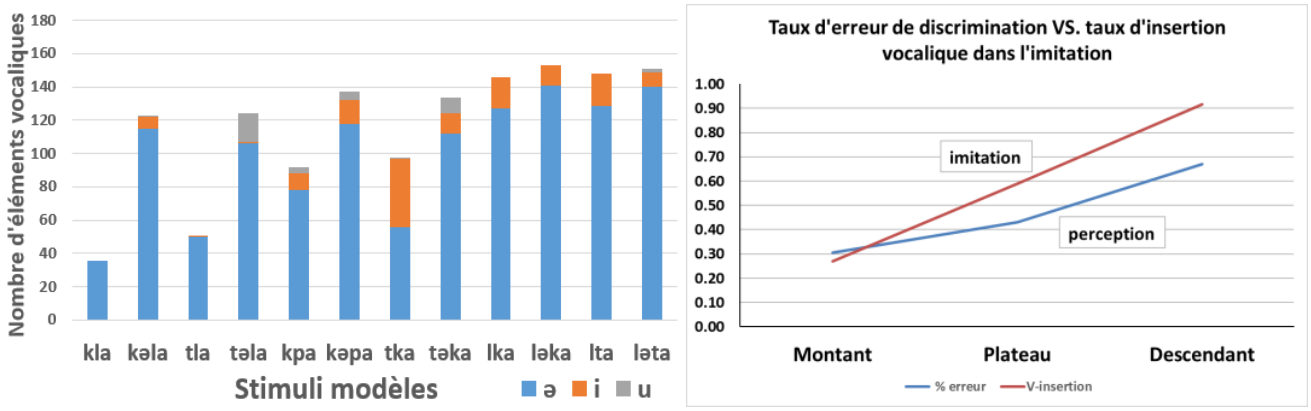


FIGURE 4 : Insertion vocalique dans les imitations pour les items C1C2a et C1əC2a (à gauche) et effets additifs entre la perception et l'imitation (à droite)

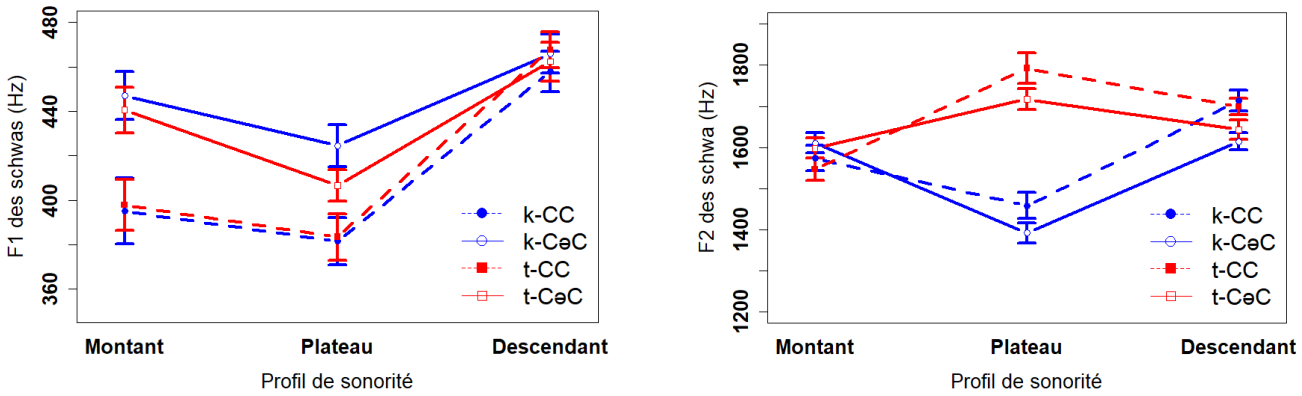


FIGURE 5 : Valeurs de F1 (à gauche) et F2 (à droite) au point médian des schwas en fonction du profil de sonorité pour les items C1C2a et C1əC2a

4 Discussion et conclusion

Dans le cadre de cette étude, nous avons examiné l'influence du PSS sur la perception et la production de clusters consonantiques non natifs en position initiale par des locuteurs sinophones. Les résultats de discrimination entre C1C2a et C1əC2a révèlent un effet notable du PSS sur la perception des clusters, avec une discrimination plus difficile pour les profils les plus marqués, suggérant une augmentation des réparations perceptives par épenthèse vocalique. Les résultats de l'expérience d'imitation montrent que ces effets PSS augmentent nettement en imitation par rapport à la perception pour les profils plateau et plus encore pour les profils descendants. Ceci suggère une additivité des effets PSS, et l'existence d'un effet PSS propre à la production, du moins pour les profils marqués plateau et descendant.

La tâche d'imitation, de par sa nature, englobe des éléments de perception et de production. Dans l'éventualité où aucun effet PSS n'était détecté en production, il serait logique de ne pas observer davantage d'épenthèse en imitation qu'en perception. Néanmoins, si un effet spécifique au PSS est

manifeste en production, nous devrions alors constater une quantité d'épenthèse en imitation supérieure à celle en perception (i.e., taux d'erreurs pour le Pattern 'différent' dans la discrimination AX). C'est exactement ce que nos observations confirment, illustrant ainsi le phénomène d'additivité des effets. De plus, cette additivité semble croître en fonction de la marque des profils de sonorité. Alors que nous ne constatons qu'une faible différence des schwas produits par rapport aux schwas perçus pour les clusters à profil montant (~4%), cette différence augmente substantiellement pour les clusters à profil plateau (16%) et plus encore les profils descendants (25%). Ces résultats suggèrent l'existence d'un effet spécifique au PSS en production, indépendamment de la perception.

Les résultats soulèvent également la question du statut de l'élément vocalique inséré dans les clusters C1C2a : s'agit-il d'un schwa épenthétique (i.e. inséré intentionnellement) ou d'un simple élément transitionnel ? En supposant que cet élément vocalique n'a pas de cible articulatoire propre, on peut émettre l'hypothèse qu'un certain nombre de caractéristiques acoustiques différeront entre ce vocoïde et le schwa canonique produit dans les formes C1əC2a. Autrement dit, si l'élément vocalique dans les formes C1C2a est le simple résultat d'un chevauchement gestuel réduit entre les deux consonnes, il devrait avoir une durée plus courte et des valeurs de F1 et F2 différentes du schwa des C1əC2 (Davidson, 2006 ; Gick & Wilson, 2006 ; Ridouane & Fougeron, 2011). En excluant l'insertion des voyelles [i, u] dans certains clusters, dont la présence suggère des épenthèses plutôt que des éléments purement transitionnels, nous observons la présence d'autres vocoïdes (semblables à des schwas) dans les séquences C1C2a, manifestant des différences significatives selon le profil de sonorité des clusters. Examinons d'abord le profil le plus marqué. Le nombre de vocoïdes produits dans les séquences /lk/ et /lt/ est pratiquement équivalent au nombre de schwas produits dans les séquences /lək/ et /lət/. De manière significative, la durée des deux schwas est quasiment identique, de même que leurs valeurs F1. Pour ces clusters très marqués, il serait donc légitime de conclure que l'élément vocalique présent entre C1 et C2 est bien un schwa épenthétique et non un simple élément transitionnel. Concernant le profil plateau, environ 60% des séquences /kp/ et /tk/ présentent des éléments vocaliques. La durée relative de ces éléments est significativement plus courte comparée au schwa canonique dans les formes correspondantes /kəp/ et /tək/. Le F1 de cet élément vocalique est significativement plus bas mais aucune différence en termes de F2 n'est observée. Les mêmes caractéristiques sont aussi observées pour le vocoïde présent entre C1 et C2 pour le profil montant, même si sa fréquence d'occurrence est moindre (environ 30%). Ces observations suggèrent que cet élément vocalique entre C1 et C2 est un schwa transitionnel. En effet, une durée plus courte et un F1 plus bas peuvent être les conséquences d'une brève période d'ouverture d'un conduit vocal plus fermé entre deux constrictions (Flemming, 2004).

Pour conclure, il est important de souligner que l'analyse du statut de l'élément vocalique dans la production des sujets sinophones est encore préliminaire et nécessite une étude plus approfondie, en particulier avec des données plus variées. Par exemple, il serait intéressant d'étudier comment le lieu d'articulation des consonnes adjacentes affecte le F2, ce qui pourrait potentiellement expliquer pourquoi il est plus élevé pour le profil descendant alors qu'il ne varie pas pour les autres profils. De même, une autre explication des données relatives à la durée pourrait être que les sujets chinois sont sensibles à la différence de durée dans la perception entre schwas illusoires et schwas réels, et qu'ils réussissent à imiter cette différence. En attendant de réaliser ces analyses plus approfondies, le point crucial à retenir est que le PSS affecte non seulement la perception des séquences consonantiques en mandarin, mais également leur production, et que cet effet se manifeste dans la fréquence, la durée et la qualité des schwas produits au sein de ces séquences.

Références

BERENT I., LENNERTZ T. & ROSSELLI M. (2012). Universal phonological restrictions and language specific repairs: Evidence from Spanish. *The Mental Lexicon,* 13, 275–305. DOI : 10.1177/0023830911417804.

BERENT I., LENNERTZ T., JUN J., MORENO M. & SMOLENSKY P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences* 105(14), 5321–5325. DOI : 10.1073/pnas.0801469105.

BERENT I., STERIADE D., LENNERTZ T. & VAKNIN V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition,* 104(3), 591–630. DOI : 10.1016/j.cognition.2006.05.015.

BOERSMA P. & WEENINK D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4.01, http://www.praat.org.

BROSELOW, E. & FINER, D. (1991). Parameter setting in second language phonology and syntax. *Second Language Research*, 7, 35–59.

BROWMAN C. P. & GOLDSTEIN L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4), 155-180. DOI : 10.1159/000261913.

CHEN X., RIDOUANE R. & HALLE P. (2022). Perception des clusters selon leur profil de sonorité : le cas des auditeurs du mandarin confrontés à des clusters russes. Proc. XXXIVe Journées d'Études sur la Parole -- JEP 2022, 183-192. DOI : 10.21437/JEP.2022-20.

CLEMENTS G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Éds.), *Papers in Laboratory Phonology*, p. 283-333. Cambridge: Cambridge University Press.

DAVIDSON L. (2000). Experimentally uncovering hidden strata in English phonology. In L. Gleitman & A.Joshi, Éds., *Proceedings of the 22nd annual conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

DAVIDSON L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics, 34*(1), 104-137. DOI : 1016/j.wocn.2005.03.004.

DRAXLER C., & JANSCH, K. (2004). SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. *International Conference on Language Resources and Evaluation.*

DUPOUX E., KAKEHI K., HIROSE Y., PALLIER C. & MEHLER J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance,* 25(6), 1568–1578. DOI : 10.1037/0096-1523.25.6.1568.

FLEMMING, E. (2004). Contrast and perceptual distinctiveness. In B. HAYES, R. KIRCHNER, & D. STERIADE, Éds., *Phonetically-Based Phonology*. Cambridge: Cambridge University Press.

GICK B. & WILSON I. (2006). Excrescent schwa and vowel laxing: Cross-linguistic: responses to conflicting articulatory targets. In L. GOLDSTEIN, D. WHQLEN & C. BEST, Éds., *Laboratory Phonology,* p. 635-660. Berlin, New York: De Gruyter Mouton. DOI : 10.1515/9783110197211.3.635.

GÓMEZ D., BERENT I., BENAVIDES-VARELA S., BION R., CATTAROSSI L., NESPOR M. & MEHLER J. (2014). Language universals at birth. *Proceedings of the National Academy of Sciences*, 111, 5837-5841. DOI : 10.1073/pnas.1318261111.

GREENBERG J. (1978). Some Generalizations Concerning Initial and Final Consonant Clusters. In E. Moravcsik (Éds.), *Universals of Human Language, v. 2*, p. 243-279. Stanford, CA: Stanford

MAÏONCHI-PINO N., TAKI Y., MAGNAN A., YOKOYAMA S., ÉCALLE J., TAKAHASHI K., HASHIZUME H. & KAWASHIMA R. (2015). Sonority-related markedness drives the misperception of unattested

onset clusters in French listeners. *L'Année psychologique,* 115, 197-222. DOI : 10.4074/S0003503314000086.

REDFORF M. A. (2008). Production constraints on learning novel onset phonotactics. *Cognition, 107*(3). DOI : 785–816. 10.1016/j.cognition.2007.11.014.

RIDOUANE R. & FOUGERON, C. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology*, 2(2), 275-30. DOI : 10.1515/labphon.2011.010.

SANTOLIN C., CRESPO-BOJORQUE P., SEBASTIAN-GALLES N., & Toro, J. M. (2023). Sensitivity to the sonority sequencing principle in rats (Rattus norvegicus). *Scientific reports*, *13*(1), 17036. DOI : 17036. 10.1038/s41598-023-44081-y.

SPRENGER-CHAROLLES L. & SIEGEL L. (1997). A longitudinal study of the effects of syllabic structure on the development of reading and spelling skills in French. *Applied Psycholinguistics*, 18, 485-505. DOI : 10.1017/S014271640001095X.

STOET G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Method*s, 42(4), 1096-1104.

STOET G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology,* 44(1), 24-31.

YIN R., VAN DE WEIJER J. & ROUND E. (2023). Frequent violation of the sonority sequencing principle in hundreds of languages: how often and by which sequences ?. *Linguistic Typology, 27*(2), 381-403.

ZHAO X. & BERENT I. (2016). Universal restrictions on syllable structure: Evidence from mandarin chinese. *Journal of psycholinguistic research,* 45(4), 795–811. DOI : 10.1007/s10936-015-9375-1.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Pertinence des pseudo-mots dans l'évaluation de l'intelligibilité : Effet du nombre ou du caractère non lexical ?

Marie Rebourg[1], Muriel Lalain[1], Alain Ghio[1], Corinne Fredouille[2], Nicolas Fakhry[1, 3], Virginie Woisard[4]

(1) Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
(2) Avignon Université, Laboratoire Informatique d'Avignon, France
(3) Service ORL, APHM, La Conception, Marseille, France
(4) Service ORL, CHU Larrey, UT2J Laboratoire de NeuroPsychoLinguistique, Toulouse, France
marie.rebourg@univ-amu.fr, muriel.lalain@univ-amu.fr, alain.ghio@univ-amu.fr, corinne.fredouille@univ-avignon.fr, nicolas.fakhry@ap-hm.fr, woisard.v@chu-toulouse.fr

RESUME

La perte d'intelligibilité constitue une plainte récurrente des patients traités pour un cancer de la cavité buccale ou de l'oropharynx. La notion d'intelligibilité par son aspect factorielle est complexe à définir, mais aussi, par extension, à évaluer avec précision. Les différents matériaux utilisés dans ces évaluations sont connus pour montrer des effets d'apprentissages imputables aux listes d'items courtes et fermées, présentes dans les batteries de tests classiques. Dans cette étude, nous évaluons l'effet d'apprentissage du matériel linguistique en comparant l'évolution des scores d'intelligibilité calculés à partir de la transcription de mots et de pseudo-mots, présentés en proportion équivalente, soit la répétition de 50 mots vs de 52 pseudo-mots. Nos résultats montrent un effet d'apprentissage des pseudo-mots lorsqu'ils sont répétés, dans les mêmes proportions que celui observé sur les mots. Ainsi, c'est la quantité de pseudo-mots qui permet de neutraliser l'effet d'apprentissage du matériel linguistique dans une évaluation de l'intelligibilité.

ABSTRACT

Relevance of pseudowords in the assessment of intelligibility: Effect of number or non-lexical character ?

Loss of intelligibility is a recurring complaint among patients treated for cancer of the oral cavity or oropharynx. The notion of intelligibility due to its multifactorial aspect is complex to define, but also, by extension, to evaluate with precision. The different materials used in these evaluations are known to show learning effects attributable to the short and closed lists of items, present in traditional test batteries. In this study, we evaluate the learning effect of linguistic material by comparing the evolution of intelligibility scores calculated from the transcription of words and pseudo-words, presented in equivalent proportion, i.e. the repetition of 50 words vs. 52 pseudowords. Our results show a learning effect on pseudowords when they are repeated, in the same proportions as that observed on words. Thus, it is the quantity of pseudo-words which makes it possible to neutralize the learning effect of linguistic material in an evaluation of intelligibility.

MOTS-CLÉS : Phonétique Clinique, Intelligibilité, Trouble de la Production de la Parole, Cancer VADS

KEYWORDS : Clinical phonetic, Intelligibility, Speech disorders, Head and Neck cancer

1 L'intelligibilité

Les traitements dont bénéficient les patients dans le cadre de la prise en charge des cancers de la cavité buccale et de l'oropharynx sont connus pour porter atteinte aux fonctionnements anatomiques de l'appareil phonatoire. Il en résulte une perte d'intelligibilité, qui constitue une plainte récurrente chez ces patients, en perte d'autonomie communicationnelle. L'évaluation de cette composante linguistique multidimensionnelle est essentielle dans le parcours de soin du patient, puisqu'elle permet de mesurer le handicap à la communication en évaluant les composantes dégradées/préservées et de mesurer l'effet du traitement préalablement établi.

L'intelligibilité, par sa dimension multiple et factorielle, représente une notion complexe à définir et à circonscrire. Plusieurs définitions et approches ont été proposées, telles que « le degré de précision avec lequel un message est compris par un auditeur » (Yorkston, Dowden et Beukelman, 1992). Toutefois, ces définitions intègrent la notion de compréhension, entrainant une confusion entre les différents niveaux d'informations linguistiques qui doivent être ciblés par la notion d'intelligibilité. Le concept de compréhension réfère à l'intégration de l'ensemble des informations pertinentes indépendantes et dépendantes du signal permettant de comprendre, de saisir le sens d'un énoncé oral en situation de communication (Lindblom, 1990). Cette acception soutient ainsi l'idée selon laquelle les processus de bas niveaux supportent des informations « dépendantes du signal » alors que les informations « indépendantes du signal » sont liées aux processus de haut niveau (Hustad, Jones et Dailey, 2003). Ainsi, la définition de l'intelligibilité doit être circonscrite au bas niveau, comme proposé par Hustad (2008) selon laquelle l'intelligibilité fait référence à « la façon dont le signal acoustique d'un locuteur peut être récupéré avec précision par un auditeur ». Cette approche permet d'évacuer la dimension de compréhensibilité (haut + bas niveaux) tout en recentrant le concept d'intelligibilité autour des informations acoustico-phonétiques, dépendantes du signal acoustique (bas niveau).

L'évaluation, la mesure, de l'intelligibilité doit donc être rigoureusement contrôlée de façon à déterminer quel matériel linguistique permet d'évaluer l'intelligibilité de quel niveau linguistique. En d'autres termes, la sélection du matériel linguistique utilisé pour évaluer l'intelligibilité doit être conditionnée par la définition précise du niveau linguistique visé par la mesure.

1.1 Mesurer l'intelligibilité – études et contexte clinique

Les études récentes concernant la mesure d'intelligibilité portent tant sur le matériel linguistique que sur la méthode d'évaluation, mais aussi sur la méthode de calcul choisie. Elles s'accordent sur la nécessité de mesurer précisément quelle composante linguistique est évaluée en fonction des différents matériaux linguistiques. Dans la lignée de Ganzeb*oom et* al. (2016), *Xue et* al. (2021) suggèrent que différentes constructions de l'intelligibilité sont nécessairement reflétées par les mesures d'intelligibilité selon qu'elles emploient des échelles analogiques ou des évaluations par transcriptions, mais aussi selon le matériel linguistique employé dans cette évaluation, distinguant les phrases, les mots et les pseudo-mots. Cette proposition ancre par conséquent l'aspect multi-dimensionnel de l'intelligibilité.

Ainsi, les études considérant des ensembles de phrases, questionnant la pertinence de ce matériel linguistique pour évaluer l'intelligibilité, ont montré que l'intégration du contexte menait davantage à une évaluation de la compréhension, puisque celles-ci intègrent les informations de haut et bas niveaux (Yorkston, Strand et Kennedy, 1996; Ganzeb*oom et* al., 2016; Xue, R. Ho*ut, et* al., 2021). Cela représente donc une évaluation perceptive du déficit global. De plus, en comparaison avec les phrases, les listes de mots se montrent comme de meilleurs candidats (*Xue et* al., 2023). Néanmoins, les listes de mots disponibles dans les batteries d'évaluation clinique

(BECD (Auzou et Rolland-Monnoury, 2006), FDA2 (Bl*anc et* al., 2014)) se montrent facilement mémorisables de par leur caractère court et fermé. La répétition de la tâche par le clinicien et un nombre d'items restreints favorisent les effets d'apprentissages et de restauration lexicale. Ceux-ci sont donc inhérents aux matériaux linguistiques de nature lexicale. De plus, ces listes se montrent peu contrôlables au regard de l'occurrence et de la position des phonèmes au sein des items. Restent donc les pseudo-mots, items non lexicaux, n'étant pas porteurs de sens et suivant les règles phonotactiques de la langue, qui se positionnent comme item candidat idéal et pertinent. Ils permettent de s'affranchir de l'intégration des informations de haut niveau, au profit des unités de bas niveau, forçant ainsi le décodage acoustico-phonétique des sons de parole. Ils peuvent être générés de façon automatique, selon un ensemble de contraintes administrant l'occurrence et la position des phonèmes. Ainsi, le seul effet de restauration qui pourrait être attendu avec ce type de matériel linguistique concernerait le niveau du phonème, soit une reconstruction basée sur les connaissances phonologiques d'organisation des unités phonémiques de la langue par l'auditeur.

Concernant les différentes méthodes d'attribution et de calcul des scores, elles peuvent reposer sur des échelles de mesure graduée (Lickert), des échelles analogiques visuelles, ou des transcriptions. Les études les plus récentes montrent que les scores les plus fiables sont obtenus par transcriptions orthographiques (*Xue et* al., 2023). Les différentes échelles prennent en compte des paramètres subjectifs, qui, mêmes s'ils sont précisés, restent empreints d'une grande variabilité, dépendant de l'auditeur-évaluateur. De plus, ces scores perceptifs globaux se montrent peu pertinents dans l'identification des composantes dégradées/préservées et des niveaux linguistiques touchés par ces altérations. Ainsi, les scores calculés de façon automatique, selon une méthodologie précise, sont perçus comme plus fiables et plus objectifs.

En ce sens, une tâche de Décodage acoustico-phonétique (ci-après DAP) permettant de collecter des transcriptions orthographiques, basée sur la perception de pseudo-mots produits, doublée d'une méthode de calcul innovante, basée sur la théorie des traits distinctifs, a été développée dans le cadre du projet de recherche C2SI (Carcinologic Speech Severity Index, Institut National pour le Cancer n°2014-135) (Astésano *et al.*, 2018; Lalain *et al.*, 2020; Woisard *et al.*, 2021). Celle-ci permet le calcul d'un score analytique en termes de distance à la cible, en nombre de traits moyens altérés par phonème (Ghio, Lalain, Giusti, *et al.*, 2020). La mise à l'épreuve de cette tâche a montré sa pertinence pour distinguer deux groupes de locuteurs – patient vs contrôle (Ghio *et al.*, 2018). Elle a également montré que le recours aux pseudo-mots permet d'attribuer des scores stables au cours du temps, contrairement à l'utilisation de mots ; ceci suggérant que les effets d'apprentissage du matériel linguistique, rencontrés dans les batteries de tests classiques, peuvent être neutralisés par une grande diversité au sein des listes de stimuli (Rebo*urg et* al., 2019; Lal*ain et* al., 2022). De plus, l'évaluation de la pertinence de cette tâche a également montré que l'effet d'expertise auditive des cliniciens était préservé. En moyenne, les scores calculés à partir des transcriptions de ces auditeurs sont plus bas que ceux des auditeurs naïfs, suggérant qu'ils sont de meilleurs décodeurs (Rebo*urg et* al., 2020). La méthode de calcul employée a également montré sa pertinence pour proposer une mesure fine et fiable représentative du déficit articulatoire et par extension, du handicap communicationnel relatif à la qualité de vie du patient.

Enfin, les études les plus récentes, au regard des différentes mesures proposées pour évaluer l'intelligibilité en contexte clinique, montrent que les mesures portant au niveau du phonème sont pertinentes pour détecter les erreurs articulatoires, dans le cadre des dysarthries (*Xue et* al., 2023). Et donc, par extension, probablement pour apprécier finement les séquelles articulatoires, et leur évolution, dans le cadre de la prise en charge clinique et orthophonique de patients après un cancer de la cavité buccale ou de l'oropharynx.

Dans la présente étude, nous interrogeons les résultats précédemment obtenus (Rebo*urg et* al., 2020) qui montrent que l'utilisation d'un très grand nombre de pseudo-mots permet de neutraliser

les effets d'apprentissage du matériel linguistique couramment utilisé pour évaluer l'intelligibilité. La présente étude questionne l'origine de cet effet : est-il davantage lié au nombre d'items, en termes de quantité, ou aux pseudo-mots en tant qu'unité linguistique non lexicale, soit sa qualité ?

2 Méthodologie

Afin de satisfaire à ces différentes questions de recherche nous avons constitué deux corpus, basés sur la production de mots (M) et de pseudo-mots (PM) isolés. Ces corpus ont été utilisés dans des tests de jugement perceptif de l'intelligibilité, visant à obtenir les transcriptions orthographiques des stimuli perçus. Celles-ci, après phonétisation, permettent le calcul des scores de Déviation Phonologique Perçue (DPP) reflétant le nombre de traits (Jakobson, Fant et Halle, 1952) moyens altérés par phonème (Ghio, Lalain, Giusti, *et al.*, 2020)

2.1 Corpus

Pour mener à bien ces travaux de recherche, deux corpus ont été constitués. (i) Un premier corpus de production de mots isolés. Il comprend les enregistrements de 20 locuteurs : 10 patients traités pour un cancer de la cavité buccale ou de l'oropharynx (Toulouse (Balaguer, 2021)) et 10 sujets contrôles (Aix-en-Provence (Rebourg, 2022)), appariés en âge et en sexe. Ils ont été enregistrés lors de la production de la liste de 50 mots de la BECD (Auzou et Rolland-Monnoury, 2006) couramment utilisée pour évaluer l'intelligibilité en contexte clinique. (ii) Un second corpus, comprenant des productions de pseudo-mots isolés, a été constitué auprès d'un autre groupe de 20 locuteurs. Également, 10 patients traités pour un cancer de la cavité buccale ou de l'oropharynx (Toulouse (Balaguer, 2021)), 10 sujets contrôles (Aix-en-Provence (Rebourg, 2022)), appareillés en âge et en sexe. Ils ont tous été enregistrés pendant la production d'une **seule et unique liste de 52 pseudo-mots**, extraite du matériel linguistique développé dans le cadre du projet C2SI, spécifiquement élaboré pour une tâche perceptive de décodage acoustico-phonétique. Ces enregistrements, traités et préparés pour satisfaire les conditions expérimentales, comptabilisent 1000 stimuli dans le corpus de Mots et 1040 stimuli pour le corpus de Pseudo-mots.

Ces travaux de recherche emploient un protocole expérimental déjà éprouvé (Ghio *et al.*, 2018; Rebourg, 2018, 2022; Rebourg *et al.*, 2019, 2020; Ghio, Lalain, Rebourg, *et al.*, 2020; Lalain *et al.*, 2022). Afin de limiter le nombre de stimuli présentés et transcrits par chaque auditeur, dans le test de perception, les groupes de locuteurs constitutifs des corpus sus-présentés ont été divisés en deux sous-groupes (A et B). Chacun de ces sous-groupes comprend 5 locuteurs patients et 5 contrôles. Ainsi, chacun des corpus (Mots et Pseudo-mots) a été divisé en 2 sous-corpus selon les groupes de locuteurs A et B. Ils comprennent respectivement 500 stimuli Mots liste BECD) et 520 stimuli Pseudo-mots (tâche de DAP, projet C2SI).
Les 2 x 500 stimuli du corpus de mots et les 2 x 520 stimuli du corpus de pseudo-mots ont respectivement été divisés en 2 x 3 listes. Chaque liste de mots (BECD) est constituée de 167 productions et chaque liste de pseudo-mots (DAP) est constituée 174 productions. Soit un total de 6 listes pour chaque corpus et 3 listes pour chaque sous-corpus (groupes de locuteurs A et B). Les listes 1 à 3 de chacun des corpus (Mots et Pseudo-mots) sont produites par le groupe de locuteurs A et les listes 4 à 6 par le groupe de locuteurs B.

2.2 Design expérimental

Dans ce test de perception, les auditeurs écoutent les stimuli produits dans un casque et transcrivent orthographiquement sur un clavier d'ordinateur ce qu'ils entendent. Chaque auditeur écoute et transcrit 3 listes BECD et 3 listes DAP, produites par un groupe de locuteurs (A ou B). Chaque liste est transcrite par 3 auditeurs différents. L'ordre de présentation de chaque liste a été contrôlé de sorte que chaque liste occupe chaque position possible. Nous obtenons donc 3

transcriptions de 3 auditeurs différents dans chacun des 3 ordres de présentation, soit 9 transcriptions par liste.

			T1	T2	T3
Groupes de locuteurs A	3 auditeurs ≠	Mots BECD	1	2	3
		Pseudo-mots	1	2	3
	3 auditeurs ≠	Mots BECD	2	3	1
		Pseudo-mots	2	3	1
	3 auditeurs ≠	Mots BECD	3	1	2
		Pseudo-mots	3	1	2

TABLE 1 : Extrait récapitulatif du design expérimental du test de jugement perceptif par DAP, chaque chiffre des colonnes T1 à T3 correspond aux identifiants des différentes listes.

Ce protocole expérimental a permis de mesurer l'évolution des scores PPD au cours de la répétition de 50 mots et de 52 pseudo-mots; scores calculés à partir des transcriptions des auditeurs dans un test de jugement perceptif de l'intelligibilité par décodage acoustico-phonétique (DAP).

2.3 Test de perception

A travers la base de données de volontaires participant aux expériences scientifiques du Centre d'Expérimentation sur la Parole (CEP) du LPL, nous avons recruté 18 auditeurs naïfs. Tous natifs de langue française, sans problème de vue ou d'audition non corrigés et ayant un bon niveau en orthographe. Chaque participant a débuté la tâche par la transcription des productions de mots isolés, puis l'a poursuivie avec la transcription des productions de pseudo-mots. Un auditeur évalue donc 10 locuteurs (5 patients et 5 contrôles), soit un total de 1020 stimuli (500 mots et 520 pseudo-mots). Les auditeurs ont été dédommagés en ticket Kadeos.
Ils ont reçu la consigne de toujours proposer une transcription, qui soit au plus près de ce qu'ils ont perçu et identifié, en respectant les règles orthographiques du français. Ce test de jugement perceptif de l'intelligibilité a été conduit au CEP (http://cep.lpl-aix.fr/), à l'aide de la station de perception PercEval (An*dré et* al., 2003). Ce design expérimental a été élaboré pour évaluer l'évolution des scores de Déviation Phonologique Perçue au cours du temps de la tâche en fonction du matériel linguistique utilisé (Mots – Pseudo-mots) à parts égales.

2.4 Traitement des données

Afin de pouvoir analyser statistiquement ces données, collectées lors des tests de jugement perceptif de l'intelligibilité par DAP, plusieurs traitements sont nécessaires. Les données brutes, telles que présentées par le logiciel d'expérimentation PercEval (An*dré et* al., 2003), sont des transcriptions orthographiques. Différentes opérations de pré-traitement sont effectuées afin que le format des données soit compatible avec les outils de calcul des scores DPP. Ces transcriptions orthographiques sont phonétisées en deux étapes (LIA-Phon, (Béchet, 2001) et Lexique.org) et sont ainsi compatibles avec la matrice de confusion utilisée pour calculer les scores de Déviation Phonologique Perçue (DPP).
Cette matrice de confusion repose sur l'attribution d'un coût, en termes de distance, basé sur la théorie des traits distinctifs (Jakobson, Fant et Halle, 1952). Elle a été développée par A. Ghio dans le cadre de l'analyse des données DAP (Ghio *et al.*, 2018; Ghio, Lalain, Giusti, *et al.*, 2020) du projet de recherche C2SI (Astésano *et al.*, 2018; Ghio, Lalain, Giusti, *et al.*, 2020). Elle permet le calcul de distance d'édition entre deux chaînes de caractères phonétiques par l'algorithme de Wagner Fischer. Celui-ci intègre la distance de Levenshtein, qui considère trois opérations

d'éditions élémentaires : la suppression, l'insertion ou la substitution d'un caractère. Ceci permet de considérer les altérations sur les deux axes syntagmatique et paradigmatique. En d'autres termes, cette méthode permet d'attribuer un score de distance entre les transcriptions phonétiques des cibles qui devaient être prononcées par les locuteurs et les transcriptions effectives phonétisées des auditeurs.
Le calcul des scores de distances cumulées repose sur la division du score donné par la matrice de confusion par le nombre de caractères de la cible phonétique. Nous obtenons alors des scores de Déviation Phonologique Perçue (DPP) en termes de distance cumulée à la cible, qui représentent le nombre moyen de traits altérés par phonème. De plus, pour chaque auditeur, un numéro a été assigné à chaque item en fonction de l'ordre de la passation, de 1 à 500 pour les mots et de 1 à 520 pour les pseudo-mots. Nous comparons donc l'évolution des scores au cours de la tâche entre deux matériaux linguistiques répétés à occurrences égales, 50 mots et 52 pseudo-mots.

3 Résultats

Cette étude a été menée afin d'évaluer si la neutralisation des effets d'apprentissage (Rebo*urg et* al., 2020) tenait davantage à la qualité du matériel linguistique : Mots vs Pseudo-mots, ou à la quantité de pseudo-mots différents. Pour rappel, dans l'expérience précédente une même liste de 50 mots et 2 listes de 52 pseudo-mots, soit 50 mots répétés 20 fois et 2080 pseudo-mots différents, constituaient les stimuli de l'expérience. Il s'agit ici de confronter la répétition d'une liste de 50 mots à la répétition d'une liste de 52 pseudo-mots.
Une analyse de variance (ANOVA) a été conduite, afin de tester les effets simples entre l'évolution des scores PPD moyen (VD) au cours de la tâche et les différents matériaux linguistiques (Mots vs Pseudo-mots). Comme précédemment (Rebo*urg et* al., 2020; Rebourg, 2022), nos résultats concernant le matériel linguistique lexical (Mots) confirment un effet de l'ordre des items dans la passation ($F(1,498) = 38.72$, $r = 0.004$, $p < 0.001$). Autrement dit, le score moyen baisse significativement au cours de la répétition de la tâche montrant un effet d'apprentissage du matériel linguistique lexical. Les résultats obtenus pour les pseudo-mots montrent également un effet significatif de l'évolution à la baisse du score DPP moyen au cours de la répétition de la tâche ($F(1,518) = 64.78$, $r = 0.006$, $p < 0.001$). Cela suggère qu'un effet d'apprentissage est induit par la répétition d'un même matériau linguistique, quelle que soit sa nature, lexicale ou non lexicale.

Ces résultats sont illustrés par la Figure 1 qui montre une relation affine forte et négative entre les scores moyens d'intelligibilité et la position de l'item dans la passation, pour les Mots et les Pseudo-mots. Cette figure se lit comme suit : en ordonnée les scores de Déviation Phonologique Perçue (PPD) ; en abscisse le numéro attribué à l'item en fonction de son ordre dans la passation (1 à 500 pour les Mots, 1 à 520 pour les Pseudo-mots) ; chaque point représente la moyenne des scores moyens attribués aux items en fonction de leur ordre dans la passation ; les droites de régression linéaires montrent l'évolution globale des scores au cours de la tâche.

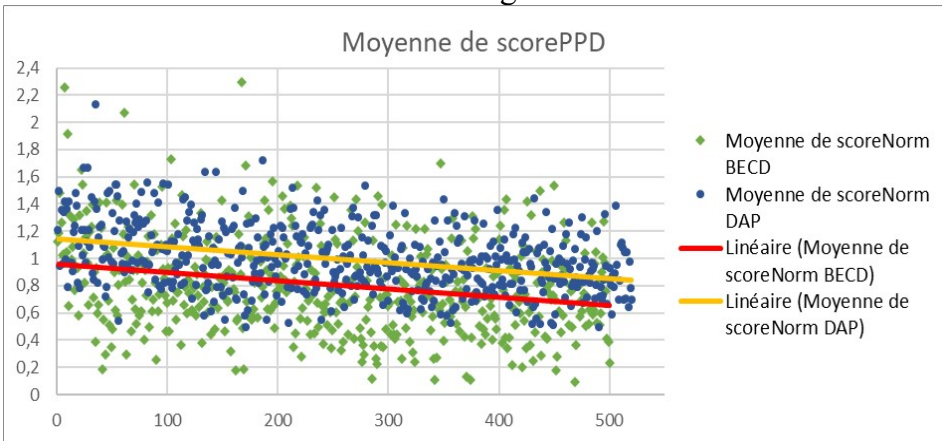


FIGURE 1 : Représentation en nuage de points des scores d'intelligibilité moyen en fonction de la position de l'item dans la passation, du type de matériel linguistique, Mots et Pseudo-Mots

La baisse de scores DPP moyens au cours de la tâche se traduit par un score moyen plus faible à la fin de la tâche, soit au 500/520ème mots/pseudo-mots, par rapport aux premiers stimuli transcrits. Pour les mots, losanges verts, droite de régression linéaire rouge, les scores moyens passent de 0,97 pour la première transcription à 0,64 pour la 500ème transcription, soit un abaissement de 0,33 traits par phonème. Cela représente une baisse de 34% du score moyen pour 500 essais. Pour les pseudo-mots, points bleus, droite de régression linéaire jaune, les scores moyens passent de 1,19 pour la première transcription à 0,86 pour la 520ème transcription, soit un abaissement de 0,33 traits par phonème, égale à celui des mots. Cela représente une baisse de 28% du score moyen. Cet effet de baisse significative des scores moyens représente un effet d'amélioration de 0,00066 trait par phonème par essai, pour les deux matériaux linguistiques. Cela révèle un effet d'apprentissage du matériel linguistique au cours de la tâche.

De plus, les scores moyens PPD calculés pour les Pseudo-mots ($\mu = 0.99$) sont significativement plus élevés ($p < 0.001$) que ceux obtenus avec une évaluation avec les Mots ($\mu = 0.80$). Cela suggère qu'une évaluation basée sur des pseudo-mots est plus stricte qu'une évaluation basée sur des mots de lexique. La taille de cet effet mesurée avec un coefficient d de Cohen ($d = -0.153$) nous indique que cette différence est faible, représentant 0.2 trait moyen d'écart par phonème.

4 Conclusion - discussion

Ces résultats révèlent un effet d'apprentissage du matériel linguistique lors de la répétition d'une tâche comprenant des listes d'items courtes et fermées, quelle que soit la nature du matériel linguistique lexical et non lexical. Mis en perspectives avec les résultats de notre précédente expérience, ces présents résultats montrent que c'est la quantité de pseudo-mots, qui permet de neutraliser cet effet d'apprentissage. En effet, aucun effet d'apprentissage au cours de la tâche n'avait été révélé par la présentation de 2080 pseudo-mots différents, contrairement à celui mis en avant ici, lors de la répétition d'une liste de 52 pseudo-mots. Ces résultats suggèrent que dans le cadre de l'évaluation clinique de la parole, un grand répertoire de pseudo-mots constitue un matériel linguistique pertinent pour neutraliser les effets d'apprentissage des listes d'items, biais fréquent des batteries de tests classiques d'évaluation de l'intelligibilité. De plus, ces résultats contribuent à valider les critères d'objectivité et de pertinence de la tâche DAP (C2SI).

Nos résultats montrent une différence significative de scores moyens entre les deux matériaux linguistiques, plus élevée pour les Pseudo-mots que pour les Mots, de 0.2 trait moyen d'écart par phonème. Cela suggère que la nouveauté de ces formes linguistiques non lexicales, jamais perçues auparavant, contrecarre l'effet de restauration lexicale. Les scores, en moyenne moins élevés, obtenus avec des mots peuvent s'expliquer par le recours aux informations de haut niveau. L'auditeur fait appel à son lexique mental dans lequel il a stocké ces mots de lexique, probablement déjà rencontrés au cours de son expérience linguistique, facilitant ainsi un effet de restauration lexicale, susceptible d'être délétère dans le cadre de l'évaluation clinique de la parole.

Néanmoins, nos résultats montrent que les scores moyens baissent au cours de la tâche pour les deux matériaux linguistiques. Et ce dans une amplitude exactement identique, 0,33 traits entre le premier et le dernier stimuli transcrit, soit en moyenne 0,00066 trait par phonème, par stimuli transcrit. Cette baisse des scores moyens proportionnellement identique entre les Mots et les Pseudo-mots suggère que les mêmes mécanismes se mettent en place lors de la perception d'unités linguistiques non lexicales telles que les pseudo-mots, analogues à ceux de la perception de mots

du lexique. L'auditeur tient à jour ses connaissances linguistiques et se montre capable d'intégrer, dans son stock de connaissances lexicales, des unités phoniques auxquelles aucun sens n'est attribué. Il réinvestit ses connaissances de haut niveau en ajoutant des stimuli à son répertoire de formes linguistiques perçues. Ceci expliquerait l'effet d'apprentissage révélé dans notre expérience. De plus, cet effet d'apprentissage des pseudo-mots questionne l'emploi d'un très grand répertoire de mots pour une évaluation perceptive. En effet, Si seul un grand nombre d'items évite les biais perceptifs, il peut être tentant de choisir des mots qui sont des éléments linguistiques familiers contrairement aux pseudo-mots considérés comme plus artificiels. Ainsi, on pourrait envisager un recours à tous les mots CVCV (4087), CCVCV (977), CVCCV (1648), CCVCCV (150) présents dans la base lexique.org. Cela représente 6862 mots, ce qui pourrait être une taille suffisante. Cependant, avec un tel dictionnaire, il serait très difficile d'obtenir des listes équivalentes de 50 mots en termes de contenu et d'équilibre phonétique, et ne permettraient pas de contrôler les effets de restauration lexicale. D'autre part, il serait aussi compliqué de maitriser des contraintes de fréquence d'apparition (mots rares vs fréquents), alors que cet aspect est crucial dans la perception du lexique (Vitevitch et Luce, 1999). Seul le recours aux pseudo mots permet l'obtention de listes phonétiquement équivalentes (contrôle de l'occurrence et de la position des phonèmes) et où la contrainte de la fréquence d'occurrence est annihilée par la nature inédite de ces unités.

De plus, la qualité des pseudo-mots, en tant qu'items linguistiques non lexicaux, permet de contrôler les effets de restauration lexicale en orientant l'auditeur dans le sens de la restauration phonémique. Les connaissances implicites de l'auditeur à propos de règles phonologiques de sa langue, en tant que système régi par des ensembles de règles d'ordonnancement et d'organisation des sons, lui confèrent cette capacité. De plus les pseudo-mots, en tant qu'unités linguistiques non lexicales, présentent l'avantage de pouvoir être générés en très grande quantité, neutralisant les effets d'apprentissage et de ne jamais avoir été entendus auparavant, neutralisant les effets de restauration lexicale, en centrant la tâche de l'auditeur sur le décodage des sons de parole, soit du décodage acoustico-phonétique. Ces qualités suggèrent qu'ils constituent des unités pertinentes pour l'évaluation des composantes phonétiques articulatoires et acoustiques préservées/dégradées, dans le cadre clinique.

L'ensemble de ces résultats souligne également la complexité de l'ensemble des mécanismes de perception de la parole. Ainsi, la compréhension et l'appréhension convenable des liens entre les concepts linguistiques fondamentaux et les méthodes d'évaluation les plus appropriées pour cibler ces niveaux constitue l'un des apports à la linguistique fondamentale. Celui-ci est permis par l'observation à travers le prisme des déficits et altérations de la parole, tout en soutenant l'objectif de l'évaluation en contexte clinique.

Reste à établir et déterminer avec précision les degrés de corrélation entre les différentes mesures, tant dans les méthodes (tâche de l'auditeur : notation par échelles, par transcriptions), que dans les procédures de calcul des scores (globale vs analytique) ainsi qu'entre les différents matériaux linguistiques (énoncés, phrases, mots, pseudo-mots). En ce sens, le développement d'un index d'évaluation de la parole qui présenterait synthétiquement les différents matériaux linguistiques associés avec les niveaux linguistiques évalués, en fonction des différentes méthodes de scorage, constituerait un outil essentiel tant à la pratique clinique qu'à la linguistique fondamentale.

Références

ANDRÉ, C. *ET AL.* (2003): «PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception», in *XVth ICPhS. ICPhS*, Barcelone, Espagne, p. 1421-1424.

ASTÉSANO, C. *ET AL.* (2018): «Carcinologic Speech Severity Index Project: A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer», in. *Language Resources and Evaluation Conference*, Miyazaki, p. 7.
AUZOU, P. ET ROLLAND-MONNOURY, V. (2006): *Batterie d'évaluation clinique de la dysarthrie*. Ortho Edition. France: ORTHO.
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*.
BECHET, F. (2001): «LIA PHON : Un système complet de phonétisation de textes», *Traitement Automatique des Langues, TAL*. (TAL - ATALA), 42(1), p. 47-67.
BLANC, E. *ET AL.* (2014): «Adaptation en français du test d'intelligibilité de la version révisée du « Frenchay Dysarthria Assessment » (FDA-2)», in *Congrès de la Société Française de Phoniatrie*. Paris, France. Disponible sur: https://hal.archives-ouvertes.fr/hal-01615204.
GANZEBOOM, M. *ET AL.* (2016): «Intelligibility of Disordered Speech: Global and Detailed Scores», in, p. 2503-2507. doi: 10.21437/Interspeech.2016-1448.
GHIO, A. *ET AL.* (2018): «Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique», in *XXXIIe Journées d'Etudes sur la Parole*. Aix-en-Provence, France: ISCA, p. 285-293. doi: 10.21437/jep.2018-33.
GHIO, A., LALAIN, M., GIUSTI, L., *ET AL.* (2020): «How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech», in *12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: ELRA, p. 1682-1687. Disponible sur: https://hal.archives-ouvertes.fr/hal-02482615.
GHIO, A., LALAIN, M., REBOURG, M., *ET AL.* (2020): «Testing intelligibility through acoustic-phonetic decoding of pseudowords : Construct and concurrent validation based on patients with head and neck cancers», *Journal of Communication Disorders*.
HUSTAD, K. C. (2008): «The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers With Dysarthria», *Journal of Speech, Language, and Hearing Research*, 51(3), p. 562-573. doi: 10.1044/1092-4388(2008/040).
HUSTAD, K. C., JONES, T. ET DAILEY, S. (2003): «Implementing Speech Supplementation Strategies: Effects on Intelligibility and Speech Rate of Individuals With Chronic Severe Dysarthria», *Journal of Speech, Language, and Hearing Research*, 46(2), p. 462-474. doi: 10.1044/1092-4388(2003/038).
JAKOBSON, R., FANT, C. G. M. ET HALLE, M. (1952): *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge, Etats-Unis d'Amérique: Acoustics Laboratory, Massachusetts Institute of Technology.
LALAIN, M. *ET AL.* (2020): «Design and Development of a Speech Intelligibility Test Based on Pseudowords in French: Why and How?», *Journal of Speech, Language, and Hearing Research*, 63(7), p. 2070-2083. doi: 10.1044/2020_JSLHR-19-00088.
LALAIN, M. *ET AL.* (2022): «Prédiction du degré d'altération de l'intelligibilité chez des patients traités pour un cancer de la cavité buccale ou de l'oropharynx», in *32ème édition des Journées d'Etudes sur la Parole*. Noirmoutier, France.
LINDBLOM, B. (1990): «On the communication process: Speaker listener interaction and the development of speech.», in *Augmentative and Alternative Communication*. (6), p. 220-230.
REBOURG, M. (2018): *Validation d'une tâche de Décodage Acoustico Phonétique : Lexicalisation, mémorisation, familiarisation*. Mémoire de Master 2 - Sciences du langage - Linguistique expérimentale. Aix-Marseille Université.
REBOURG, M. *ET AL.* (2019): «Pertinence de l'utilisation de non mots pour évaluer l'intelligibilité», in *Journées de Phonétique Clinique*. Mons, Belgium (Questions de Phonétique Clinique), p. 172. Disponible sur: https://hal.archives-ouvertes.fr/hal-02098845.
REBOURG, M. *ET AL.* (2020): «Évaluer l'intelligibilité, mots ou pseudo-mots ? Comparaison entre deux groupes d'auditeurs», in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e*

édition), (TALN, 27e édition), (RÉCITAL, 22e édition). Nancy, France: ATALA, p. 543-551. Disponible sur: https://hal.archives-ouvertes.fr/hal-02798584.
REBOURG, M. (2022): *Évaluation de l'intelligibilité après un cancer ORL : approche perceptive par décodage acoustico-phonétique et mesures acoustiques*. These de doctorat. Aix-Marseille. Disponible sur: https://www.theses.fr/2022AIXM0247 (Consulté le: 10 février 2024).
VITEVITCH, M. S. ET LUCE, P. A. (1999): «Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition», *Journal of Memory and Language*, 40(3), p. 374-408. doi: 10.1006/jmla.1998.2618.
WOISARD, V. *ET AL.* (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*. Springer Verlag, 55(1), p. 173-190. doi: 10.1007/s10579-020-09496-3.
XUE, W., HOUT, R., *ET AL.* (2021): «Assessing speech intelligibility of pathological speech: test types, ratings and transcription measures», *Clinical Linguistics & Phonetics*, 37. doi: 10.1080/02699206.2021.2009918.
XUE, W., HOUT, R. V., *ET AL.* (2021): «Speech Intelligibility of Dysarthric Speech: Human Scores and Acoustic-Phonetic Features», in *Interspeech 2021*. *Interspeech 2021*, ISCA, p. 2911-2915. doi: 10.21437/Interspeech.2021-1189.
XUE, W. *ET AL.* (2023): «Assessing speech intelligibility of pathological speech in sentences and word lists: The contribution of phoneme-level measures», *Journal of Communication Disorders*, 102, p. 106301. doi: 10.1016/j.jcomdis.2023.106301.
YORKSTON, K. M., DOWDEN, P. A. ET BEUKELMAN, D. R. (1992): «Intelligibility measurement as a tool in the clinical management of dysarthric speakers», in *Intelligibility in speech Disorders : Theory, measurement and management*. Raymond D. Kent. Madison, Wisconsin: John Benjamins Publishing Company, p. 265-286. Disponible sur: https://benjamins.com/catalog/sspcl.1.08yor.
YORKSTON, K. M., STRAND, E. A. ET KENNEDY, M. R. T. (1996): «Comprehensibility of Dysarthric Speech», *American Journal of Speech-Language Pathology*. American Speech-Language-Hearing Association, 5(1), p. 55-66. doi: 10.1044/1058-0360.0501.55.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Peut-on marquer un focus contrastif par le geste manuel en suppléance vocale ?

Delphine Charuau Nathalie Henrich Bernardoni Silvain Gerber Olivier Perrotin
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France
delphinecharuau1@gmail.com, olivier.perrotin@gipsa-lab.grenoble-inp.fr

RÉSUMÉ

Un paradigme expérimental élicitant la focalisation sur une syllabe a été élaboré dans une tâche de conversion chuchotement-parole avec contrôle manuel de l'intonation. Deux interfaces de contrôle intonatif ont été testées : contrôle isométrique par pression du doigt et isotonique par rotation du poignet. La réalisation de la focalisation par le geste a été observée, démontrant un transfert du contrôle naturel vers manuel de l'intonation. Les résultats sont également discutés en fonction de la position de la syllabe dans l'énoncé, et en fonction de l'interface de contrôle gestuel employée.

ABSTRACT

Can a contrastive focus be achieved using hand gestures in a voice substitution paradigm ?

An experimental paradigm to elicit a focus on a syllable was developed, in a whisper-to-speech conversion task with manual control of intonation. Two interfaces for controlling intonation were tested : an isometric control by finger pressure and an isotonic control by wrist rotation. A succesful realisation of focus with gesture was observed, demonstrating a transfer from natural to manual control of intonation. Results are also discussed in terms of syllable position in the utterance, and with regards to the gestural control interface employed.

MOTS-CLÉS : Focus contrastif, Contrôle chironomique, Conversion chuchotement-parole, Suppléance vocale.

KEYWORDS: Contrastive focus, Chironomic control, Whisper-to-speech conversion, Voice substitution.

1 Introduction

Notre étude porte sur le contrôle de l'intonation dans le cadre d'une dégradation ou d'une absence des capacités phonatoires chez des patients laryngectomisés. Les solutions médicales actuelles pour remplacer la source vocale défectueuse ou absente consistent à injecter une source sonore artificielle dans le conduit vocal, souvent à l'aide d'un électrolarynx (Liu & Ng, 2007; Fuchs *et al.*, 2016; Kaye *et al.*, 2017; Ahmadi *et al.*, 2018). Ce vibrateur génère une source vocale de substitution sur laquelle l'utilisateur peut articuler la parole normalement. Il est également possible d'utiliser un microphone pour capter la parole non vocalisée (par exemple un chuchotement), et d'y réintroduire une phonation en temps-réel par synthèse vocale à partir d'un modèle de source glottique filtré par la réponse du conduit vocal mesurée (Perrotin & McLoughlin, 2020). La voix reconstruite est ensuite diffusée en temps-réel sur un haut-parleur. La principale limite à la fois des électrolarynx et des systèmes de

synthèse est le peu d'information disponible pour reconstruire l'intonation. Par défaut, ces systèmes génèrent une intonation relativement constante, privant ainsi l'expression parlée d'une partie de l'information prosodique nécessaire à la fois à la structuration du discours (Mertens, 2008; Di Cristo, 2016) et à l'expression d'attitudes et d'émotions (Ward, 2019).

La synthèse vocale performative consiste à proposer à l'utilisateur un contrôle temps-réel de certains paramètres de la voix à générer, à l'image d'un instrument de musique numérique. En particulier, un axe d'étude a exploré l'utilisation de la chironomie, c'est-à-dire la gestuelle de la main, pour le contrôle de l'intonation dans la synthèse vocale (d'Alessandro, 2022). Ainsi, un tel paradigme permet d'aller chercher l'information d'intonation auprès d'un geste non-vocal (la main), en faisant l'hypothèse forte que l'utilisateur est capable de transférer une production de l'intonation implicite lorsque produite par la vibration des plis vocaux vers une production explicite par le geste de la main. Cette hypothèse a été validée dans des *tâches d'imitation*, où des utilisateurs ont reproduit le contour intonatif de phrases données en contrôlant la fréquence fondamentale de synthétiseurs vocaux par la position d'un stylet sur une tablette graphique, à la fois sur des tâches de parole (d'Alessandro *et al.*, 2011) et de chant (d'Alessandro *et al.*, 2014). Le transfert inverse, du contrôle manuel de l'intonation vers le contrôle naturel, a aussi été observé et s'est montré efficace dans l'apprentissage de l'intonation du français (Xiao *et al.*, 2022) et de l'anglais (Xiao *et al.*, 2023) en tant que langues étrangères, toujours en tâches d'imitation.

Dans la lignée de ces recherches, nous avons introduit la synthèse performative dans une solution de suppléance vocale basée sur la conversion chuchotement-parole en temps-réel. L'utilisateur doit alors à la fois articuler le message avec son conduit vocal et contrôler l'intonation de manière synchrone à l'aide du geste manuel (Perrotin & McLoughlin, 2020; Ardaillon *et al.*, 2022). Néanmoins, le paradigme de suppléance vocale diffère des études précédentes selon trois aspects. D'abord, si l'hypothèse de transfert du contrôle intonatif entre plis vocaux et geste de la main a été démontré sur des tâches d'imitations, ces dernières sont peu fréquentes en situation de communication orale. La question du transfert du contrôle intonatif sur des tâches que nous appellerons *tâches de production*, i.e. où l'utilisateur doit produire des contours intonatifs avec un but communicationnel mais sans références immédiates, reste donc ouverte. Par ailleurs, les synthétiseurs performatifs développés pour les études en tâche d'imitation demandent un contrôle de l'intonation uniquement, le contenu phonétique étant pré-défini (Feugère *et al.*, 2017; Locqueville *et al.*, 2020). En suppléance vocale, le contrôle simultané de l'articulation et de l'intonation doit être pris en compte. Enfin, ces mêmes instruments de synthèse proposent un contrôle de l'intonation par la position d'un objet (stylet ou doigt) sur une surface. Si cela introduit une modalité visuelle dans la représentation de l'intonation qui a été démontrée comme prépondérante dans le contrôle (Perrotin & D'alessandro, 2016), celle-ci n'est pas souhaitable dans des situations de communication orale, pour des raisons ergonomiques.

Dans cet article, nous cherchons donc à évaluer la capacité de locuteurs à externaliser le contrôle de l'intonation pour la réalisation de fonctions prosodiques dans un contexte communicationnel de suppléance vocale, c'est-à-dire dans des *tâches de production*, en synchronie avec l'articulation, et en utilisant des interfaces gestuelles ne sollicitant pas la modalité visuelle. Nous nous intéressons ici à la focalisation contrastive, qui se caractérise par la mise en évidence d'un ou plusieurs mots d'un énoncé, jugés comme les plus informatifs. Un focus contrastif est marqué par une augmentation notable de l'intensité et de la fréquence fondamentale f_o sur le ou les mots d'intérêt (Jun & Fougeron, 2000; Grice *et al.*, 2017), pouvant s'accompagner d'un allongement temporel de la syllabe portant le focus (Dahan & Bernard, 1996; Astésano *et al.*, 2004). La combinaison de ces paramètres contribue à la perception auditive du focus et permet une évaluation objective en termes de variations de durée syllabique et de fréquence fondamentale. Des variations fines de ces paramètres s'opèrent au niveau

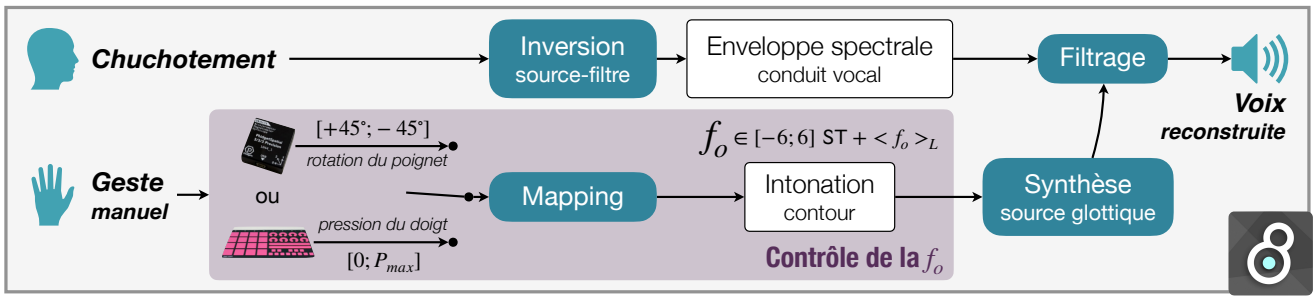


FIGURE 1 – Schéma du système de suppléance vocale avec contrôle manuel de l'intonation.

syllabique de manière complémentaire, lorsqu'il s'agit de marquer le focus sur l'ensemble du mot (Astésano *et al.*, 2004). En effet, les variations d'intensité et de f_o sont plutôt localisées sur la première syllabe ou au milieu du mot, tandis que les variations de durée arrivent à la fin du mot. Il s'agit de voir si les locuteurs reproduisent des dépendances à la position de la syllabe en suppléance vocale.

Nous posons une première hypothèse selon laquelle les locuteurs pourront combiner un contrôle implicite du focus sur le plan articulatoire, en allongeant la durée de la syllabe à mettre en relief, avec un contrôle explicite sur le plan gestuel, en élevant le contour intonatif sur la syllabe d'intérêt, par le biais de l'interface manuelle. Notre seconde hypothèse porte sur un transfert de l'effet de la position de la syllabe sur la f_o ainsi que sur la durée des syllabes. Le protocole expérimental pour l'élicitation d'un focus contrastif et l'évaluation de sa réalisation par le geste manuel est décrit en Section 2. Les résultats sont discutés en Section 3.

2 Matériel and méthodes

2.1 Système de suppléance vocale

Le système utilisé dans ces travaux et présenté en Fig. 1 est composé d'un module de conversion chuchotement-parole et d'interfaces gestuelles pour le contrôle de l'intonation, détaillés ci-dessous.

Conversion chuchotement-parole : Le système est une extension de la méthode proposée par Perrotin & McLoughlin (2020) qui consiste en : 1) la décomposition source-filtre du chuchotement par la méthode GFM-IAIF (Perrotin & McLoughlin, 2019), pour isoler l'enveloppe spectrale du conduit vocal du bruit coloré correspondant à la source sonore du chuchotement ; 2) la génération d'un signal de source glottique par le modèle LF (Fant *et al.*, 1994), de fréquence fondamentale f_o donnée ; 3) le filtrage du signal de source par l'enveloppe spectrale du conduit vocal. Ces trois étapes sont implémentées en temps-réel sur la plate-forme Max/MSP (Cycling74, 2024).

Interfaces gestuelles : La f_o utilisée en synthèse est contrôlée linéairement sur une échelle en demi-tons (ST), sur un intervalle d'une octave (± 6 ST) autour de la f_o moyenne de la voix du locuteur, notée $< f_o >_L$. Celle-ci est mesurée lors de la première phase d'entraînement du protocole sur de la parole naturelle (voir Section 2.2). Nous avons proposé deux types de geste pour le contrôle de l'intonation : un geste de *pression* isométrique et un geste de *rotation* isotonique. Un contrôle intonatif par geste de pression du pouce est déjà proposé dans la solution commercialisée et très usitée de l'électrolarynx Trutone (2024). Le geste de rotation du poignet s'inspire des gestes de battement qui peuvent accompagner la focalisation (Leonard & Cummins, 2011). Le geste de *pression* est réalisé sur

Scénario							Condition
– Participant :	Le	loup	doux	a suivi	le beau	loup .	*Pré*
– Expérimentatrice :	Le loup doux			a suivi	le beau chien ?		*Question*
– Participant :	Le	loup	doux	a suivi	le beau	loup .	*Post*

(a) Scénario pour l'expérience. La syllabe soulignée est celle ciblée par la question de l'expérimentatrice.

Syllabe		Énoncé			Contraste
cible	*non-cible*	*Sujet (S)*	*Verbe (V)*	*Objet (O)*	*Mot changé dans la question*
S1	O2	Lou du Mans	a suivi	le loup doux.	Jean
S2	O3	Le loup doux	a suivi	le beau loup.	chat
S3	O1	Le beau loup	a suivi	Lou du Mans.	chien
O1	S3	Le beau loup	a suivi	Lou du Mans.	Jean
O2	S1	Lou du Mans	a suivi	le loup doux.	chat
O3	S2	Le loup doux	a suivi	le beau loup.	chien

(b) Corpus d'énoncés.

TABLE 1 – Corpus (bas) et exemple de scénario sur un des énoncés (haut).

une tablette Sensel de la marque Morph (2024), qui mesure la pression de l'index de la main préférée de l'utilisateur, d'une pression nulle à une pression maximale P_{max}, respectivement associées à -6 et $+6$ ST autour de $< f_o >_L$. Le geste de *rotation* est réalisé à l'aide d'un accéléromètre 1044_1B de la marque Phidget (2024), tenu dans la main préférée de l'utilisateur, dont l'avant-bas est posé à l'horizontale sur un accoudoir. L'accéléromètre mesure le mouvement haut/bas du poignet de la main en degrés de rotation, $0°$ étant la position horizontale du poignet correspondant à $< f_o >_L$. Les rotations $+45°$ et $-45°$ sont respectivement associées à -6 et $+6$ ST autour de $< f_o >_L$. Ainsi, une descente du poignet est liée à une augmentation de f_o.

2.2 Protocole expérimental

Scénario : Nous avons construit un scénario permettant l'induction d'un focus contrastif sans donner d'instruction explicite, suivant les travaux de Dohen & Lœvenbruck (2009). Il s'agit d'une tâche de parole sous la forme d'interactions simulées entre le participant et l'expérimentatrice. L'interaction comporte trois tours de parole, résumés en Table 1a, dont le texte s'affiche au fur et à mesure sur l'écran disposé face au participant. Le participant a d'abord la consigne de lire un énoncé affiché. Ensuite, une question pré-enregistrée par l'expérimentatrice est affichée et jouée au participant. Cette question reprend l'énoncé du participant en y changeant un mot pour simuler une erreur de compréhension. Enfin, le participant a pour consigne de répéter l'énoncé initial qui s'affiche à l'écran. On appelle *Pré* et *Post* la première et deuxième répétition de la phrase par le participant.

Corpus : Le corpus est constitué de 3 phrases de 9 syllabes de type Sujet-Verbe-Objet (SVO), avec 3 syllabes par constituant. Le constituant verbal est fixe pour toutes les phrases et les constituants sujet et objet sont chacun composés de 3 mots monosyllabiques entièrement voisés (pour s'affranchir d'une décision de voisement dans la conversion chuchotement-parole). Chacun de ces deux constituants contient une syllabe /lu/ et chaque phrase apparaît deux fois dans le corpus, où soit le premier /lu/ soit le deuxième est la syllabe *cible*, c'est-à-dire changée par l'expérimentatrice. On appellera l'autre syllabe /lu/ *non-cible*. Au total, le corpus est composé de 6 énoncés, possédant chacun la syllabe *cible* à une ***position dans l'énoncé*** différente, comme indiqué en Table 1b.

Au final, lors de la production d'un scénario (Table 1a), la syllabe /lu/ est prononcée 4 fois par le

participant, selon 4 ***statuts de la syllabe*** : *Pré non-cible* (rouge clair), *Pré cible* (rouge foncé), *Post non-cible* (vert clair) et *Post cible* (vert foncé). Alors qu'aucune autre consigne que de lire chaque énoncé est donnée, nous faisons l'hypothèse que le participant produira naturellement un focus contrastif dans la condition *Post cible* uniquement. Dans la suite, on appellera *tâche de production* la réalisation des 6 scénarios correspondant à chacun des énoncés, répétés 3 fois chacun. L'ordre de présentation des scénarios et répétitions est aléatoire.

Productions vocales : L'expérience est divisée en trois phases associées à trois ***modes de production***, chacune précédée d'une ou plusieurs activités de familiarisation au protocole et/ou au contrôle. Dans la première, le participant utilise sa *voix* naturelle et commence par 3 répétitions de 6 scénarios sur des énoncés autres que ceux présentés en Table 1a pour se familiariser avec la tâche. C'est dans cette phase que $< f_o >_L$ est mesuré. Ensuite, le participant réalise la *tâche de production*. En deuxième et troisième phase, il utilise le système de conversion chuchotement-parole en contrôle par *rotation* puis *pression* ou inversement. L'ordre de passage de ces deux dernières phases est attribué aléatoirement selon le participant. Pour chacune de ces phases, le participant commence par la lecture du texte MonPage 2 (Pommée, 2021, p. 114) qui lui permet de se familiariser à l'interface par un contrôle libre de l'intonation. Le deuxième entraînement est une tâche d'imitation, où il s'agit de reproduire 6 phrases enregistrées par d'Alessandro *et al.* (2011) en imitant leur intonation, avec 3 répétitions chacune. Cela permet d'apprendre au participant à contrôler l'intonation. Enfin, la phase se termine par la réalisation de la *tâche de production*.

Conditions expérimentales : L'expérience s'est déroulée en chambre anéchoïque au laboratoire. Chaque participant est assis face à un écran sur lequel s'affiche les consignes et les supports des tâches de parole. Un casque audio fermé Beyerdynamic DT797 équipé d'un micro est utilisé pour capter la voix de l'utilisateur et lui restituer la voix de synthèse en temps-réel, ainsi que les stimuli sonores. Le participant devait passer manuellement au scénario suivant et pouvait faire des pauses librement à ces moments-là. L'expérience complète durait environ 1h15 et était rémunérée 15€ en bon d'achat dans un grand magasin.

Participants : Nous avons enregistré 16 locuteurs (âge médian = 24.5 ans ; Q1 = 22.5 ; Q3 = 27), de langue maternelle française et sans trouble rapporté de la parole, de l'audition et de la motricité du bras et de la main. Le protocole expérimental a été approuvé par le comité d'éthique de l'université Grenoble Alpes (CERGA-Avis-2023-21) et respecte le Règlement Général sur la Protection des Données.

2.3 Traitement des données

Extraction des données : L'alignement texte-parole des fichiers audio a été réalisé à l'aide de l'application Astali (Loria, 2016). La segmentation en syllabes ainsi que leurs annotations (*cible*/ *non-cible*) ont été réalisées manuellement sur Praat. Pour chaque syllabe, nous reportons sa durée relative par rapport à la durée de l'énoncé pour lequel les pauses ont été exclues, appelée D_r. Chaque énoncé ayant 9 syllabes, la durée relative moyenne d'une syllabe est de 11%. La f_o de la voix naturelle a été mesurée automatiquement par la fonction `To pitch` de Praat et la f_o contrôlée par le geste est fournie directement par le système. Ces valeurs sont exprimées en ST. Pour s'affranchir de l'effet du locuteur et de la production vocale, nous soustrayons à chaque trajectoire de f_o mesurée la médiane de f_o calculée pour l'ensemble des productions du locuteur avec la production vocale correspondante. Nous appelons f_{o_c} la fréquence fondamentale centrée résultante, exprimée aussi en ST. Dans cette étude, nous reportons le pic de f_{o_c} sur les syllabes d'intérêt (*cible* et *non-cible*).

Analyses statistiques : Nous avons étudié l'impact du ***statut de la syllabe***, de sa ***position dans l'énoncé***, et de son ***mode de production*** sur la durée relative et sur le pic de f_{oc}. Pour la durée relative, compte tenu du fait que ses valeurs sont bornées dans l'intervalle (0;1), nous avons appliqué une régression beta avec effet aléatoire et utilisé la fonction `glmmTMB` du package `glmmTMB` du logiciel R. Pour le pic de f_{oc}, nous avons appliqué un modèle linéaire mixte et utilisé la fonction `lme` de la librairie `nlme` du logiciel R. Dans les deux cas, le participant et le numéro de répétitions ont été ajoutés comme effets aléatoires du modèle. Nous avons utilisé ensuite la fonction `glht` de la librairie `multcomp` du logiciel R pour réaliser des comparaisons multiples d'où sont issues les p-values données ci-après. Les résultats sont considérés comme significatifs si $p < 0.05$.

3 Résultats

3.1 Réalisation de la focalisation

La figure 2 rend compte des valeurs du pic du f_{oc} sur les syllabe /lu/ et de leur durée relative, selon leur ***position*** au sein des constituants et selon le ***mode de production***. Il convient de préciser que pour une même position, les syllabes *cible* et *non-cible* appartiennent à des énoncés différents.

Validité du protocole : Aucune différence significative n'est observée entre les syllabes *Pré non-cible* (rouge clair) et *Pré cible* (rouge foncé), tant en termes de variation de f_o qu'en termes de durée, et ce, quel que soit le ***mode de production***. Ces résultats indiquent qu'il n'y a pas d'anticipation du focus et que nous avons bien induit une causalité entre question et focus observés en condition *Post* ci-après.

Réalisation de la focalisation : La figure 2 montre que le focus est réalisé d'une part, par une augmentation de l'intonation sur la syllabe *Post cible*, et d'autre part, par un allongement sa durée relative, quel que soit le ***mode de production***. En condition *voix*, les locuteurs ont tendance à marquer le focus en augmentant, en médiane sur l'ensemble des syllabes, le pic de f_{oc} de 1.64 ST sur la syllabe *Post cible* (vert foncé), par rapport à la syllabe *Pré cible*. Toutefois, sur chaque syllabe, cette différence n'est significative que lorsque le focus tombe sur la deuxième syllabe du constituant objet (O2). En revanche, durant les tâches de production avec interface, une augmentation significative du pic de f_{oc} entre l'ensemble des syllabes *Pré cible* et *Post cible*, de 3.28 ST pour le contrôle par *pression* et de 4.59 ST pour le contrôle par *rotation*, rend compte de la volonté des locuteurs de marquer le focus sur la syllabe cible. En outre, l'ensemble des syllabes *Post cible* sont significativement allongées de 3.8% en parole naturelle, et de 6% lors des tâches de parole impliquant un contrôle manuel de l'intonation. Au regard de ces résultats, il apparaît que la focalisation est réalisée sur le plan articulatoire (durée), ainsi que par le geste manuel (f_{oc}). Par ailleurs, la hausse du pic de f_{oc} est particulièrement marquée sur les interfaces.

3.2 Effet de la position de la syllabe

Afin de vérifier les effets de la ***position dans l'énoncé*** de la syllabe sur la réalisation du focus, nous comparons la durée relative et le pic de f_{oc} sur les syllabes *Pré cible* et *Post cible* en *voix* selon leur position au sein des constituants, puis en production *pression* et *rotation*.

Variations prosodiques en voix naturelle : La comparaison des valeurs de f_{oc} pour les 6 syllabes *Pré cible* selon leur position fait ressortir 11 combinaisons significatives sur les 15 testées. L'analyse

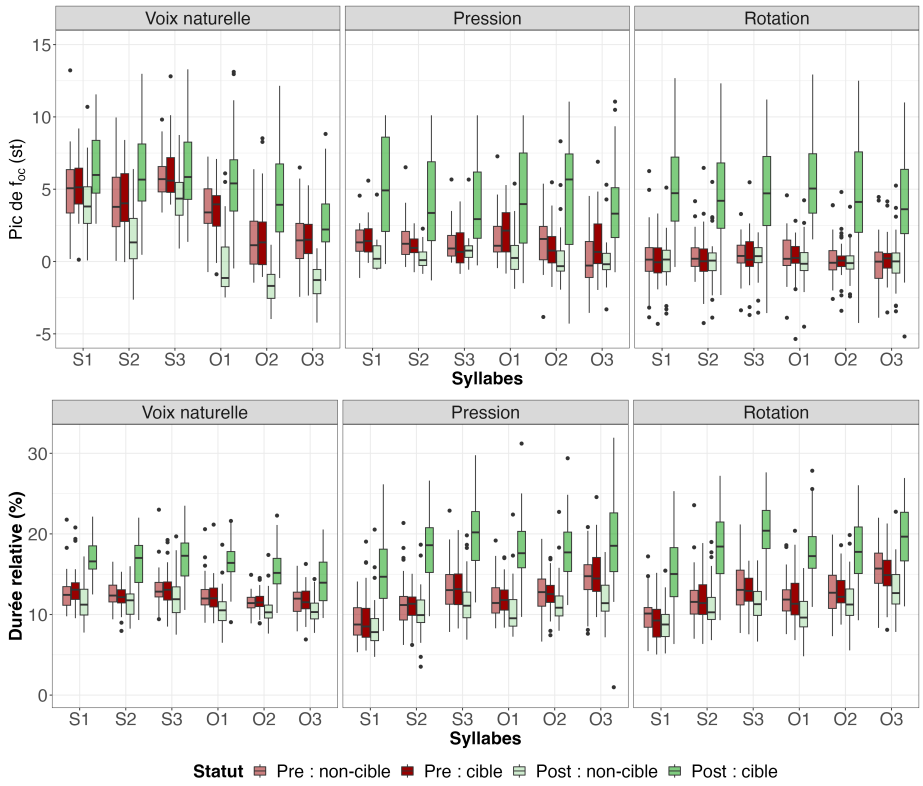


FIGURE 2 – Pic de f_{oc} (haut) et durée relative (bas) des syllabes /lu/ selon le ***statut de la syllabe*** et le ***mode de production***.

statistique révèle une variabilité significative du f_{oc} entre l'ensemble des syllabes du constituant objet. Cette variabilité s'atténue au sein du constituant sujet pour lequel une différence significative n'est observée qu'entre S2 et S3. Le nombre de combinaisons significatives rend compte d'une forte dépendance de la variation de f_{oc} à la position de la syllabe en condition *Pré*. Nous constatons donc une variation de f_{oc} des syllabes *Pré cible* selon leur position dans les constituants et dans la phrase, notamment une baisse significative de f_{oc} à la fin du constituant objet, coïncidant avec l'intonation descendante en fin de phrase. En revanche, la position de la syllabe ne présente que peu d'effet sur la variation du f_{oc} des syllabes *Post cible* : sur 15 combinaisons testées, seules 6 s'avèrent significatives. Nous n'observons aucune différence significative entre les syllabes d'un même constituant, sujet et objet, excepté entre O2 et O3. La réduction du nombre de différences significatives entre les syllabes *Pré cible* et *Post cible* suggère donc un plafonnement de f_{oc} lors de la réalisation du focus.

En parole naturelle, la durée des syllabes *Pré cible* reste relativement constante tout au long des constituants. Nous relevons une constance similaire dans la durée des syllabes *Post cible* au sein du constituant sujet. Une légère baisse de la durée de la syllabe initiale du constituant objet est relevée, tandis que la durée des syllabes en deuxièmes et troisièmes positions de ce constituant décroît significativement. De manière générale, les syllabes sont plutôt isochrones, exceptées pour les deux dernières syllabes de la phrase. Nous ne relevons pas de dépendance à la position de la syllabe.

Variations prosodiques en contrôle de l'intonation : Lors des tâches de parole impliquant un contrôle manuel de l'intonation, nous n'observons aucun effet de la position de la syllabe sur le contour de f_{oc}. En effet, aucune différence significative n'est observée selon la position syllabique, aussi bien pour les syllabes *Pré cible* que pour les syllabes *Post cible*, bien que nous retrouvons un plafonnement de la focalisation pour les syllabes *Post cible*, à l'instar de ce qui a été constaté en *voix*. Dans le contexte d'un contrôle externe de l'intonation, en dehors de la réalisation du focus, il y a

donc peu de mouvements intonatifs durant la production de la phrase. Les participants se concentrent exclusivement sur la tâche de focalisation, au détriment des autres fonctions de la prosodie.

Lors d'un contrôle par *rotation*, la durée relative des syllabes *Pré cible* augmente significativement au fur et à mesure que l'on s'approche de la fin du constituant, aussi bien sujet qu'objet. Si nous observons un schéma similaire lors d'un contrôle par *pression*, seule la hausse de durée entre les syllabes *Pré cible* initiale et finale du constituant objet est significative. Les syllabes *Post cible* montrent également une augmentation significative de leur durée au fur et à mesure de leur avancée au sein du constituant sujet, quelle que soit l'interface employée. Néanmoins, si une hausse similaire est également observée pour les syllabes du constituant objet, elle n'est pas significative. Nous observons un schéma d'allongement des syllabes au sein de chaque constituant, qui se confirme à la fois dans les contextes avec et sans focalisation, quel que soit le contrôle manuel exercé (*pression* et *rotation*). Ces allongements spécifiques aux interfaces pourraient être imputables au chuchotement, qui ralentit naturellement le débit de parole (Schwartz, 1967; Houle & Levi, 2020), et/ou à la hausse de la charge cognitive, induite non seulement par l'externalisation de l'intonation, mais également par la coordination entre l'articulation et le geste manuel.

4 Conclusion

Au regard des résultats présentés, nous pouvons conclure au succès du transfert de la production du focus à travers la variation de f_{oc} et de la durée sur la syllabe cible. En effet, lors des tâches de production impliquant un contrôle manuel de l'intonation, tous les locuteurs ont clairement explicité le focus en augmentant les valeurs de f_{oc} à l'endroit attendu. Ceci indique que les locuteurs ont non seulement perçu l'importance de f_{oc} dans la réalisation du focus, mais aussi qu'ils ont su utiliser les interfaces pour mettre en relief la syllabe cible. Ce comportement a été observé chez tous nos locuteurs. Par ailleurs, une augmentation significative de la durée des syllabes *Post cible* a également été constatée, à l'instar des données en parole naturelle attestées dans la littérature.

Contrairement aux observations faites en *voix*, lors des tâches de production impliquant un contrôle gestuel, nous ne remarquons aucune variation significative de f_o en dehors de la focalisation. Hormis lors de la réalisation du focus, la voix de synthèse est relativement monotone. Nous ne relevons que peu ou pas de variation de f_o, y compris à la fin des énoncés, où l'on pourrait s'attendre à une intonation descendante. Si les locuteurs ont montré une aptitude à reproduire précisément les contours intonatifs en tâche d'imitation (d'Alessandro *et al.*, 2011), dans notre tâche de production, les participants semblent s'être concentrés exclusivement sur la tâche de focalisation, au détriment des autres fonctions de la prosodie. La méthode employée se révèle néanmoins encourageante. En effet, cette étude met en évidence la capacité des participants à 1) comprendre qu'il fallait produire un focus ; 2) intégrer que le focus s'exprime en partie par une augmentation de f_o ; 3) planifier et mettre en œuvre ce contrôle par le biais des interfaces. Cependant, il apparaît que l'utilisation des interfaces ait été limitée à une seule fonction prosodique. Ceci suggère qu'un apprentissage complet de toutes les fonctions de la communication pourrait nécessiter un entraînement prolongé.

Au-delà de l'étude d'autres fonctions prosodiques, il conviendrait d'examiner le phénomène de ralentissement de la production des syllabes, observé en conditions *pression* et *rotation*, possiblement attribuable à la charge cognitive engendrée par la coordination entre l'articulation et le geste manuel, et/ou au chuchotement. Cette question suscite un certain intérêt dans le cadre de travaux futurs.

Références

AHMADI F., NOORIAN F., NOVAKOVIC D. & VAN SCHAIK A. (2018). A pneumatic Bionic Voice prosthesis—Pre-clinical trials of controlling the voice onset and offset. *PLOS ONE*, **13**(2), e0192257. DOI : 10.1371/journal.pone.0192257.

ARDAILLON L., HENRICH N. & PERROTIN O. (2022). Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In *Interspeech 2022*, p. 2253–2257 : ISCA. DOI : 10.21437/Interspeech.2022-10675.

ASTÉSANO C., MAGNE C., MOREL M., COQUILLON A., ESPESSER R., BESSON M. R. & LACHERET-DUJOUR A. (2004). Marquage acoustique du focus contrastif non codé syntaxiquement en français. In *25èmes Journées d'Études sur la Parole*, p.ã, Fès, Maroc : AFCP.

CYCLING74 (2024). Max 8, http://cycling74.com.

DAHAN D. & BERNARD J.-M. (1996). Interspeaker variability in emphatic accent production in french. *Language and Speech*, **39**(4), 341–374. DOI : 10.1177/002383099603900402.

D'ALESSANDRO C. (2022). Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant. In *Actes des Journées d'Etudes sur la Parole (JEP)*, p. 625–636, Noirmoutiers, France : ISCA. DOI : 10.21437/JEP.2022-66.

D'ALESSANDRO C., FEUGÈRE L., LE BEUX S., PERROTIN O. & RILLIARD A. (2014). Drawing melodies : Evaluation of chironomic singing synthesis. *The Journal of the Acoustical Society of America*, **135**(6), 3601–3612. DOI : 10.1121/1.4875718.

DI CRISTO A. (2016). *Les musiques du français parlé : essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain*. Volume 1 de Études de linguistique française. De Gruyter. DOI : 10.1515/9783110479645.

DOHEN M. & LŒVENBRUCK H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, **52**(2-3), 177–206. PMID : 19624029, DOI : 10.1177/0023830909103166.

D'ALESSANDRO C., RILLIARD A. & LE BEUX S. (2011). Chironomic stylization of intonation. *The Journal of the Acoustical Society of America*, **129**(3), 1594–1604. DOI : 10.1121/1.3531802.

FANT G., KRUCKENBERG A., LILJENCRANTS J. & BAVEGARD M. (1994). Voice source parameters in continuous speech. transformation of lf-parameters. In *International Conference on Spoken Language Processing (ICSLP)*, p. 1451–1454, Yokohama, Japan : ISCA.

FEUGÈRE L., D'ALESSANDRO C., DOVAL B. & PERROTIN O. (2017). Cantor digitalis : Chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, **2**. DOI : 10.1186/s13636-016-0098-5.

FUCHS A. K., HAGMULLER M. & KUBIN G. (2016). The New Bionic Electro-Larynx Speech System. *IEEE Journal of Selected Topics in Signal Processing*, **10**(5), 952–961. DOI : 10.1109/JSTSP.2016.2535970.

GRICE M., RITTER S., NIEMANN H. & ROETTGER T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, **64**, 90–107. Mechanisms of regulation in speech, DOI : https ://doi.org/10.1016/j.wocn.2017.03.003.

HOULE N. & LEVI S. V. (2020). Acoustic differences between voiced and whispered speech in gender diverse speakers. *The Journal of the Acoustical Society of America*, **148**(6), 4002. DOI : 10.1121/10.0002952.

JUN S.-A. & FOUGERON C. (2000). A phonological model of french intonation. *Intonation : Analysis, Modelling and Technology*, p. 209–242. DOI : 10.1007/978-94-011-4317-2_10.

KAYE R., TANG C. G. & SINCLAIR C. F. (2017). The electrolarynx : voice restoration after total laryngectomy. *Medical Devices : Evidence and Research*, **Volume 10**, 133–140. DOI : 10.2147/MDER.S133225.

LEONARD T. & CUMMINS F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, **26**(10), 1457–1471. DOI : 10.1080/01690965.2010.500218.

LIU H. & NG M. L. (2007). Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, **34**(3), 327–332. DOI : 10.1016/j.anl.2006.11.010.

LOCQUEVILLE G., D'ALESSANDRO C., DELALEZ S., DOVAL B. & XIAO X. (2020). Voks : Digital instruments for chironomic control of voice samples. *Speech Communication*, **125**, 97–113. DOI : 10.1016/j.specom.2020.10.002.

LORIA (2016). Astali. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

MERTENS P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique*, **56**(1), 97–124. DOI : 10.3917/tl.056.0097.

MORPH (2024). Surface tactile, https://morph.sensel.com.

PERROTIN O. & D'ALESSANDRO C. (2016). Seeing, Listening, Drawing : Interferences between Sensorimotor Modalities in the Use of a Tablet Musical Interface. *ACM Transactions on Applied Perception*, **14**(2), 1–19. DOI : 10.1145/2990501.

PERROTIN O. & MCLOUGHLIN I. V. (2019). A spectral glottal flow model for source-filter separation of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, ICASSP '19, p. 7160–7164, Brighton, UK : IEEE. DOI : 10.1109/ICASSP.2019.8682625.

PERROTIN O. & MCLOUGHLIN I. V. (2020). Glottal Flow Synthesis for Whisper-to-Speech Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 889–900. DOI : 10.1109/TASLP.2020.2971417.

PHIDGET (2024). Accéléromètre, https://www.phidgets.com/docs/Accelerometer_Guide.

POMMÉE T. (2021). *Les mesures d'intelligibilité : État de l'art, considérations pratiques pour l'applicabilité clinique et explorations acoustiques*. Thèse de doctorat, Université Toulouse III Paul Sabatier.

SCHWARTZ M. F. (1967). Syllable duration in oral and whispered reading. *The Journal of the Acoustical Society of America*, **41**(5), 1367–1369. DOI : 10.1121/1.1910487.

TRUTONE (2024). Electrolarynx, https://www.atosmedical.com/products/provox-trutone-emote-2.

WARD N. G. (2019). *Prosodic Patterns in English Conversation*. Cambridge University Press. DOI : 10.1017/9781316848265.

XIAO X., AUDIBERT N., LOCQUEVILLE G., D'ALESSANDRO C., KÜHNERT B., KLEINBERGER R. & PILLOT-LOISEAU C. (2022). Évaluation de la stylisation chironomique pour l'apprentissage de l'intonation du français L2. In *Actes des Journées d'Etudes sur la Parole (JEP)*, Journées d'Études sur la Parole " Parole, Geste, Musique : des unités à leur organisation ", p. 465–473, Noirmoutier, France : AFCP. HAL : hal-03838095.

XIAO X., KUHNERT B., AUDIBERT N., LOCQUEVILLE G., PILLOT-LOISEAU C., ZHANG H. & D'ALESSANDRO C. (2023). Performative Vocal Synthesis for Foreign Language Intonation Practice. In *CHI '23 : CHI Conference on Human Factors in Computing Systems*, p. 1–9, Hamburg, Germany : ACM. DOI : 10.1145/3544548.3581210, HAL : hal-04113924.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Réductions temporelles en français parlé : Où peut-on trouver les zones de réduction ?

Yaru Wu[1,2,3] Kim Gerdes[2] Martine Adda-Decker[2,3]
(1) UR4255 CRISCO, Université de Caen Normandie, France
(2) LISN (CNRS), Université Paris-Saclay, 91405 Orsay cedex, France
(3) Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
yaru.wu@unicaen.fr, gerdes@lisn.fr, madda@limsi.fr

RÉSUMÉ

Cet article examine la réduction dans la parole continue en français, ainsi que les différents facteurs qui contribuent au phénomène, tels que le style de parole, le débit de parole, la catégorie de mots, la position du phone dans le mot et la position du mot dans les groupes syntaxiques. L'étude utilise trois corpus de parole continue en français, couvrant la parole journalistique formelle, la parole journalistique informelle et la parole familière. La méthode utilisée comprend l'alignement forcé et l'étiquetage automatique des zones de réduction. Les résultats suggèrent que la réduction de la parole est présente dans tous les styles de parole, mais moins fréquente dans la parole formelle, et que la réduction est plus susceptible d'être observée lorsque le débit de parole est plus élevé. La position médiane des mots ou des groupes syntaxiques a tendance à favoriser la réduction.

ABSTRACT

Temporal Reductions in Spoken French : where can we find the reduction zones ?

This article investigates speech reduction in continuous speech in French, as well as different factors that contribute to the phenomenon, such as speech style, speech rate, word category, phone position in the word, and location of words within syntactic groups. The study uses three corpora of continuous speech in French, covering formal, less formal, and casual speech. The method used involves forced alignment and the labeling of reduction zones. The results suggest that speech reduction is present in all speech styles but less common in formal speech and that reduction is more likely to be observed in speech utterances with a high speech rate. The middle position of either words or syntactic groups tends to favor reduction.

MOTS-CLÉS : variation, réduction, style de parole, découpage syntaxique, parole continue.

KEYWORDS: variation, reduction, speech style, syntactic chunking, continuous speech.

1 Introduction

La variation de la parole peut être observée de manière continue, en particulier dans la parole familière (Duez, 1997; Ernestus, 2000; Meunier & Espesser, 2011). Dans la communication quotidienne, les mots et les phrases ne sont pas toujours articulés comme on s'y attend dans leur forme canonique complète (c'est-à-dire la forme de référence). Dans la parole naturelle, les mots sont souvent hypo-articulés, ce qui donne des productions avec des segments affaiblis ou avec moins de segments que prévu. En effet, si nous comparons la réalisation phonétique des mots à leurs prononciations canoniques, nous pouvons trouver des parties de mots supprimées ou fusionnées dans le flux continu

de la parole. En outre, les résidus articulatoires et les détails phonétiques peuvent souvent être retracés dans des séquences de mots réduites (Kohler, 1999; Hawkins, 2003), même lorsque des segments sont supprimés (Niebuhr & Kohler, 2011). Par exemple, Johnson (2004) a montré que la réduction massive était courante dans l'anglais américain conversationnel en parole continue. Schuppler *et al.* (2012) ont constaté que seulement 11,7% des occurrences totales de /t/ en fin de mot sont prononcées de manière canonique en néerlandais conversationnel. Un exemple typique de réduction temporelle de la parole, s'étendant sur plusieurs segments contigus dans la parole familière française, concerne la séquence de mots « je ne sais pas » (/ʒənəsɛpa/, 4 syllabes en français dans la forme canonique), souvent prononcée comme une séquence monosyllabique similaire à [ʃpa]. Afin de mieux comprendre la réduction de la parole en parole continue, nous étudions différents facteurs de variation contribuant à la réduction de la parole, à savoir le style de parole, le débit de parole, la catégorie de mots, la position du phone dans le mot et la position des mots dans les groupes syntaxiques.

La réduction est largement considérée comme un phénomène qui se produit dans la parole familière (voir par exemple Johnson, 2004; Pluymaekers *et al.*, 2005; Schuppler *et al.*, 2012). Cependant, ce phénomène peut également être observé dans des contextes plus formels (Adda-Decker & Lamel, 2017). Nous nous attendons à observer moins de réductions dans une parole plus formelle. En outre, on sait peu de choses sur la façon dont la réduction est distribuée, pour différents styles de parole, en fonction de différents facteurs linguistiques. Le débit de parole s'est avéré être un facteur important pour la réalisation/suppression des occlusives alvéolaires à l'intérieur des mots en anglais spontané dans Raymond *et al.* (2006). Dans cet article, nous souhaitons étudier l'influence du débit de parole sur la réduction de la parole en français continu, couvrant la parole journalistique formelle, la parole journalistique informelle et la parole familière. Outre les deux facteurs de variation précédents, nous avons également ajouté à notre étude des analyses sur l'influence de la catégorie de mots, de la position du phone dans le mot et de la position des mots au sein des groupes syntaxiques, facteurs peu étudiés dans la littérature. Pour ce faire, nous utilisons le découpage syntaxique pour regrouper des suites de mots et étudier les phénomènes de réduction. De plus, Gendrot *et al.* (2016) ont montré que le découpage syntaxique pouvait aider à détecter automatiquement la hiérarchie prosodique dans un corpus de parole journalistique. Nos analyses sur les groupes syntaxiques correspondent donc dans une large mesure aux analyses sur les groupes prosodiques.

Dans ce qui suit, le corpus, l'annotation de la réduction et le traitement des données sont décrits dans la section 2. Les résultats sont présentés dans la section 3, avant les conclusions de la section 4.

2 Méthode

2.1 Corpus et alignement

Trois grands corpus de parole continue en français ont été utilisés pour notre étude, c'est-à-dire le corpus de parole journalistique formelle ESTER (Galliano *et al.*, 2006), le corpus de parole journalistique informelle ETAPE (Gravier *et al.*, 2012), et le corpus de parole familière NCCFr (Torreira *et al.*, 2010). Le corpus ESTER couvre environ 100 heures d'émissions radiophoniques en français. Le corpus ETAPE contient environ 42,5 heures de débats et de conversations à la radio et à la télévision. Le *Nijmegen Corpus of Casual French* (NCCFr) contient 35 heures de conversations informelles entre amis. Le système de transcription de la parole LISN (anciennement LIMSI, Gauvain *et al.*, 2002) a été utilisé pour segmenter automatiquement les données en mots et en phones, en mode d'alignement forcé. Avec cette méthode, la durée minimale d'un segment est de 30 ms (Adda-Decker & Lamel,

2000). L'association optimale des segments de parole avec une transcription phonémique (obtenue via un lexique de prononciation) a été choisie en fonction de la transcription du segment au niveau du mot et du modèle acoustique. Les transcriptions orthographiques ont été fournies au système de transcription en mode d'alignement forcé, ce qui nous a permis d'obtenir les frontières des phones et des mots. Le système a sélectionné automatiquement la prononciation la plus adaptée pour chaque mot à partir du dictionnaire de prononciation. Les pauses, les hésitations ou les respirations étaient également détectées automatiquement. Par la suite, nous nous référons aux pauses, hésitations ou respirations comme étant des pauses en général.

2.2 Localisation des zones de réduction

Au-delà de la réduction segmentale (par exemple, la réduction des voyelles) qui étudie la réalisation acoustique des segments affaiblis, nous définissons la réduction segmentale comme une réduction temporelle (ou contraction) de phones contigus, dont les durées restent inférieures à 40 ms chacun. Nous nous concentrons ici sur les séquences de segments qui sont réduites (i.e. temporellement courtes). La réduction temporelle est localisée grâce aux caractéristiques du système d'alignement forcé. Tout d'abord, l'alignement forcé fait correspondre les segments de parole les plus adaptés sur la base des modèles acoustiques et des variations de prononciation fournies dans le dictionnaire de prononciation. Deuxièmement, le système de transcription de la parole du LISN génère des phones d'une durée minimale de 30 ms, ce qui correspond à 3 cadres (Adda-Decker & Lamel, 2000). Lorsque des mots ou des séquences de mots sont produits avec une réduction (par exemple, des phones fusionnés ou supprimés), le système force toujours l'alignement de tous les phones présents dans la prononciation sélectionnée. Les variantes de prononciation comprennent généralement la voyelle schwa facultative et les consonnes de liaison. Pour des séquences de mots comme « par exemple » (/paʁɛgzɑ̃pl/) prononcées comme des séquences de phones similaires à [paʁɑ̃p], l'alignement mettra en évidence la réduction qui s'est produite dans la parole en forçant la présence de tous les segments ([paʁɛgzɑ̃pl]), mais en leur attribuant des durées très courtes. Par conséquent, le résultat de l'alignement forcé sera une séquence de segments d'une durée minimale de 30 (ou 40ms) afin de placer tous les phones de son modèle acoustique correspondant à la prononciation complète [paʁɛgzɑ̃pl].

Cette caractéristique spécifique du système d'alignement (c'est-à-dire la génération de segments courts consécutifs (30 ou 40 ms) pour les mots ou les séquences de mots réduits) peut être exploitée en tant qu'indicateur de réduction. Nous considérons qu'une séquence de plusieurs segments courts consécutifs (30 ou 40 ms) générée par l'alignement forcé est une zone de réduction, et plus il y a de segments dans la zone donnée, plus la réduction est considérée comme importante. Afin de mieux comprendre la réduction et le degré de réduction, nous avons décidé de diviser la réduction en quatre catégories. Les segments courts au sein d'une séquence de 3 segments courts consécutifs ou plus, de 30 ou 40 ms, sont étiquetés « 3 ». De même, les segments courts au sein d'une séquence de 2 segments courts consécutifs de 30 ou 40 ms sont étiquetés « 2 » ; les segments courts de 30 ms (c'est-à-dire la durée minimale autorisée par le système) qui ne sont pas précédés ou suivis d'un autre segment court de 30 ou 40 ms sont étiquetés « 1 ». Tous les autres phones sont étiquetés « 0 ». Le degré de réduction est donc composé de quatre catégories, à savoir « 0 », « 1 », « 2 » et « 3 ». Plus le chiffre est élevé, plus la zone de réduction est importante. Cette méthode ascendante permet d'identifier les mots et les séquences de mots qui sont réduits et de localiser les zones de réduction. Par exemple, le mot « ministre » en français (/ministʁ/) peut être réalisé sous forme de séquences phonétiques similaires à [miz] dans des mots fortement réduits. Avec l'alignement forcé, le système est « forcé » d'aligner les séquences entières ([ministʁ]) et nous obtenons 3 segments courts consécutifs ou plus de 30 ou 40 ms (c'est-à-dire le degré « 3 »). Ensuite, nous avons regroupé les trois niveaux

de réduction présentés dans cette section (c'est-à-dire « 1 », « 2 », « 3 ») dans un groupe nommé « Réduit » et « 0 » est appelé « Normal » (c'est-à-dire sans réduction). Le taux de réduction est donc la proportion de segments réduits (c'est-à-dire 1, 2, 3) par rapport à l'ensemble des points de données (c'est-à-dire 0, 1, 2, 3).

2.3 Préparation des données

Le débit de parole a été mesuré pour chaque énoncé entre les pauses (unités interpausales) et exclut donc les pauses. Le débit de parole est donc de 0 pour les pauses. Le calcul du débit basé sur les unités interpausales nous permet de vérifier si la réduction se produit davantage dans les énoncés interpausaux ayant un débit de parole plus élevé. Les mots ont également été étiquetés automatiquement en fonction de leurs catégories grammaticales de manière traditionnelle, sur la base d'un dictionnaire français des formes fléchies (Lefff, Sagot, 2010). Les mots ambigus obtiennent des étiquettes multiples, utilisées comme telles dans les règles de combinaison. Prenons par exemple le segment ambigu « les portions », qui peut être interprété comme un déterminant/nom ('les portions') ou comme un pronom/verbe ('(nous) les portions'). Le segment est annoté comme $det - pron\ noun - verb$ et regroupé sans réellement résoudre l'ambiguïté. Les cas d'ambiguïtés dont la résolution aboutirait en fait à un découpage différent sont suffisamment rares pour être ignorés. Le marquage de la partie du discours basé sur Lefff a permis de regrouper les tokens en trois catégories plus larges : 'gram' pour les mots grammaticaux, 'lex' pour les mots lexicaux et 'ponct' pour la ponctuation. [1]

La position du phone dans le mot a également été étudiée dans cette étude. Nous distinguons les phones situés au début des mots (« Deb ») des phones situés au milieu (« Mil ») ou à la fin (« Fin ») du mot. Les mots ne contenant qu'un seul phone ont été étiquetés « 1Ph ». Selon Gendrot *et al.* (2016), il est utile de détecter automatiquement la hiérarchie prosodique à travers le découpage syntaxique en utilisant un corpus de parole journalistique. Dans cet article, nous utilisons un système similaire basé sur des règles pour le découpage syntaxique qui regroupe les mots de manière à maximiser les groupes de mots avec un nombre donné de syllabes.

Le nombre de syllabes pour chaque groupe syntaxique a été calculé sur la base de la transcription et d'un système basé sur des règles, comptant les schwas finaux des mots comme une demi-syllabe. Nous avons exécuté de manière itérative l'algorithme de découpage NLTK [2] basé sur des règles avec pour objectif d'obtenir un nombre maximum de syllabes fixé à 7 syllabes, conformément à la littérature (Gendrot *et al.*, 2016). Lorsque l'algorithme atteint ou dépasse l'objectif donné, c.-à-d. le nombre maximal de sept syllabes, le chunk n'est plus qualifié pour être combiné avec les mots voisins. Nous obtenons ainsi une annotation BILU pour chaque mot (et donc aussi pour chaque phone). [3]

3 Résultats

Dans cette section, nous présentons d'abord la distribution de la réduction temporelle pour les trois corpus. Ensuite, nous présentons l'impact du débit de parole, de la catégorie de mots, de la position du phone dans le mot, de la position du mot dans le groupe syntaxique et du style de parole. En ce

1. Les catégories grammaticales de Lefff sont cl, det, pre, pro, pri, coo, aux, et csu ; les catégories lexicales sont nc, np, adj, v, adv ; $ponct$, $parent$, $epsilon$, et b pour la ponctuation.
2. NLTK est une bibliothèque Python spécialement conçue pour le traitement naturel du langage.
3. « B » signifie « Begin », représentant les premiers mots du groupe syntaxique ; « I » signifie « In », indiquant que le mot est situé au milieu du groupe syntaxique ; « L »signifie « Last », représentant le dernier mot du groupe syntaxique ; « U » représente les groupes syntaxiques qui ne contiennent qu'un seul mot.

qui concerne les analyses statistiques, nous avons utilisé le modèle linéaire généralisé (Nelder & Wedderburn, 1972) dans R (R Development Core Team, 2019). Les facteurs mentionnés ci-dessus ont été inclus comme effets fixes dans le modèle. En outre, nous avons inclus le nombre de syllabes dans le mot et la fréquence de mots comme variables de contrôle.

3.1 Distribution de la réduction

Le taux global de réduction temporelle est présenté dans cette section. Le tableau 1 montre le taux de durée pour les phones de 30 ou 40 ms (c'est-à-dire des segments extrêmement courts, donc réduits) et pour les phones de 50 ms+ (c'est-à-dire 50 ms ou plus). Le taux de réduction temporelle le plus élevé est observé pour le corpus de parole familière (∼30%), suivi par le corpus journalistique informel ETAPE (22%) et le corpus journalistique formel ESTER (18%). En d'autres termes, moins le style de parole est formel, plus la réduction est présente dans la parole.

	Durée : 30∼40ms	Durée : 50+ms
ESTER	18%	82%
ETAPE	22%	78%
NCCFr	30%	70%

TABLE 1 – Le taux de durée pour les phones (1) de 30 ou 40 ms et (2) de 50 ms ou plus.

Comme indiqué dans la section 2.2, les phones ont également obtenu une étiquette de réduction selon les conventions. Les trois corpus mis en commun, notre base de données comprend 7,6 millions de phones, dont 6,2 millions ne présentent aucune réduction. Plus précisément, 900 000 phones ont une réduction de degré « 1 », 360 000 de degré « 2 », et seulement 157 000 de degré « 3 », ce qui montre une distribution déséquilibrée des degrés de réduction.

Degré de réduction	**ESTER**		**ETAPE**		**NCCFr**	
	Occ.	%	Occ.	%	Occ.	%
0	3058667	82,1	2029469	77,67	893096	69,76
1	429280	11,52	304808	11,67	170410	13,31
2	142062	3,81	123430	4,72	93752	7,32
3	95337	**2,56**	155273	**5,94**	123066	**9,61**

TABLE 2 – Distribution de la réduction pour le corpus de parole journalistique formelle ESTER, le corpus de parole journalistique informelle ETAPE et le corpus de parole familière NCCFr.

Le tableau 2 présente la distribution de la réduction pour différents styles de parole, à savoir la parole journalistique formelle ESTER, la parole journalistique informelle ETAPE et la parole familière NCCFr. La parole familière (NCCFr) présente un taux de réduction de degré « 3 » (en gras) plus élevé que la parole journalistique formelle (ESTER) et la parole journalistique informelle (ETAPE) : ESTER 2,56 % vs ETAPE 5,94 % vs NCCFr 9,61 %. Les mêmes tendances sont observées pour les degrés « 1 » et « 2 » en ce qui concerne les styles de parole.

Il est intéressant de noter que les taux de réduction des segments étiquetés « 1 » restent relativement constants à travers les styles, alors qu'une augmentation importante peut être observée pour les segments étiquetés « 2 » et surtout pour ceux étiquetés « 3 » lorsque l'on passe de la parole formelle (ESTER) à la parole familière (NCCFr). Cela suggère que la réduction tend à impliquer des séquences plus longues plutôt que de multiples segments d'un seul phone pour une parole moins formelle.

La figure 1 illustre la distribution des durées pour les trois styles de parole (de gauche à droite : (1) la parole journalistique formelle ESTER, (2) la parole journalistique informelle ETAPE, (3) la parole familière NCCFr). La durée des phones en millisecondes (ms) est indiquée sur l'axe des

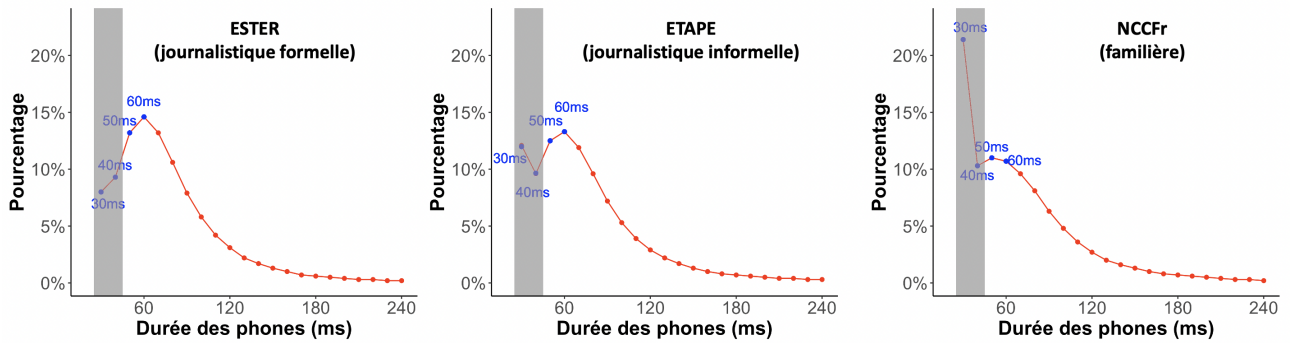


FIGURE 1 – Distribution de la durée des segments dans les trois corpus (de gauche à droite : parole journalistique formelle ESTER, parole journalistique informelle ETAPE, parole familière NCCFr).

abscisses. Des schémas différents de distribution des durées des phones sont observés pour les trois styles de parole. Pour la parole journalistique formelle ESTER, la distribution des durées de phones correspond à une courbe globale en forme de cloche et le pic de la courbe est situé à 60 ms (voir l'abscisse). Cela suggère que la durée la plus fréquemment observée dans la parole journalistique formelle est de 60 ms. Pour la parole journalistique informelle ETAPE, la proportion de la durée minimale autorisée par le système (30 ms) a augmenté de 5%. En ce qui concerne la parole familière NCCFr, le pic de la courbe correspond à la durée minimale de 30 ms, ce qui suggère que la durée la plus fréquemment observée ici est de 30 ms (>20 %). Ces observations sont cohérentes avec les résultats trouvés dans Adda-Decker & Lamel (2017) sur la durée des phones dans la parole préparée et spontanée (en français et en anglais).

Nos analyses sur la distribution de la durée nous permettent de mieux comprendre la spontanéité des trois types de parole et donc le phénomène de réduction dans la production de la parole. Dans ce qui suit, nous nous concentrerons sur ces zones grises et nous étudierons les facteurs de variation qui conditionnent la réduction.

La figure 2 présente les résultats sur le débit de parole. Les figures 3 à 5 illustrent le taux de réduction en fonction des catégories de mots, de la position du segment dans le mot et de la position du mot dans le groupe syntaxique, pour les trois styles de parole. Pour chaque figure, les résultats sur la parole journalistique formelle (ESTER) sont présentés sur le panneau de gauche, suivis de la parole journalistique informelle (ETAPE) au milieu et de la parole familière (NCCFr) sur le panneau de droite.

3.2 Débit de parole

La figure 2 présente le taux de réduction en fonction du débit de parole pour les trois styles de parole, à savoir la parole journalistique formelle ESTER, la parole journalistique informelle ETAPE et la parole familière NCCFr. Les résultats montrent que la réduction de la parole (« Réduit ») est associée à un débit de parole plus élevé que celui observé dans « Normal ». Cette tendance est observée pour les trois corpus. Les résultats du GLM confirment qu'il est plus probable d'observer la réduction de parole avec un débit de parole plus élevé [log odds ratio = 0,0776, $|Z|$ = 383,27, p<0,001].

3.3 Catégorie de mots

Le taux de réduction en fonction des catégories de mots (c.-à-d. les mots grammaticaux par rapport aux mots lexicaux) est présenté dans la figure 3. Les mots grammaticaux (« gram ») sont plus réduits que

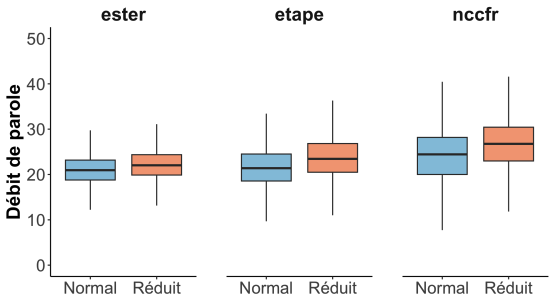


FIGURE 2 – Réduction en fonction du débit de parole pour les trois styles de parole (de gauche à droite : parole journalistique formelle ESTER, parole journalistique informelle ETAPE, parole familière NCCFr).

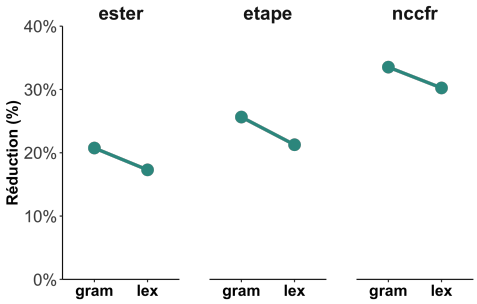


FIGURE 3 – Taux de réduction en fonction des catégories de mots (mots grammaticaux (gram) vs. mots lexicaux (lex)) pour les trois corpus (de gauche à droite : parole journalistique formelle ESTER, parole journalistique informelle ETAPE, parole familière NCCFr).

les mots lexicaux (« lex »). Ce résultat n'est pas surprenant, étant donné que les mots grammaticaux sont plus fréquemment utilisés dans la parole continue et donc plus enclins à la réduction. Les résultats du GLM confirment que la probabilité d'observer une réduction diminue significativement pour « lex » [log odds ratio = -0,1911, $|Z|$= 63,94, p<0,001], par rapport à celle observée pour « gram ».

3.4 Position du phone dans le mot

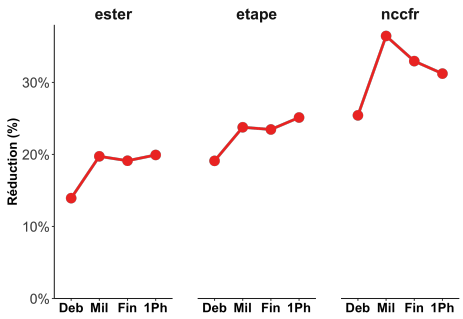


FIGURE 4 – Taux de réduction en fonction de la position du segment dans le mot (Début (« Deb ») vs. Milieu (« Mil ») vs. Fin (« Fin ») du mot vs. mot avec un phone (« 1ph »)) pour les trois corpus (de gauche à droite : parole journalistique formelle ESTER, parole journalistique informelle ETAPE, parole familière NCCFr).

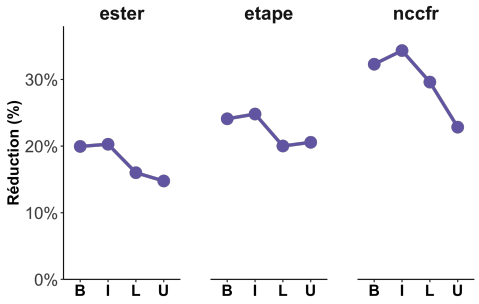


FIGURE 5 – Taux de réduction en fonction de la position du mot dans le groupe syntaxique : Début (B : *Begin*) vs. Milieu (I : *In*) vs. Fin (L : *Last*) du groupe syntaxique vs. groupe syntaxique contenant un seul mot (U : *Unique*), pour les trois corpus (de gauche à droite : parole journalistique formelle ESTER, parole journalistique informelle ETAPE, parole familière NCCFr).

La figure 4 montre le taux de réduction des phones en fonction de la position du phone dans le mot (« Deb », « Mil », « Fin », « 1Ph »). Les résultats montrent que, de manière générale, la position qui favorise le plus la réduction est « Mil ». Les résultats du GLM montrent que, tout en considérant d'autres facteurs de variation, la probabilité d'observer une réduction diminue significativement pour « Deb » [log odds ratio = -0,5401, $|Z|$ = 203,96, p<0, 001], pour « Fin » [log odds ratio = -0,1913,

$|Z|$ = 76,85, p<0,001] et pour « 1Ph » [log odds ratio = -0,4052, $|Z|$ = 97,31, p<0,001], par rapport à celle observée pour « Mil ».

3.5 Position du mot dans le groupe syntaxique

La figure 5 illustre le taux de réduction en fonction de la position du mot dans le groupe syntaxique. Il est intéressant de noter que, comme le montre la figure 4, la position qui favorise le plus la réduction est « I » (milieu). Plus précisément, les phones à l'intérieur des mots qui se situent au milieu d'un groupe syntaxique sont plus susceptibles de subir la réduction. Les résultats du GLM confirment que la probabilité d'observer la réduction diminue significativement pour « B » [log odds ratio = -0,0654, $|Z|$= 23,11, p<0,001], pour « L » [log odds ratio = -0,2532, $|Z|$= 108,25, p<0,001] et pour « U » [log odds ratio = -0,1747, $|Z|$= 49,35, p<0,001], par rapport à ce qui est observé pour « I ». Il convient de mentionner que le groupe « U » est composé de mots multisyllabiques tels que « aujourd'hui », « également », « effectivement », ainsi que de mots monosyllabiques tels que « et » et « tout ».

3.6 Style de parole

L'influence du style de parole sur la réduction peut être observée à partir des figures 2 à 5. Les résultats du GLM confirment que la probabilité d'observer une réduction augmente de manière significative pour la parole journalistique informelle [log odds ratio = 0,1882, $|Z|$= 89,59, p<0,001] et pour la parole familière [log odds ratio = 0,4298, $|Z|$= 159,98, p<0,001], par rapport à ce qui a été observé pour la parole journalistique formelle.

4 Conclusions

Cette étude porte sur la réduction de la parole et sur les facteurs susceptibles de favoriser la réduction en parole continue en français. Trois grands corpus de styles de parole différents (du formel au familier) ont été analysés. Les données de parole ont été automatiquement segmentées en mots et en phones, en mode d'alignement forcé. L'alignement automatique permet de localiser de manière innovante et efficace les zones de réduction de la parole. Les résultats montrent que le débit de parole, la catégorie de mots, la position du phone dans le mot, la position du mot dans le groupe syntaxique et le style de parole ont tous un impact significatif sur la réduction de la parole. Plus précisément, il est plus probable d'observer la réduction dans la parole produite avec un débit de parole élevé. En outre, il est plus probable d'observer la réduction dans les mots grammaticaux que dans les mots lexicaux. La position interne des mots ou des groupes syntaxiques favorise également la réduction. Enfin, le style de parole joue un rôle important en termes de réduction de la parole. Les contextes moins formels ont tendance à déclencher plus de réductions. En effet, la réduction peut prendre la forme d'une fusion, d'une suppression ou d'une recombinaison de segments ou de syllabes. Ces mécanismes peuvent potentiellement être liés à divers processus phonologiques, ce qui pourrait être intéressant sur le plan phonologique. Dans l'ensemble, cet article nous permet d'acquérir des connaissances sur la réduction et sur sa localisation en parole continue. En outre, notre recherche peut fournir des indications précieuses sur la manière dont les segments de la parole peuvent être réduits tout en restant compréhensibles pour les auditeurs.

Références

ADDA-DECKER M. & LAMEL L. (2000). The use of lexica in automatic speech recognition. In *Lexicon Development for Speech and Language Processing*, p. 235–266. Springer.

ADDA-DECKER M. & LAMEL L. (2017). Discovering speech reductions across speaking styles and languages. In *Rethinking reduction : Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Walter de Gruyter GmbH & Co KG.

DUEZ D. (1997). Acoustic markers of political power. *Journal of Psycholinguistic Research*, **26**(6), 641–654.

ERNESTUS M. T. C. (2000). *Voice assimilation and segment reduction in casual Dutch : A corpus-based study of the phonology-phonetics interface*. Thèse de doctorat, LOT, Utrecht.

GALLIANO S., 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 LREC*, volume 6, p. 315–320.

GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The limsi broadcast news transcription system. *Speech communication*, **37**(1), 89–108.

GENDROT C., GERDES K. & ADDA-DECKER M. (2016). Détection automatique d'une hiérarchie prosodique dans un corpus de parole journalistique. *Langue Francaise*, **191**, 123–147.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC-Eighth international conference on Language Resources and Evaluation*.

HAWKINS S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of phonetics*, **31**(3-4), 373–405.

JOHNSON K. (2004). Massive reduction in conversational american english. In *Spontaneous speech : Data and analysis. Proceedings of the 1st session of the 10th international symposium*, p. 29–54 : Tokyo, Japan : The National International Institute for Japanese Language.

KOHLER K. J. (1999). Articulatory prosodies in german reduced speech. p. 89–92.

MEUNIER C. & ESPESSER R. (2011). Vowel reduction in conversational speech in french : The role of lexical factors. *Journal of Phonetics*, **39**(3), 271–278.

NELDER J. A. & WEDDERBURN R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, **135**(3), 370–384.

NIEBUHR O. & KOHLER K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, **39**(3), 319–329.

PLUYMAEKERS M., ERNESTUS M. & BAAYEN R. H. (2005). Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, **118**(4), 2561–2569.

R DEVELOPMENT CORE TEAM (2019). *R : A La nguage and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAYMOND W. D., DAUTRICOURT R. & HUME E. (2006). Word-internal/t, d/deletion in spontaneous speech : Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language variation and change*, **18**(1), 55–97.

SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

SCHUPPLER B., VAN DOMMELEN W. A., KOREMAN J. & ERNESTUS M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final/t : Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, **40**(4), 595–607.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**(3), 201–212.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien

Antoine Caubrière[1] Elodie Gauthier[2]
(1) Orange Innovation, 2 Av. de Belle Fontaine, 35510 Cesson-Sévigné
(2) Orange Innovation, 2 Av. Pierre Marzin, 22300 Lannion
antoine.caubriere@orange.com elodie.gauthier@orange.com

RÉSUMÉ

Les approches auto-supervisées ont conduit à des avancées majeures dans le domaine de l'apprentissage profond. Par l'exploitation d'une grande quantité de données non annotées, ces approches ont notamment permis des améliorations dans des contextes peu dotés. Toutefois, les langues africaines restent majoritairement sous-représentées dans les jeux de données de préentraînement publiquement distribués. Dans ces travaux, nous préentraînons des modèles de parole auto-supervisés multilingues à partir de langues subsahariennes exclusivement. Nous étudions la pertinence des représentations apprises sur la tâche de reconnaissance de parole, en utilisant le jeu d'évaluation FLEURS-102. Notre modèle HuBERT$_{base}$ obtient des résultats similaires face à l'approche multilingue w2v-bert de FLEURS, tout en étant plus efficient, avec 6 fois moins de paramètres et 7 fois moins de données. Nous présentons aussi un second modèle exploitant une sous-sélection équilibrée des données initiales, obtenant des performances compétitives avec près de 80 fois moins de données de préentraînement.

ABSTRACT

Multilingual speech representation by self-supervised learning for sub-Saharan languages.

Self-supervised approaches are now unmissable and represent a major advance in deep learning. While self-supervised approaches have shown strong gains in a low-resource setting by leveraging the large amount of unlabeled data available on the web, languages spoken in sub-Saharan Africa (SSA) are still underrepresented in the datasets used in publicly available pre-trained models. In this paper, we build a multilingual pre-trained SSL model that uses only speech data in local languages spoken in SSA. We conducted experiments for downstream speech recognition task on the SSA subset of the FLEURS-102 dataset. Experiments conducted on speech recognition shown that our model, based on the HuBERT$_{base}$ architecture, obtains competitive results on the FLEURS dataset compared to the multilingual pre-trained w2v-bert-51 model, while being more efficient by using 7x less data and 6x less parameters. We trained another model with 80x less data, by using an equilibrated data selection.

MOTS-CLÉS : Apprentissage auto-supervisé, Langues subsahariennes, Reconnaissance de la parole multilingue, HuBERT.

KEYWORDS: Self-supervised representation, African languages, Multilingual ASR, HuBERT.

1 Introduction

Récemment, les approches auto-supervisées ont montré leur potentiel pour la mise en place de systèmes de reconnaissance de la parole performants (Chung *et al.*, 2021; Conneau *et al.*, 2021; Pratap *et al.*, 2023). Ce type d'approche permet l'exploitation d'une grande quantité de données non transcrites pour l'apprentissage d'une représentation dense de la parole. Elles sont plus riches que certaines caractéristiques classiques comme les MFCC ou les bancs de filtres. Un modèle pré-entraîné de façon auto-supervisé peut ensuite être utilisé soit comme un encodeur de parole, dont les paramètres seront adaptés, soit comme un extracteur de caractéristiques qui sera figé. Indépendamment de l'utilisation finale d'un modèle pré-entraîné, les données exploitées pour son apprentissage impacteront ses performances sur les tâches finales (Zhao & Zhang, 2022).

Les travaux de (Pires *et al.*, 2019) ont montré que le transfert d'apprentissage d'une langue bien dotée vers une langue sous-dotée est plus efficace lorsque les langues partagent des caractéristiques typologiques similaires. Toutefois, (Joshi *et al.*, 2020) fait état que 48% des caractéristiques typologiques répertoriées par le projet de classification WALS [1] (*World Atlas of Language Structures*) n'apparaissent pas dans les jeux de données. En complément, la plupart des modèles multilingues accessibles publiquement sont entraînés sur quelques langues seulement, ce qui entraîne leur sur-représentation au détriment d'autres langues (Valk & Alumäe, 2021; Babu *et al.*, 2022; Conneau *et al.*, 2023; Zhang *et al.*, 2023). Encore sous-dotées de nos jours, les langues africaines, de par leurs richesses et la présence de caractéristiques uniques, sont fortement impactées par cette situation (Clements & Rialland, 2007; Yadav & Sitaram, 2022).

Ces dernières années, l'intérêt des langues africaines est grandissant au sein de la communauté du traitement des langues. En surpassant des modèles multilingues pré-appris en majorité sur des données en anglais, plusieurs études ont montré l'intérêt d'un modèle pré-appris principalement sur les langues africaines (Ogueji *et al.*, 2021; Adelani *et al.*, 2022; Dossou *et al.*, 2022; Adebara *et al.*, 2022). En traitement de la parole, de nouveaux challenges et ressources sont publiés (Gutkin *et al.*, 2020; Sikasote & Anastasopoulos, 2021; Boito *et al.*, 2022; Olatunji *et al.*, 2023; Wanjawa *et al.*, 2023). Dans le cadre de la tâche de reconnaissance de la parole, les travaux de (Ritchie *et al.*, 2022) ont permis de meilleures performances en exploitant une approche multilingue auto-supervisée par rapport à une approche plus classique.

En phase avec tous ces travaux, nous proposons dans ce papier un modèle de représentation de la parole multilingue centré sur les langues africaines. Plus particulièrement, en utilisant des données exclusivement issues de la région subsaharienne, nous pré-entraînons un modèle fait pour être adapté à des tâches de traitement de la parole pour ces langues. Nous pré-entraînons également un second modèle pour lequel nous équilibrons les données du jeu d'apprentissage en fonction des langues et du genre des locuteurs. L'objectif est de produire un modèle non orienté vers un groupe de langues ou de locuteurs en raison d'une sur-représentation. Dans le cadre de nos expérimentations, nous nous attachons à résoudre la tâche de reconnaissance de la parole pour évaluer la pertinence des représentations fournies par nos modèles auto-supervisés.

Dans ce papier, nous présentons tout d'abord le jeu de données brutes – non transcrit –, que nous avons construit, avant d'apporter des détails sur l'architecture employée. Nous poursuivons ensuite par la description de nos expériences, ainsi que des résultats obtenus. Nous terminons en comparant nos systèmes avec le modèle de référence proposé par (Conneau *et al.*, 2023), avant de conclure.

1. https ://wals.info/feature

2 Jeux de données

2.1 Données non transcrites

Nous avons collecté, sur le web, des données issues de plusieurs sources émises dans des pays d'Afrique subsaharienne. Ces données correspondent à des journaux d'informations diffusés en ligne, qui traitent de divers sujets comme la politique, l'environnement, la santé, l'éducation, l'actualité régionale. Ces sujets sont abordés sous forme d'interviews, de débats et d'émissions. Ces journaux d'information s'engagent à favoriser la liberté d'expression, la mixité et le dialogue entre les cultures et à contribuer à l'égalité entre femmes et hommes. Dans ces travaux, néanmoins, ces données ne sont utilisées qu'à des fins de pré-apprentissage, afin de construire une représentation acoustique multilingue pertinente pour le traitement des langues parlées dans la zone. La teneur lexicale et sémantique des enregistrements audio n'est ainsi jamais exploitée. Les enregistrements récoltés permettent de couvrir un ensemble de 21 langues et variantes.

Parmi ces données, nous pouvons trouver des enregistrements en environnement contrôlé (studio), des interviews bruitées (extérieur), ainsi que des éléments non relatifs à la parole comme de la musique. Notre jeu de données mélange ainsi de la parole préparée, de la parole spontanée, ainsi que des segments de parole bruitée. Afin d'isoler les segments de parole, nous avons appliqué une détection d'activité vocale à l'aide de l'outil *pyannote* (Bredin, 2023). Nous effectuons aussi des pré-traitements permettant l'uniformisation des enregistrements (format, fréquence d'échantillonnage, ...).

L'ensemble de nos pré-traitements nous permet de construire un jeu de données de plus de 59 500h de parole exploitable pour un apprentissage auto-supervisé.

Nous donnons la répartition brute par langue dans la figure 1. Les langues sont identifiées par leur code de langue issu de la norme ISO 639-3 et les variantes sont considérées comme des langues distinctes. La catégorie "Unknown" correspond à des segments provenant d'enregistrements mixant plusieurs des 21 langues considérées, avec la présence d'alternance codique sur une partie des segments. Nous n'avons pas appliqué d'algorithme d'identification de langue. La langue française "FRA" correspond à des données accentuées produites par des locuteurs africains.

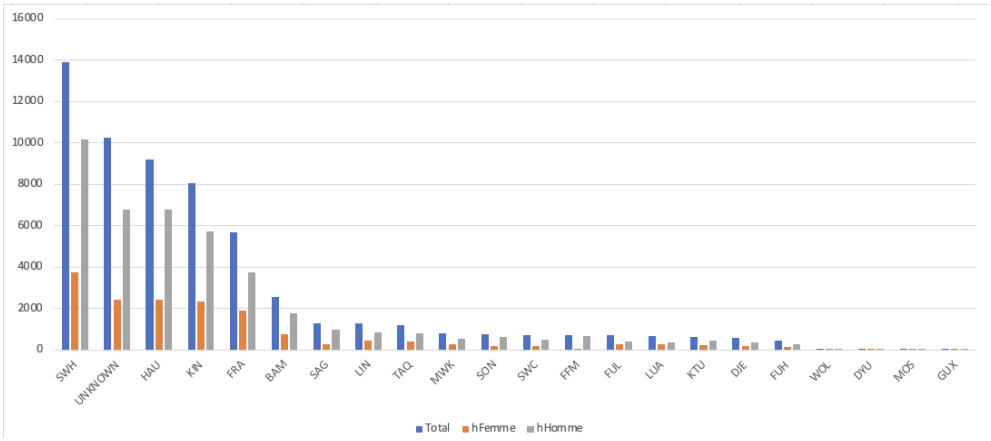


FIGURE 1 – Répartition brute du jeu d'apprentissage en heures par langues / variantes.

En complément des données brutes, nous produisons une sous-sélection de notre jeu de données la

plus équilibrée possible. Cette répartition équilibrée vise à produire un modèle auto-supervisé le moins orienté possible vers un groupement de langues sur-représentées dans les données d'apprentissage. Pour chaque langue, nous regroupons, dans la mesure du possible, 400 heures de segments de parole équilibrées selon les critères de langues et de genre des locuteurs. Nous appliquons un algorithme de détection du genre basé sur l'apprentissage fin d'un modèle XLSR-53 (Conneau *et al.*, 2021), sur librispeech (Panayotov *et al.*, 2015), pour produire l'annotation en genre des segments de notre corpus non supervisé [2]. Nous considérons les variantes d'une même langue comme appartenant à la langue et nous accumulons en quantité équivalente par variantes.

Cet équilibrage nous permet de construire un second jeu de données totalisant 5660 heures de parole. Nous donnons la répartition du jeu de données équilibré par langues dans la figure 2.

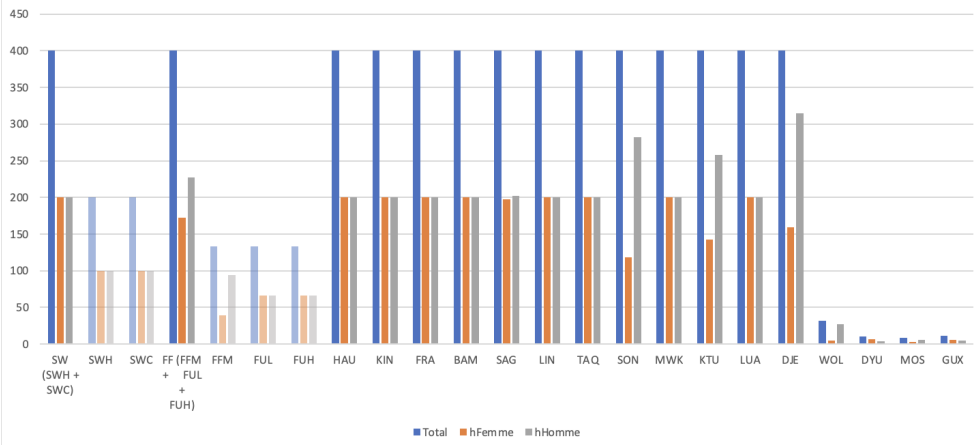


FIGURE 2 – Répartition équilibrée du jeu d'apprentissage en heures par langues.

2.2 Données transcrites

Dans le cadre de cette étude, nous exploitons le jeu d'évaluation FLEURS (Conneau *et al.*, 2023). Il s'agit d'un corpus au sein duquel 102 langues sont représentées et pour lesquelles environ 12h de parole transcrite est fournie (par langue). Nous nous concentrons plus particulièrement sur le sous-ensemble de données relatif à la région subsaharienne (FLEURS$_{SSA}$). Ces données sont composées de segments de parole couvrant 20 langues différentes, dont 5 sont communes aux données non transcrites. Dans ce papier, nous exploitons les transcriptions normalisées fournies par le corpus.

3 Système

3.1 Pré-apprentissage

Le système mis en place dans cette étude correspond à l'architecture "*base*" de l'approche HuBERT (Hsu *et al.*, 2021). Cette approche est auto-supervisée par l'annotation automatique en étiquette cible.

2. https ://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

Cette annotation est produite pour l'ensemble des segments de parole du corpus non transcrit, à l'aide de l'algorithme des K-moyennes.

L'architecture "*base*" correspond à un encodeur convolutif complété par 12 couches de type *transformers* produisant des plongements de taille 768. Elles possèdent 8 têtes d'attention et des linéaires internes de 3 072 unités. Cet ensemble de paramètres conduit à un modèle HuBERT d'environ 95 millions de paramètres.

L'apprentissage d'un système de ce type s'effectue en plusieurs étapes :

1. Apprendre un modèle non supervisé à partir des K-moyennes sur 100 classes, en exploitant les descripteurs acoustiques MFCC.
2. Créer des étiquettes auto-supervisées avec les K-moyennes sur l'ensemble d'entraînement et de validation.
3. Apprendre de zéro le modèle *transformer* avec les étiquettes auto-supervisées.
4. Extraire les représentations neuronales intermédiaires des *transformers* ("*base*" = couche 6).
5. Reproduire les étapes 1. à 3. avec ces représentations neuronales et 500 classes.

3.2 Apprentissage fin

Lors de l'apprentissage fin, nous considérons le modèle pré-entraîné comme un encodeur de la parole qui ne sera pas gelé. Nous exploitons ses représentations neuronales, de taille 768, issues de la dernière couche *transformer*. Nous ajoutons par-dessus deux couches linéaires de 1 024 unités, ainsi qu'une couche de sortie softmax dont la taille dépend des langues traitées par le système final.

4 Expériences

Nous effectuons l'ensemble du préentraînement de notre système à l'aide de l'outil Fairseq (Ott *et al.*, 2019). Nous avons parallélisé les calculs sur 4 GPU A40. Le pré-entraînement brut (sur environ 60k heures de données) a duré pendant un peu plus de 35 jours (soit plus de 3 360 heures de calcul GPU) tandis que le préentraînement sur données équilibrées (sur environ 5k heures de données) a duré un plus plus de 12 jours (soit plus de 1 150 heures de calcul GPU). Dans les deux cas, et notamment en raison de limite de mémoire, nous exploitons 600 heures du jeu d'entraînement pour effectuer l'apprentissage du modèle des K-moyennes basé sur les représentations internes des *transformers*. Nous sélectionnons ces 600 heures aléatoirement en conservant la contrainte de proportionnalité des langues et des genres représentés. Nous nommons les modèles pré-entraînés en fonction de la quantité d'heures non-supervisée utilisée. Nous apprenons ainsi le modèle 60k sur les quelques 59 500 heures récoltées et le modèle 5k sur la sous-sélection équilibrée de ces quelques 5 600 heures de parole. Les deux modèles pré-entraînés sont disponibles publiquement sur la plate-forme Hugging Face [3].

L'apprentissage fin n'est pas dépendant du modèle pré-entraîné utilisé. Pour chacune des deux expérimentations, nous utilisons l'outil SpeechBrain (Ravanelli *et al.*, 2021) afin d'entraîner les système de reconnaissance de la parole pour chacune des langues du corpus FLEURS$_{SSA}$. En moyenne, et en fonction de la langue, le temps d'affinement de l'entraînement prend 10 heures sur un seul GPU RTX 3090. L'ensemble des apprentissages supervisés par langues sont regroupés sous les systèmes *60k* et *5k*.

3. Les URLs seront données lors de la version finale de l'article.

En complément, nous réalisons un apprentissage conjoint sur l'ensemble des 20 langues, sur un seul GPU RTX 3090 pendant 48 heures. Ce premier apprentissage supervisé sert de base commune pour un transfert d'apprentissage vers un apprentissage fin sur chacune des langues du corpus FLEURS$_{SSA}$. Les systèmes supervisés obtenus à la suite de cet apprentissage conjoint sont regroupés sous les systèmes *60k*$_{joint}$ et *5k*$_{joint}$.

4.1 Résultats

Nous compilons dans la table 1, les résultats de nos expérimentations sur l'ensemble de test de FLEURS$_{SSA}$, exprimés en taux d'erreur sur les caractères (CER) et sur les mots (WER). Nous regroupons les langues entre celles vues lors du pré-apprentissage et celles non vues. Nous ajoutons aussi, pour chacun des groupes, la moyenne des scores, ainsi que la moyenne globale (toutes langues confondues). Afin de se comparer le plus justement possible, nous appliquons une méthodologie similaire aux travaux de (Conneau *et al.*, 2023). Ainsi, nous n'exploitons pas de modèles de langue pour réorganiser les hypothèses émises par le système.

	CER				WER			
	60k	5k	60k$_{joint}$	5k$_{joint}$	60k	5k	60k$_{joint}$	5k$_{joint}$
Langues vues								
Peul	21,2	21,2	17,8	**17,7**	61,9	60,6	56,4	**55,4**
Haoussa	10,5	11,2	**9,0**	10,1	32,5	35,6	**29,4**	33,8
Lingala	8,7	8,7	**6,9**	7,4	24,7	24,2	**20,9**	21,4
Swahili	7,1	8,6	**5,5**	6,5	23,8	28,8	**20,3**	24,4
Wolof	19,4	19,2	**17,0**	17,3	55,0	54,2	50,7	**50,1**
Moyenne	13,4	13,8	**11,2**	11,8	39,6	40,7	**35,5**	37,0
Langues non vues								
Afrikaans	23,3	23,8	20,3	**19,9**	68,4	68,3	62,6	**61,1**
Amharique	15,9	15,5	14,9	**14,3**	52,7	51,4	49,0	**47,6**
Luganda	11,5	11,7	**10,7**	11,1	52,8	53,3	**50,3**	52,0
Igbo	19,7	20,9	**17,2**	**17,2**	57,5	57,9	52,9	**52,4**
Kamba	16,1	16,3	**15,6**	15,9	53,9	**53,7**	**53,7**	54,3
Luo	9,9	10,2	**8,2**	8,4	38,9	38,5	34,9	**34,3**
Sotho du Nord	13,5	14,4	**11,7**	11,5	43,2	44,6	38,9	**38,6**
Chewa (Nyanja)	13,3	13,7	**10,9**	11,3	54,2	54,5	48,3	**48,1**
Oromo	22,8	22,9	**20,1**	21,2	78,1	77,4	74,8	**74,3**
Shona	11,6	11,2	**8,3**	8,7	50,2	48,2	**39,3**	39,7
Somali	21,6	21,9	**19,7**	20,0	64,9	64,5	**60,3**	60,6
Umbundu	21,7	21,7	**18,8**	20,7	61,7	60,8	**54,2**	57,0
Xhosa	11,9	12,4	**9,9**	10,1	51,6	52,3	**45,9**	47,1
Yoruba	24,3	25,0	**23,5**	23,8	67,5	68,0	**65,7**	66,6
Zoulou	12,2	12,4	**9,6**	10,0	53,4	53,0	**44,9**	46,1
Moyenne	16,6	16,9	**14,6**	14,9	56,6	56,4	**51,7**	52,0
Moyenne globale	15,8	16,1	**13,8**	14,1	52,3	52,5	**47,7**	48,2

TABLE 1 – Résultats obtenus sur les 20 langues subsahariennes de l'ensemble de test issu de FLEURS.

Les résultats présentés dans le tableau 1 montrent l'apport bénéfique de l'utilisation conjointe de données transcrites pour alimenter les modèles multilingues. Pour l'ensemble des langues, les performances sont bien meilleures après application du transfert d'apprentissage depuis les données conjointes. Pour les modèles 60k, nous notons respectivement une amélioration relative de 12,6% et de 8,8% sur les taux de CER et de WER moyens. Dans le cas des modèles 5k, il s'agit d'une amélioration relative de 12,4% (CER) et de 8,1% (WER).

Un second résultat intéressant de ces expérimentations concerne la comparaison entre les performances du modèle 60k$_{joint}$ et du modèle 5k$_{joint}$. 55 000 heures de données supplémentaires et plus de 22 jours de préentraînement GPU ont été consommés afin que le modèle 60k$_{joint}$ converge, pour finalement n'observer qu'un gain relatif de 1,8% (CER) et 1,0% (WER).Ce constat incite à s'interroger sur la quantité de données utiles ainsi que sur leur répartition lors du processus d'apprentissage, dans le cadre d'une approche multilingue. Ceci est d'autant plus intéressant au regard des questionnements sur la frugalité des approches, en termes de coût énergétique notamment.

Par comparaison des deux modèles "$_{joint}$", en ce qui concerne les langues vues lors du pré-apprentissage, nous remarquons une dégradation des performances sur les langues les plus représentées dans l'approche 60k (houaussa, lingala, swahili). En contrepartie, nous observons de meilleures performances sur les langues qui étaient sous-représentées dans l'approche 60k (peul et wolof). Ces résultats suggèrent que l'apprentissage du modèle 5k conduit à des représentations de la parole mieux réparties entre les langues.

4.2 Comparaison à FLEURS

En complément de nos résultats, nous confrontons ici nos systèmes aux résultats publiés par (Conneau *et al.*, 2023). Dans la mesure où les auteurs ne fournissent pas le détail des scores [4], nous effectuons la comparaison uniquement en termes de score moyen sur les caractères (CER), sur les 20 langues du sous-ensemble subsaharien proposé dans FLEURS$_{SSA}$. L'approche considérée par (Conneau *et al.*, 2023) est un w2v-bert de 600M de paramètres, pré-appris sur plus de 400 000 heures de parole.

	5k	60k	5k$_{joint}$	60k$_{joint}$	FLEURS$_{w2v-bert}$
Moyenne	16.1	15.8	14.1	13.8	13.6

TABLE 2 – Scores moyens sur l'ensemble test des données FLEURS$_{SSA}$.

Les résultats obtenus avec le système 60k$_{joint}$ montrent des performances très proches du modèle w2v-bert de FLEURS, tout en exploitant 7 fois moins de données et 6 fois moins de paramètres. Notre modèle 5k$_{joint}$, bien qu'obtenant des résultats inférieurs de 3,5%, se montre particulièrement compétitif eu égard au volume des données de préentrainement (80 fois moins de données que w2v-bert). L'utilisation de données exclusivement en langues locales, représentatives des parlers en Afrique subsaharienne, permet de réaliser un pré-apprentissage bien plus efficient. En ciblant des aspects particuliers – ici les particularités inhérentes au contexte africain –, ces résultats tendent à démontrer la pertinence d'une modélisation à l'échelle, face à l'apprentissage de modèles surdimensionnés.

4. Les résultats diffèrent entre la version déposée sur arXiv et la version IEEE du papier. Nous considérons ici la version officiellement validée et publiée par IEEE.

5 Conclusion

Dans ce papier, nous pré-entraînons le premier modèle multilingue auto-supervisé, librement partagé, appris exclusivement sur 21 langues et variantes africaines. Ces langues présentent des caractéristiques riches et non observées dans les autres langues du monde, notamment les langues occidentales. Nous confrontons notre modèle au jeu d'évaluation FLEURS qui propose un ensemble composé de langues parlées en Afrique subsaharienne. Face à leur modèle de référence, nous obtenons des résultats très similaires, tout en proposant une approche bien plus modérée en termes de coût d'apprentissage (près de 7 fois moins de données et 6 fois moins de paramètres). Dans ce même objectif de frugalité et d'efficience, nous proposons un second modèle (5k) entraîné à partir d'un sous-ensemble équilibré, en langue et en genre, de nos données. Le dimensionnement de ce modèle a permis de réduire de 2 200 heures le calcul GPU nécessaire, tout en contenant la perte de performance (dégradation moyenne de 0,3 points de CER face à notre modèle 60k, et 0,3 points face au modèle w2v-bert de FLEURS). Ce modèle se montre ainsi être un bon compromis entre matière d'efficience. Lors de travaux futurs, une analyse approfondie de ces systèmes sera menée, notamment au travers de l'étude de l'impact de l'équilibrage du jeu d'apprentissage sur la qualité des représentations dans un contexte multilingue, ainsi que de l'étude de l'impact du genre des locuteurs sur les performances des modèles de reconnaissance de la parole.

Références

ADEBARA I., ELMADANY A., ABDUL-MAGEED M. & INCIARTE A. A. (2022). Serengeti : Massively multilingual language models for Africa. *arXiv preprint arXiv :2212.10785*.

ADELANI D., NEUBIG G., RUDER S., RIJHWANI S., BEUKMAN M., PALEN-MICHEL C., LIGNOS C., ALABI J., MUHAMMAD S., NABENDE P., DIONE C. M. B., BUKULA A., MABUYA R., DOSSOU B. F. P., SIBANDA B., BUZAABA H., MUKIIBI J., KALIPE G., MBAYE D., TAYLOR A., KABORE F., EMEZUE C. C., AREMU A., OGAYO P., GITAU C., MUNKOH-BUABENG E., MEMDJOKAM KOAGNE V., TAPO A. A., MACUCWA T., MARIVATE V., ELVIS M. T., GWADABE T., ADEWUMI T., AHIA O., NAKATUMBA-NABENDE J., MOKONO N. L., EZEANI I., CHUKWUNEKE C., OLUWASEUN ADEYEMI M., HACHEME G. Q., ABDULMUMIN I., OGUNDEPO O., YOUSUF O., MOTEU T. & KLAKOW D. (2022). MasakhaNER 2.0 : Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4488–4508, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.298.

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J., BAEVSKI A., CONNEAU A. & AULI M. (2022). XLS-R : Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, p. 2278–2282. DOI : 10.21437/Interspeech.2022-143.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BOITO M. Z., BOUGARES F., BARBIER F., GAHBICHE S., BARRAULT L., ROUVIER M. & ESTÈVE Y. (2022). Speech resources in the tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2066–2071.

BREDIN H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. In *Proc. Interspeech 2023*.

CHUNG Y.-A., ZHANG Y., HAN W., CHIU C.-C., QIN J., PANG R. & WU Y. (2021). w2v-BERT : Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 244–250. DOI : 10.1109/ASRU51503.2021.9688253.

CLEMENTS G. N. & RIALLAND A. (2007). *Africa as a phonological area*, p. 36–85.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

CONNEAU A., MA M., KHANUJA S., ZHANG Y., AXELROD V., DALMIA S., RIESA J., RIVERA C. & BAPNA A. (2023). Fleurs : Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 798–805 : IEEE.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOSSOU B. F., TONJA A. L., YOUSUF O., OSEI S., OPPONG A., SHODE I., AWOYOMI O. O. & EMEZUE C. (2022). Afrolm : A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, p. 52–64.

GUTKIN A., DEMIRSAHIN I., KJARTANSSON O., RIVERA C. E. & TÚBÒSÚN K. (2020). Developing an open-source corpus of yoruba speech. In *Proc. of Interspeech 2020*, p. 404–408, October 25–29, Shanghai, China, 2020.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460. DOI : 10.1109/TASLP.2021.3122291.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv :2004.09095*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

OGUEJI K., ZHU Y. & LIN J. (2021). Small data ? no problem ! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, p. 116–126.

OLATUNJI T., AFONJA T., YADAVALLI A., EMEZUE C. C., SINGH S., DOSSOU B. F. P., OSUCHUKWU J., OSEI S., TONJA A. L., ETORI N. & MBATAKU C. (2023). AfriSpeech-200 : Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, **11**, 1669–1685. DOI : 10.1162/tacl_a_00627.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. *CoRR*, **abs/1904.01038**.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual BERT ? In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, p. 4996–5001, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1493.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M., BAEVSKI A., ADI Y., ZHANG X., HSU W.-N., CONNEAU A. & AULI M. (2023). Scaling speech technology to 1,000+ languages.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). Speechbrain : A general-purpose speech toolkit.

RITCHIE S., CHENG Y.-C., CHEN M., MATHEWS R., VAN ESCH D., LI B. & SIM K. C., Éds. (2022). *Large vocabulary speech recognition for languages of Africa : multilingual modeling and self-supervised learning*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SIKASOTE C. & ANASTASOPOULOS A. (2021). Bembaspeech : A speech recognition corpus for the bemba language.

VALK J. & ALUMÄE T. (2021). Voxlingua107 : a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, p. 652–658 : IEEE.

WANJAWA B. W., WANZARE L. D. A., INDEDE F., MCONYANGO O., MUCHEMI L. & OMBUI E. (2023). Kenswquad—a question answering dataset for swahili low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **22**(4). DOI : 10.1145/3578553.

YADAV H. & SITARAM S. (2022). A survey of multilingual models for automatic speech recognition.

ZHANG Y., HAN W., QIN J., WANG Y., BAPNA A., CHEN Z., CHEN N., LI B., AXELROD V., WANG G., MENG Z., HU K., ROSENBERG A., PRABHAVALKAR R., PARK D. S., HAGHANI P., RIESA J., PERNG G., SOLTAU H., STROHMAN T., RAMABHADRAN B., SAINATH T., MORENO P., CHIU C.-C., SCHALKWYK J., BEAUFAYS F. & WU Y. (2023). Google usm : Scaling automatic speech recognition beyond 100 languages.

ZHAO J. & ZHANG W.-Q. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1227–1241. DOI : 10.1109/JSTSP.2022.3184480.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Retour auditif interne de la production de parole : mesures préliminaires de la vibration osseuse par accélérométrie et comparaison au son aérien

Raphaël Vancheri Coriandre Vilain
Nathalie Henrich Bernardoni Pierre Baraduc

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France
[prenom].[nom]@gipsa-lab.fr

RÉSUMÉ

Lorsqu'on parle, le retour auditif se décompose en une voie aérienne et une voie interne ou 'par conduction osseuse'. Un locuteur entend les deux composantes, contrairement au récepteur. Alors que la moitié du signal cochléaire est interne, on connaît mal l'information qu'il véhicule et comment elle impacte le contrôle moteur oral. Dans cette étude, nous considérons deux indicateurs du signal auditif interne pendant la production de parole, la vibration des dents de la mâchoire supérieure et le son enregistré près du tympan. Une méthode de conversion de voix nous permet d'évaluer les différences informationnelles entre voix aérienne et voix "osseuse" interne. Comme observé précédemment par la simple méthode d'enregistrement péritympanique, la somme des retours acoustiques aérien et interne amène une lisibilité supérieure des trajectoires formantiques qui pourrait faciliter le contrôle de la production de parole.

ABSTRACT

Internal acoustic feedback during speech production : preliminary measure of bone vibration during speech with an accelerometer and comparison to aerial-conducted sound.

During speech, the auditory feedback involves both an aerial component picked up by the external ear, and an internal vibration : the 'bone conduction' component. While a speaker hears both components, a listener only hears the aerial part. Although half of the cochlear signal comes from internal conduction, the information it conveys, and how it impacts oral motor control, is still unclear. In this study, we considered two proxies of the internal auditory signal : the vibration of the upper teeth and the sound emitted next to the eardrum. A voice conversion method allows to evaluate the informational differences between aerial and internal bone voice. As preceedingly observed with the peritympanic recording method, the summation of internal and aerial feedback leads to clearer formantic trajectories, which may facilitate speech motor control.

MOTS-CLÉS : conduction osseuse, production de parole, perception, contrôle moteur.

KEYWORDS: bone conduction, speech production, perception, motor control.

1 Introduction

Durant la production de parole, les tissus mous péri-oraux ainsi que les os sous-jacents transmettent un signal acoustique interne jusqu'à la cochlée de manière directe (vibration du rocher) et indirecte

(vibration des osselets, vibration du tympan). Ce signal interne est éminemment complexe à objectiver du fait de la quasi impossibilité de mesurer directement les vibrations acoustiques au niveau de la cochlée. L'essentiel de nos connaissances sur la conduction interne (communément mais improprement appelée conduction osseuse) provient d'études sur l'animal (p. ex. Tonndorf *et al.* 1966) ou sur des cadavres (Eeg-Olofsson *et al.*, 2008), qui ont permis de quantifier la conduction des vibrations dans les différentes structures physiologiques (Stenfelt & Goode, 2005), mais bien évidemment pas pendant une vibration laryngée. Pörschmann (2000), à l'aide d'une méthode de masquage, a étudié le premier le spectre sonore de la parole interne, et mis en évidence une différence entre phones voisés et non-voisés. Plus récemment, Reinfeldt *et al.* (2010) ont utilisé l'enregistrement direct des vibrations du conduit auditif comme un indicateur du signal par conduction osseuse, et observé les différences de spectre sonore entre une petite sélection de voyelles ou consonnes voisées isolées. Nous avons précédemment utilisé cette méthode pour décrire le contenu spectral de signaux aériens et péri-tympaniques pendant la production de parole continue (Baraduc & Vilain, 2022). Toutefois ces travaux n'ont mesuré que la vibration des tissus mous ; une partie du signal interne restait inconnue. Dans cet article, nous décrivons une méthode pour dépasser cette limite, et les résultats préliminaires obtenus sur 3 sujets pilotes.

2 Matériel et méthodes

2.1 Sujets

Trois sujets francophones du laboratoire (1 femme, 2 hommes, 24-49 ans) parmi les auteurs, ont participé sans compensation à cette expérience en tant que sujets pilotes.

2.2 Dispositif expérimental

Les sujets ont été équipés d'un capteur acoustique intra-auriculaire (microphone capillaire Etymotic ER-7C) permettant d'enregistrer le signal sonore péritympanique. Ce signal est isolé acoustiquement de l'environnement extérieur par une boite fabriquée au laboratoire, placée contre l'oreille du sujet et écrantant les signaux extérieurs d'au moins 30 dB sur l'ensemble du spectre sonore ; son volume est suffisant pour éviter les résonances (effet d'occlusion) que généreraient de simples bouchons d'oreille. Les sujets sont également équipés d'un accéléromètre miniature (PCB Piezotronics 352A73 conditionné par une carte d'acquisition Data Translation) fixé par une résine photopolymérisable sur une dent de la mâchoire supérieure (incisive, canine ou prémolaire en fonction de configuration de la surface dentaire). Cet accéléromètre mesure les vibrations locales des os, après leur passage à travers le ligament alvéolo-dentaire et la dentine. Trois microphones aériens (BK 4189) sont également utilisés pour mesurer les signaux acoustiques : à 50 cm de la bouche du sujet (dans l'axe), à 2 cm de l'oreille droite (dans l'espace extérieur), à 2 cm de l'oreille gauche (dans la boîte isolante). L'acquisition des données est effectuée sous Matlab à l'aide de la PsychoPhysics Toolbox (pour les données audio) et d'une toolbox DataTranslation pour les données d'accélérométrie.

2.3 Protocole

Après une étape d'équilibration perceptive ne se rapportant pas aux résultats présentés ici, les sujets devaient lire à voix haute les 100 premières phrases du corpus FHarvard (Aubanel *et al.*, 2020), présentées sur un moniteur placé en face d'eux.

2.4 Analyse des données

Les signaux ont été synchronisés par intercorrélation (avec pour signal de référence l'enregistrement du microphone placé face au sujet), puis nous avons édité parallèlement tous les signaux afin d'en enlever les répétitions et erreurs. Les signaux d'enregistrement intra-auriculaire et d'accélérométrie ont été amplifiés (de 15 dB et 40 dB respectivement) ; ce dernier a préalablement été filtré passe-haut à 5 Hz afin d'en retirer la composante continue. Enfin, les deux signaux ont également été débruités par soustraction spectrale (-15 dB, par Audacity). Les enregistrements ont été ensuite segmentés sous Praat au moyen d'un outil d'alignement phonétique automatique, EasyAlign (Goldman, 2010) ; les résultats ont été manuellement vérifiés et édités au besoin. Afin de mieux cerner les différences en terme d'information portée par les signaux aérien et osseux, nous avons cherché à déterminer dans quelle mesure ils étaient interconvertibles. Cette conversion de voix a été réalisée par régression par mélange de gaussiennes à partir de coefficients mel-cepstraux (trames de 5 ms), grâce à un code partagé par T. Hueber (Hueber & Bailly, 2016), après sous-échantillonage à 16 kHz. Les différences spectrales en dB ont été calculées dans la bande 0–5 kHz. L'enveloppe spectrale de la parole est aisément calculée à partir des coefficients mel-cepstraux ; la resynthèse peut s'effectuer avec un filtre MLSA. Ces calculs ont été réalisés avec SPTK 4.0 (`github.com/sp-nitech/SPTK/releases`). La visualisation des résultats a été obtenue sous Matlab, en utilisant en particulier la fonction `spectrogram`.

3 Résultats

3.1 Observations de surface : signal aérien, osseux, et des tissus mous

Le signal de l'accéléromètre reflète-t-il des propriétés de la conduction interne des vibrations de parole ? En quoi diffère-t-il du l'enregistrement péritympanique, censé refléter surtout la vibration des tissus mous pendant la vocalisation ? La figure 1 donne un exemple de spectrogrammes des signaux de parole, selon qu'ils sont aériens ou tirés de l'accéléromètre (pour simplifier, nous utiliserons le raccourci abusif de "signal osseux" dans la suite des résultats), ou du signal aérien. Comme observé précédemment avec l'enregistrement péritympanique (Baraduc & Vilain, 2022), le signal interne permet de mieux suivre certaines transitions formantiques. Toutefois, alors que le signal des tissus mous était particulièrement différent pendant les fricatives, on peut déjà remarquer sur ces exemples que ce n'est pas le cas pour le signal osseux. Une comparaison directe des deux méthodes d'estimation du retour auditif interne n'est malheureusement pas possible sur ce jeu de données, les enregistrements péritympaniques ont été particulièrement décevants pour deux sujets (cf. infra).

Une deuxième observation a trait à un aspect technique important de l'estimation du retour par conduction interne. Les tissus mous ayant tendance à absorber les vibrations de haute fréquence, l'enregistrement péritympanique est délicat au-dessus de 4 kHz. Ceci n'est pas le cas avec l'accéléromètre, dont la plage de mesure s'étend jusqu'à 40 kHz. En fait, la densité spectrale de la parole osseuse n'est

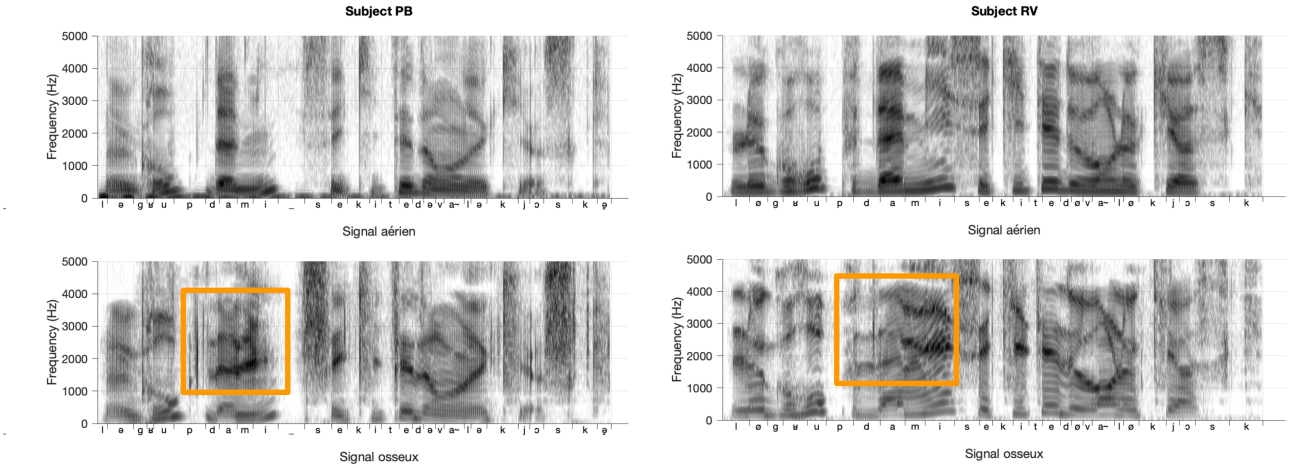


FIGURE 1 – Spectrogrammes de parole chez deux sujets, pour la phrase *"le groupe d'amis s'est quitté devant le kiosque"*. En haut, les spectrogrammes du signal aérien ; en bas les spectrogrammes du signal interne. Comme vu précédemment avec les enregistrements péritympaniques, la trajectoire de F2 est particulièrement lisible dans le signal interne pendant *d'amis* (encadré orange).

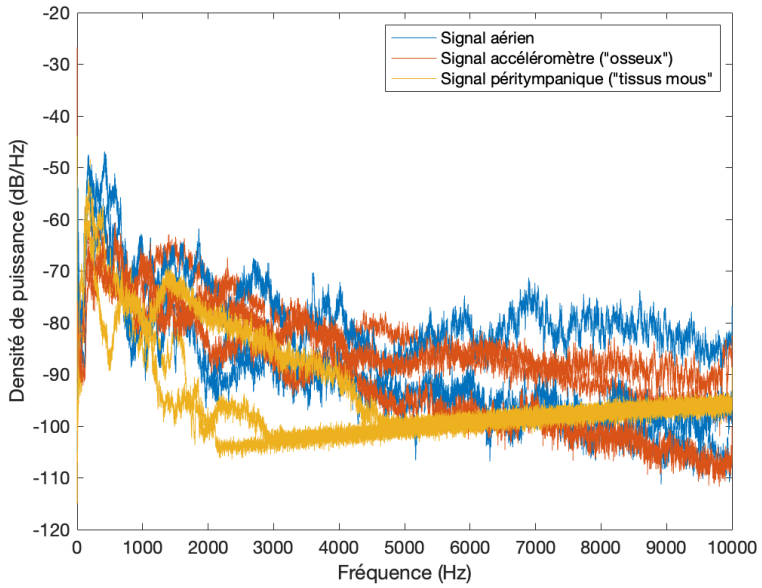


FIGURE 2 – Densité de puissance du signal de parole, pour les 3 sujets de l'expérience pilote. En bleu, me signal aérien ; en rouge le signal de l'accéléromètre ("osseux") ; en jaune le signal péritympanique ("tissus mous")

pas différente de la parole aérienne, comme on le voit figure 2. Par ailleurs, on remarque que pour un sujet, la densité de puissance péritympanique n'est pas très différente des autres signaux aériens ou osseux, mais est tronquée à 4,5 kHz par le plancher de bruit du microphone. L'amortissement des vibrations hautes fréquences par les tissus mous ne semble pas, ou au moins pas toujours, être responsable de cette difficulté à estimer le retour auditif interne de la parole dans le haut du spectre. Quoi qu'il en soit, en pratique, le signal de l'accéléromètre permet une bien meilleure estimation de la conduction interne des hautes fréquences, comme on peut l'observer également sur la figure 1.

3.2 Analyse de différences informationnelles

Afin de quantifier si le signal "osseux" porte une information redondante avec le signal aérien, nous avons utilisé une méthode de conversion de voix (cf. §2.4), comme dans nos travaux précédents. Brièvement, le principe en est le suivant : les différences spectrales entre signal "osseux" et signal aérien converti en "osseux" sont révélatrices de l'information *spécifique* portée par le signal osseux, puisqu'elles correspondent à des parties du spectre qu'on peut difficilement prédire à partir de la voix aérienne. L'argument est symétrique et l'information spécifique portée par le signal aérien peut être estimée par la conversion de voix réciproque "osseux"→ aérien.

La figure 3 permet de comparer des spectrogrammes de parole aérienne et "osseuse", avec une mise en évidence par la couleur des parties du spectre difficilement convertibles, donc spécifiques de chaque signal. Cet exemple peut être comparé à une figure similaire de (Baraduc & Vilain, 2022). On remarque que les transitions formantiques sont plus lisibles dans le signal "osseux", ce qui est particulièrement clair dans l'extrait illustré *"son dernier c[ongé]"*, pour lequel les /o~/, /d/, et la séquence /nj/ ont effectivement une mesure de spécificité osseuse élevée.

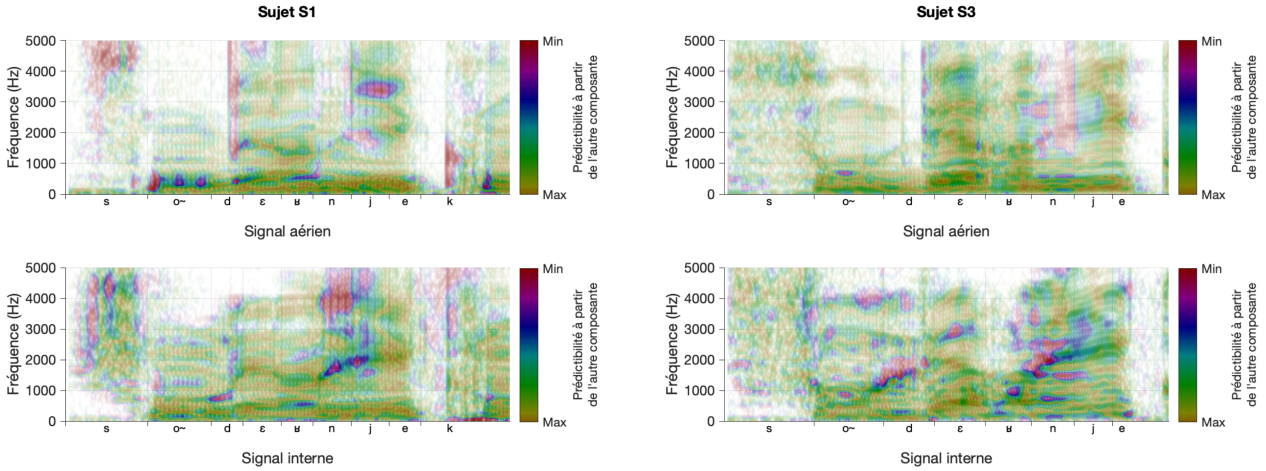


FIGURE 3 – Spectrogrammes du signal aérien (en haut) et du signal "osseux" (en bas), correspondant au début de la prononciation de la phrase *"Son dernier congé dura deux semaines"*. Les couleurs indiquent les différences statistiquement non interconvertibles (selon le code couleur représenté à droite). On remarque que les voyelles et consonnes nasales, les occlusives voisées ou même le /ʁ/ ont une signature osseuse particulière (et plus claire).

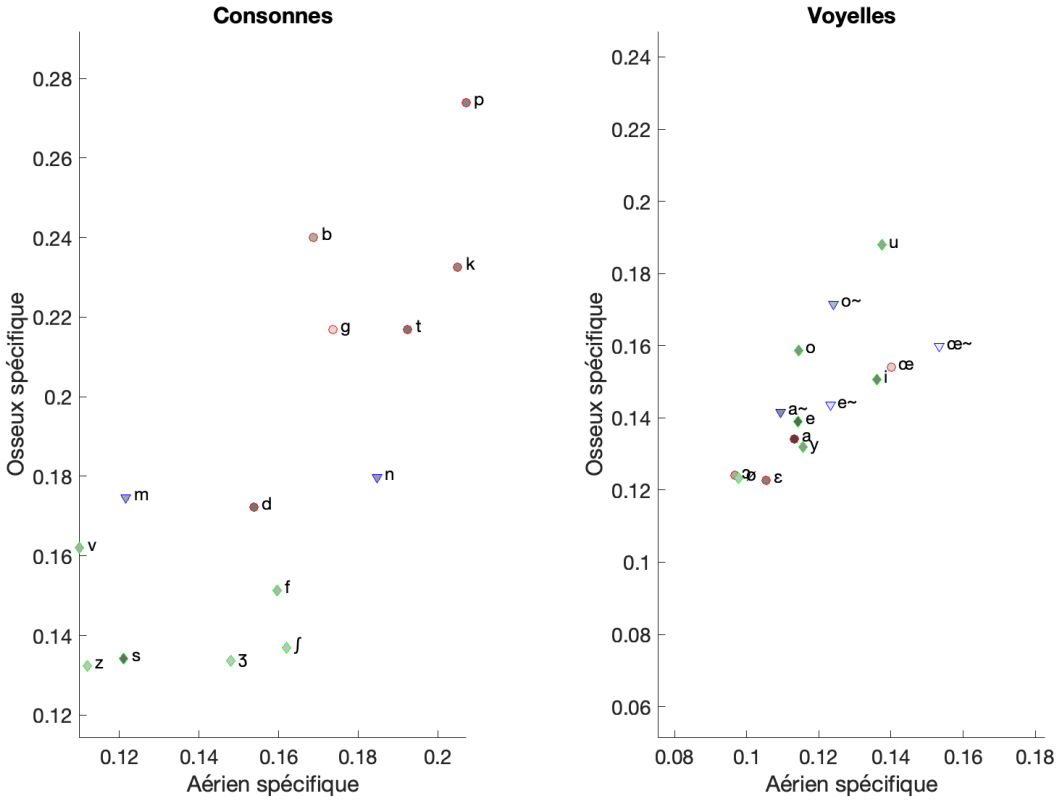


FIGURE 4 – Distribution des mesures de spécificité du signal aérien (axe horizontal) et du signal "osseux" (axe vertical) pour une sélection de phonèmes, consonnes (panneau de gauche) ou voyelles (panneau de droite). La spécificité est définie comme la distance maximale entre enveloppes spectrales du signal réel et du signal prédit par conversion depuis l'autre source. On remarque que les occlusives sont bien différenciées, les voyelles et consonnes nasales fermées également. Les fricatives portent peu d'information spécifique à la voie osseuse, en moyenne.

Pour donner un aperçu global, nous avons segmenté les signaux de parole et quantifié pour chaque phonème l'information spécifique portée par le signal aérien ou "osseux", en considérant comme mesure de spécificité la différence maximale d'enveloppe spectrale entre signal réel et signal prédit (converti), pendant la durée du phonème. Ces valeurs sont variables selon le contexte de coarticulation et le sujet, et ici, le nombre de participants étant limité, nous avons représenté la moyenne inter-sujets sur tout le jeu de données dans le graphe de la figure 4.

Les résultats confortent l'idée que le signal "osseux" est souvent différemment informatif du signal aérien. Les occlusives sont particulièrement différentes et différemment informatives dans les deux signaux. Une étude plus détaillée révèle que les occlusives non voisées portent souvent des traces de formants visibles dans le bruit de plosion, voire avant ; ceci est plus variable dans les consonnes voisées, dont certains exemples sont particulièrement clairs quant à la transition formantique en cours, d'autres moins. Les voyelles fermées, particulièrement lorsqu'elles sont postérieures ou nasales, ont aussi une signature assez spécifique à chaque composante auditive. Enfin, on remarque que les fricatives portent peu d'information spécifique dans le signal osseux, à l'inverse de ce que nous avions constaté précédemment dans le signal péritympanique. L'origine de cette différence devra être élucidée, mais il est possible qu'elle soit liée à la position de l'accéléromètre par rapport à la cavité orale postérieure, la résonance de cette cavité excitée par le bruit aéro-acoustique étant nette dans les enregistrements péritympaniques.

4 Discussion

Ces premiers résultats prolongent notre précédente description des différences entre retour acoustique aérien de la parole naturelle, et retour par conduction interne. Ils sont un aperçu des développements méthodologiques en cours qui devraient nous permettre de mieux caractériser le retour auditif de sa propre voix, en associant deux techniques complémentaires d'évaluation de la vibration interne. Si cet article décrit des travaux préliminaires, les techniques mises en œuvre ont également des limites que nous rappelons ici.

Une première limite tient au placement des capteurs. Si la mesure de la vibration des tissus mous est a priori idéale près du tympan, la mesure de la vibration des os de la tête serait probablement meilleure à une position plus proche de l'os temporal ; ceci est toutefois difficilement compatible avec un positionnement simple de l'accéléromètre. Par ailleurs il faut rappeler que la mesure intègre l'élasticité du ligament odonto-alvéolaire, dont les caractéristiques mécaniques pourraient affecter le signal (le graphe de la figure 2 étant toutefois assez rassurant).

Ensuite, notre mesure de différence informationnelle par la conversion de voix a également un certain nombre de limites. Tout d'abord, elle peut être biaisée par le modeste corpus d'entraînement (seulement 5 minutes de parole). Nous travaillons actuellement sur une mesure dérivée de la théorie de l'information, qui sera indépendante de la spécification d'un modèle. D'autre part, nous avons résumé dans la figure 4 les différences spectrales en considérant leur somme sur la bande 0–5 kHz, ce qui ne rend pas compte du caractère éventuellement crucial d'une information sonore particulière à l'intérieur de ce spectre.

Néanmoins, ces premiers résultats d'estimation directe de la vibration osseuse nous semblent très encourageants. Un des intérêts d'évaluer le retour interne par une méthode d'accélérométrie est de libérer le sujet d'un capteur situé dans l'oreille (et d'une boîte acoustiquement isolée), ce qui permet

de simplifier les expériences sur la parole et la voix, voire la pratique musicale au sens large. A court terme, nous chercherons toutefois à mieux caractériser les différences entre retour interne purement osseux et conduction via les tissus mous, et évaluer leur variabilité interindividuelle.

Remerciements

Les auteurs remercient Thomas Hueber pour sa disponibilité, ses conseils et le partage de son code Matlab de conversion de voix.

Références

AUBANEL V., BAYARD C., STRAUSS A. & SCHWARTZ J. (2020). The Fharvard corpus : A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication*, **124**, 68–74. DOI : https ://doi.org/10.1016/j.specom.2020.07.004.

BARADUC P. & VILAIN C. (2022). Retours acoustiques de la production de parole : caractérisation des différences informationnelles entre le son aérien et le son par conduction osseuse. In *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, p. 980–988. DOI : 10.21437/JEP.2022-104.

EEG-OLOFSSON M., STENFELT S., TJELLSTRÖM A. & GRANSTRÖM G. (2008). Transmission of bone-conducted sound in the human skull measured by cochlear vibrations. *International Journal of Audiology*, **47**(12), 761–769. DOI : doi :10.1080/14992020802311216.

GOLDMAN J.-P. (2010). Easyalign : a friendly automatic phonetic alignment tool under praat. In *Proc. Interspeech 2011*, p. 3233–3236. DOI : 10.21437/Interspeech.2011-815.

HUEBER T. & BAILLY G. (2016). Statistical conversion of silent articulation into audible speech using full-covariance hmm. *Computer Speech Language*, **36**, 274–293. DOI : https ://doi.org/10.1016/j.csl.2015.03.005.

PÖRSCHMANN C. (2000). Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice. *Acta Acustica united with Acustica*, **86**(6).

REINFELDT S., ÖSTLI P., HÅKANSSON B. & STENFELT S. (2010). Hearing one's own voice during phoneme vocalization—Transmission by air and bone conduction. *The Journal of the Acoustical Society of America*, **128**(2), 751–762. DOI : 10.1121/1.3458855.

STENFELT S. & GOODE R. L. (2005). Bone-Conducted Sound : Physiological and Clinical Aspects. *Otology & Neurotology*, **26**(6), 1245–1261. DOI : 10.1097/01.mao.0000187236.10842.d5.

TONNDORF J., GREENFIELD E. C. & KAUFMAN R. S. (1966). The Relative Efficiency of Air and Bone Conduction in Cats. *Acta Oto-Laryngologica*, **61**(sup213), 105–123. DOI : 10.3109/00016486609120802.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Synthèse de gestes communicatifs via STARGATE

Louis ABEL[1] Vincent COLOTTE[1] Slim OUNI[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
louis.abel@loria.fr, vincent.colotte@loria.fr, slim.ouni@loria.fr

RÉSUMÉ

La synthèse de gestes liés à la parole est un domaine de recherche en pleine expansion. Cependant, les nouveaux systèmes utilisent souvent des architectures complexes, les rendant souvent inadaptés à leur utilisation dans des agents conversationnels incarnés ou dans d'autres domaines de recherche comme la linguistique, où le lien entre la parole et les gestes est difficile à étudier manuellement. Cet article présente STARGATE, une nouvelle architecture tirant parti de l'autorégression pour fournir des capacités en temps réel, mais aussi des convolutions de graphe couplées à l'attention pour incorporer des connaissances structurelles explicites et permettre une forte compréhension spatiale et temporelle du geste. Nous avons démontré que notre modèle est capable de générer des gestes convaincants en surpassant l'état de l'art dans une étude quantitative, tout en obtenant des scores légèrement meilleurs en termes de cohérence et de crédibilité des gestes générés liés à la parole sur une étude perceptive.

ABSTRACT

Co-Speech gestures synthesis using STARGATE

Co-speech gestures synthesis is a growing field of research. However, new systems often use complex or heavy architecture, making them unsuitable for incorporation into Embodied Conversational Agents (ECAs) or for interpretation in other research fields such as linguistics, where the link between speech and gestures is difficult to research manually. This paper presents STARGATE, a novel architecture for Spatio-Temporal Autoregressive Graph from Audio-Text Embeddings. The model takes advantage of autoregression to provide real-time capabilities, but also graph convolutions coupled with attention to incorporate explicit structural prior knowledge and enable efficient spatial and temporal processing. The model was evaluated against state-of-the-art models in both perceptive and quantitative studies. We demonstrated that our model is capable of generating convincing gestures by outperforming the state-of-the-art in a quantitative study, while achieving slightly better scores in terms of consistency and credibility of the generated gestures related to speech.

MOTS-CLÉS : Apprentissage profond, Synthèse de gestes, Synthèse audiovisuel de la parole.

KEYWORDS: Deep learning, Gestures synthesis, Audiovisual speech synthesis.

1 Introduction

La synthèse de gestes à partir de la parole est un domaine émergent qui a fait l'objet d'une attention particulière au cours des dernières années. Bien que les mécanismes mettant en œuvre la relation entre la production des gestes et la parole sont encore peu connus, des progrès considérables ont été faits dans le développement de techniques de génération de gestes à partir de la parole. L'omniprésence des gestes dans la communication humaine souligne leur importance dans les interactions humaines naturelles. Afin de capturer l'essence des gestes humains et de les intégrer dans des systèmes de communication artificiels, plusieurs chercheurs ont analysé et tenté de classifier les gestes. Au départ, des systèmes basés sur des règles ont été utilisés pour développer des agents conversationnels incarnés

(ECA) (Cassell, 2001), en s'appuyant sur les connaissances des neurosciences et de la linguistique. Les premiers systèmes étaient rudimentaires et souvent incompatibles avec les conclusions émises dans la littérature. L'absence d'un système de classification unifié pour les gestes (par exemple (McNeill, 1992; Kendon, 2004; Boutet, 2008)) et les conclusions disparates concernant la relation entre les gestes et la parole ((So *et al.*, 2009; de Ruiter *et al.*, 2012; Krauss & Hadar, 1999)) au sein de ces cadres ont entravé l'élaboration de règles cohérentes et fiables. Ces dernières années, les approches basées sur les données sont apparues comme une voie prometteuse pour extraire implicitement les modèles et les règles complexes qui régissent la relation entre la parole et le geste. Ces approches utilisent une variété d'architectures, allant des simples autoencodeurs (Takeuchi *et al.*, 2017; Kucherenko *et al.*, 2019) aux autoencodeurs variationnels (VAE) et aux VAE conditionnels (Li *et al.*, 2021; Lu *et al.*, 2023), pour couvrir une plus large distribution de gestes et un meilleur conditionnement à partir de la parole. Les systèmes basés sur la diffusion (Alexanderson *et al.*, 2023; Zhao *et al.*, 2023; Zhang *et al.*, 2023; Deichler *et al.*, 2023) ont également fait l'objet d'une attention particulière, produisant des séquences de gestes de haute qualité. Dans la littérature, le travail de (Alexanderson *et al.*, 2020), StyleGestures, se distingue par son architecture autorégressive innovante utilisant les flux de normalisation (Henter *et al.*, 2020). Les flux de normalisation sont un type spécialisé de réseau de neurones qui permet de modéliser efficacement des distributions complexes. Cette structure de réseau particulière a été largement reconnue comme une référence pour l'évaluation des performances des systèmes de synthèse gestuelle, comme en témoigne son adoption massive dans les recherches ultérieures (Li *et al.*, 2021; Alexanderson *et al.*, 2023; Ao *et al.*, 2022). Il a notamment été sélectionné comme référence pour le GENEA Challenge 2020 (Kucherenko *et al.*, 2021), un défi visant à faire progresser l'état de l'art en matière de synthèse gestuelle. Compte tenu de son architecture autorégressive et de son utilisation intensive en tant que référence, nous avons adopté StyleGestures comme référence de l'état de l'art pour nos comparaisons.

Malgré les recherches approfondies sur la synthèse de gestes liés à la parole, il n'existe que peu d'études sur l'explicabilité et l'interprétabilité de ces méthodes de génération des gestes, indépendamment de leur cohérence ou de leur complexité. Ce manque d'explication pose un défi important dans un domaine qui recherche de nouveaux cadres théoriques pour approfondir la relation complexe entre la parole et le geste. Pour relever ce défi, nous explorons des mécanismes plus simples et plus interprétables, tels que les convolutions de graphe (Kipf & Welling, 2016). Inspirées par leur application réussie dans le domaine de la synthèse de mouvement, sans l'implication de la parole (par exemple, marcher, danser, se battre), les convolutions de graphe sont prometteuses pour améliorer notre compréhension du comportement des réseaux d'apprentissage profond et pour permettre la création de représentations latentes des gestes plus intéressantes.

Inspirés par ces avancées, nous proposons une nouvelle architecture de réseau qui vise à remédier aux limites susmentionnées de la synthèse de gestes. L'architecture que nous proposons vise à atteindre trois objectifs clés :

- Générer des gestes convaincants ;
- Intégrer les convolutions de graphes pour obtenir une représentation latente plus explicite ;
- S'adapter aux applications en temps réel, comme les ECA, notamment en utilisation une architecture autorégressive ;

Dans les sections suivantes, nous décrivons notre architecture et les mécanismes utilisés, suivis d'une évaluation complète de notre modèle face au modèle StyleGestures. Enfin nous discutons des résultats obtenus et concluons en explorant les directions futures potentielles qu'ouvre notre modèle.

2 Méthodes

Nous proposons une nouvelle architecture appelée STARGATE (pour Spatio-Temporal Auto-Regressive Graph from Audio-Text Embeddings), illustré dans la figure 1. Nous suivons une structure encodeur-décodeur, avec une approche autorégressive par blocs. Cela se traduit par un réseau qui prend 3 modalités différentes en entrée :

— **Audio** : Une fenêtre de 1s de parole passée et de 1s de parole future ;
— **Texte** : Une fenêtre de 1s de texte passé et de 1s de texte futur ;
— **Gestes** : Un historique de 1s de gestes passés ;

L'existence d'une fenêtre contextuelle aussi longue est motivée par le fait que les gestes sont une modalité lente, avec une durée moyenne de 1 à 2 secondes selon que le geste se réfère à un seul mot ou à une phrase complète (Ferré, 2010). Chaque modalité dispose d'un encodeur dédié pour produire un espace latent particulier à celle-ci, qui est ensuite fusionné, via une simple concaténation, pour créer une représentation multimodale de la parole/des gestes. Cette représentation est finalement décodée en un bloc de nouvelles poses. Dans ce contexte, une pose définit les rotations de chaque point du corps à un instant donné, et ainsi une série temporelle de poses définit un mouvement. Nous avons choisi d'utiliser une sortie par bloc, c'est-à-dire un groupe de 16 poses. Contrairement à un réseau où chaque étape produit une pose, notre réseau produit bloc par bloc. Outre une implémentation plus efficace en parallélisant partiellement les calculs du bloc courant, cela laisse plus de liberté au réseau pour générer des gestes, l'historique autorégressif ne portant que sur le bloc courant.

2.1 Encodeurs de parole

La parole peut être séparée en deux composantes principales : le contenu acoustique et le contenu linguistique. Le signal acoustique produit pendant la parole contient de nombreuses informations telles que la prosodie (composée de l'énergie, de la hauteur, des rythmes) ou l'état émotionnel. Quant à lui, le contenu linguistique fait partie du signal acoustique, mais avec une représentation phonétique de ce qui a été prononcé, donnant une représentation plus explicite des informations sémantiques contenues dans le texte. Le texte est en effet une source d'information cruciale pour modéliser les gestes iconiques, déictiques et métaphoriques, qui sont tous directement liés au contenu sémantique, tandis que la dernière catégorie de gestes, les gestes dits rythmiques ou de battement, sont quant à eux plutôt liés au signal acoustique (McNeill, 1992). Les deux modalités (acoustique et textuelle) sont donc nécessaires pour générer des gestes dynamiques et significatifs. Dans notre architecture, nous utilisons ces deux modalités par le biais de deux encodeurs convolutionnels (CNN) similaires, mais distincts. Celui dédié à l'audio utilise les MFCCs comme entrées, tandis que celui du texte utilise des *embeddings* BERT (Devlin *et al.*, 2018).

2.2 Encodeur de gestes

Grâce à l'approche autorégressive, les gestes en tant que sorties peuvent être utilisés comme entrée pour la prédiction suivante. Ainsi la troisième modalité d'entrée est le mouvement qui permet de conserver une bonne cohérence pour les trajectoires des gestes, mais aussi de créer par la suite une représentation multimodale parole-gestes. Notre encodeur de mouvement est basé sur les travaux de (Zhou *et al.*, 2021). Le mouvement d'entrée est représenté sous forme d'*exponential map* (Grassia, 1998) pour 17 nœuds, ayant l'avantage d'être une représentation continue de la rotation par rapport aux angles d'Euler, et sont plus compactes que les quaternions (3 valeurs vs 4).

2.2.1 Réseau de neurones de graphe

Pour pouvoir à la fois générer des gestes convaincants et expliquer comment le réseau produit sa représentation latente des gestes, nous avons utilisé plusieurs mécanismes à l'intérieur de notre

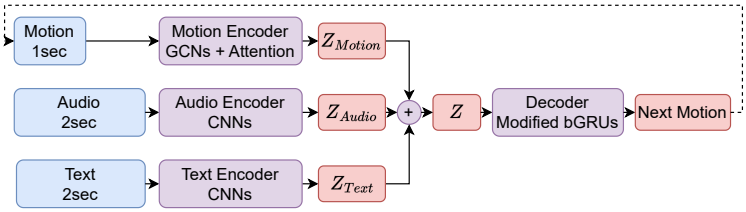


FIGURE 1 – Vue d'ensemble du réseau STARGATE avec sa structure encodeur-décodeur. Notre réseau utilise trois encodeurs distincts pour traiter les trois modalités séparément et un décodeur unique pour générer le mouvement à partir d'une représentation multimodale latente.

encodeur de geste. Le premier est l'utilisation de réseaux de convolution de graphe (GCN) (Kipf & Welling, 2016) au lieu des CNN classiques. Dans le cas d'un graphe, le calcul de la convolution est régi par une matrice d'adjacence, indiquant quels nœuds sont voisins et quel est le poids de chaque connexion. Dans notre cas, nous avons utilisé le bloc ST-GCN (pour Spatio-Temporal Graph Convolution Network) de (Zhou *et al.*, 2021), qui effectue à la fois une transformation spatiale à l'aide d'une convolution de graphe et un traitement temporel à l'aide d'un réseau de convolution temporel (TCN). Celui-ci utilise 3 matrices d'adjacence, l'une définissant des liens de boucle (l'information reste sur le même nœud), une définissant des liens de voisinages directs (lien hanche - colonne par exemple), la dernière définissant des liens de voisinages symétriques (lien main gauche - main droite par exemple). Motivés par le fait que nous voulons créer une nouvelle représentation des gestes, les matrices d'adjacence dans le réseau sont initialisées au début de l'entraînement, comme une connaissance initiale, mais sont modifiables et servent de paramètres pour que le réseau modifie les matrices pour la tâche de synthèse des gestes, en créant potentiellement de nouveaux liens intéressants entre les nœuds.

À notre connaissance, il s'agit du premier travail dans le domaine de la synthèse des gestes lié à la parole qui utilise des convolutions de graphe pour injecter une représentation plus explicite du mouvement.

2.2.2 Mécanisme d'attention

Dans l'implémentation de ST-GCN de (Zhou *et al.*, 2021) et la nôtre, il existe un mécanisme d'auto-attention sur les matrices d'adjacence avant la convolution de graphe. Les données d'entrée passent par un bloc d'auto-attention, afin de créer une "matrice d'attention", une pour chaque matrice d'adjacence. Ces matrices d'attention sont ajoutées aux matrices d'adjacence initiales pour produire ce que nous appelons des "matrices d'adjacence dynamiques". Ceci est motivé par le fait que même si nous laissons le réseau apporter de petites modifications aux matrices d'adjacence pendant l'apprentissage, au moment de l'inférence, elles resteront statiques. Ce mécanisme d'attention permet d'introduire des modifications dynamiques au moment de l'inférence, afin que le réseau accorde plus d'attention à certaines parties du corps pour chaque groupe d'images générées. Ainsi le réseau traite le geste de façon temporelle (TCN) et spatiale (GCN) avec un focus sur les parties du corps fourni par le mécanisme d'attention.

2.3 Décodeur de gestes

Les espaces latents audio, textuels et gestuels sont ensuite combinés pour produire un espace latent multimodal qui est transmis au décodeur de mouvement. Le décodeur, qui se compose de réseaux de neurones récurrents (RNNs) empilés (dans notre cas, les GRUs (Cho *et al.*, 2014)), produira le prochain bloc de pose, ces poses sont ensuite utilisées comme entrée de l'encodeur de mouvement. L'inconvénient majeur de l'autorégression est de ne travailler qu'avec des informations antérieures

FIGURE 2 – Un exemple d'ECA utilisant des mouvements générés par STARGATE. Cet exemple illustre sa capacité à générer différents types de gestes, tels que des gestes iconiques.

et de ne pas pouvoir analyser une séquence complète de gestes. Par conséquent, nous n'avons pas pu utiliser les GRU bidirectionnels pour obtenir une compréhension approfondie de l'ensemble des séquences de gestes. Cependant, motivés par les avantages que cela pourrait apporter et comme nous générons des séquences de poses, nous pouvons utiliser les GRUs bidirectionnelles sur ces séquences partielles. L'introduction de ce mécanisme de bidirectionnalité locale a pour but de permettre au réseau d'apprendre la relation entre les informations passées et futures présentes dans la représentation multimodale.

3 Entrainement

3.1 Corpus : BEAT

Nous avons entraîné tous nos modèles sur le corpus de données BEAT (Liu *et al.*, 2022). Il fournit à la fois des données de grande qualité et en grand volume. Dans notre cas, nous n'avons utilisé qu'un seul locuteur, car nous pensons que l'utilisation de plusieurs locuteurs sans fournir leur identité dans le modèle pourrait entraîner une confusion entre les styles de gestes. Nous disposons ainsi de 4 heures de données réparties entre les ensembles d'entraînement, de validation et de test avec un ratio 90/5/5. Nous n'avons pas utilisé les mouvements des doigts, car ils représentent une énorme quantité de données à traiter et à comprendre par le réseau. Le prétraitement des données audio et de mouvement suit le protocole et le code proposés par StyleGestures (Alexanderson *et al.*, 2020). Nous avons également augmenté les données en utilisant une stratégie miroir (symétrie axiale au niveau de la colonne vertébrale), permettant de doubler les données utilisables. Notre entraînement a duré environ 25 époques, soit 5h sur une machine équipée d'une carte NVIDIA RTX A5000.

3.2 Fonction objectif (*Loss*)

Notre modèle est entraîné pour minimiser deux termes : une fonction de Huber (Huber, 1992) (mixant une distance L1 et L2) sur les *exponential map* et une fonction de Huber sur les positions (dérivé des *exponential map*). Ceci est motivé par le fait que la seule minimisation de l'erreur sur les *exponential map* comme objectif du réseau donnera une importance égale à chaque articulation du squelette, cependant, comme le squelette est intrinsèquement une hiérarchie, nous voulons un contrôle le plus optimal possible des hanches et de la colonne vertébrale, car ils auront un impact sur tous les effecteurs finaux, ce qui nous est fourni en minimisant l'erreur sur les positions, qui sont calculées en traversant toute la hiérarchie, propageant les erreurs potentielles des *exponential map*. La fonction objectif est donc définie comme suit :

$$Loss = \mathcal{H}(r, \hat{r}) + \mathcal{H}(p, \hat{p})$$

avec p et r respectivement, les positions et les exponentielles de la référence, $\hat{p}$ et $\hat{r}$ respectivement les positions et les exponentielles de la génération et $\mathcal{H}$ la fonction de Huber.

	G ?	A ?	T ?	FGD ↓	Performance ↓ [Temps par image ↓] 5s	10s	30s	80s
StyleGestures	✗	✓	✗	14,15	7,76s [90ms]	12,90s [70ms]	31,05s [50ms]	80,07s [26ms]
STARGATE	✓	✓	✓	10,58	6,51s [37ms]	8,31s [17ms]	13,40s [8ms]	23,78s [5ms]
STARGATE Audio uniquement	✓	✓	✗	**8,61**	**3,49s [19ms]**	3,98s [8ms]	6,13s [3ms]	10,68s [2ms]

TABLE 1 – Résultats de la comparaison quantitative utilisant le FGD pour mesurer la qualité des gestes et des temps de traitements de chaque modèle (pour des générations de différentes durées). G, A, T représentent respectivement l'utilisation de graphe, audio et texte. Les valeurs en gras représentent le meilleur modèle. Calculé sur une machine équipé d'une carte NVIDIA RTX A6000 Laptop.

4 Évaluation

4.1 Métriques quantitatives

Dans cette partie, nous présentons l'évaluation de notre modèle ainsi qu'une variante sans texte (et son encodeur) nommée "Audio uniquement". Ces deux modèles sont comparés à StyleGestures.

Distance de gestes de Frechet (FGD). La meilleure tentative d'obtenir une métrique objective pour la synthèse de gestes est inspirée par la "Frechet Inception Distance" (FID) dans la synthèse d'images (Heusel *et al.*, 2017), adaptée dans (Yoon *et al.*, 2020) pour créer la FGD. Cette métrique est donc une distance de Frechet calculée sur un espace latent produit par un réseau d'inception. Nous avons réentraîné le réseau d'inception proposé sur nos données, car notre sortie différait significativement du réseau disponible. Ce réseau est un autoencodeur entraîné à transformer un mouvement d'entrée en une représentation latente compressée, puis recréer le mouvement initial. La métrique est donc basée sur une distance de Frechet calculée entre l'espace latent issu du mouvement de référence et l'espace latent issu du mouvement prédit. L'utilisation d'un réseau d'inception comme évaluateur permet d'obtenir une mesure plus proche de la perception humaine (Yoon *et al.*, 2020).

Comme nous pouvons le voir dans le tableau 1, les deux variantes de STARGATE sont plus performantes que StyleGestures, la variante "Audio uniquement" étant le meilleur modèle en ce qui concerne la FGD.

Performance. Bien que la performance n'est généralement pas une préoccupation majeure lors de la conception d'un modèle de synthèse de gestes, nous avons cherché à développer un réseau capable de fonctionner dans des scénarios en temps réel (par exemple pour les ECAs), en générant des gestes convaincants aussi rapidement que possible. Pour évaluer les performances dans ce contexte, nous avons effectué des tests qui prennent en compte les étapes de prétraitement, étant donné qu'elles peuvent avoir un impact significatif sur la charge de calcul (comme les calculs de BERT). Par conséquent, toutes les durées indiquées sont basées sur une entrée brute audio/texte, avec une taille de batch de 1. Nous indiquons la durée par image, les modèles ayant une sortie de longueur différente.

De même que pour la FGD, les deux variantes de STARGATE sont systématiquement plus rapides que StyleGestures. De plus, tous nos modèles sont capables de fonctionner en temps réel. Nous pouvons également observer que les performances de StyleGestures ne s'améliorent pas avec l'augmentation de la longueur de l'entrée, alors que les modèles STARGATE sont meilleurs pour le traitement de séquences plus longues, en prenant avantage du fait que les entrées audio et texte ne font pas parties de la boucle autorégressive, permettant le calcul en parallèle des 2 latent space sur toute la séquence.

Modèle	Naturel ↑ (Sans son)	Crédibilité ↑	Cohérence ↑
Référence	6,19 ± 0,28	5,27 ± 0,23	5,16 ± 0,23
Permuté	N/A	4,92 ± 0,20	4,77 ± 0,22
StyleGestures	5,97 ± 0,25	4,87 ± 0,22	4,70 ± 0,23
STARGATE	5,89 ± 0,28	5,0 ± 0,20	4,85 ± 0,22

TABLE 2 – Résultats de notre évaluation MOS, nous indiquons la moyenne et un intervalle de confiance à 95% pour chaque aspect.

4.2 Évaluation subjective

Pour mieux évaluer notre modèle, nous avons procédé à une évaluation subjective via un score d'opinion moyen (MOS) afin d'évaluer la qualité globale des gestes générés par notre nouvelle architecture. Nous avons adapté le protocole d'évaluation du GENEA Challenge 2020 (Kucherenko *et al.*, 2021) sur les énoncés pour avoir une compréhension plus claire des aspects évalués pour chacune des questions.

L'évaluation a été divisée en deux parties. La première partie consistait à visionner des vidéos sans audio et à répondre à la question suivante : *"How human-like does the gesture motion appear ?"* La seconde partie consistait à regarder des vidéos avec le son et à répondre à deux questions : *"How credible are the gestures with respect to the speech ?"* et *"How consistent are the gestures with respect to the speech ?"*, ces 2 aspects jugent les aspects sémantiques, le premier sa réalisation (si un geste est bien réalisé), le second sa cohérence (un geste peut ne pas correspondre à ce qui est dit). Les participants devaient évaluer chaque question sur une échelle de 1 à 7. Nous avons évalué quatre systèmes de génération de gestes dans cette étude : Référence (vérité terrain), Permuté (mouvement synthétique avec un audio différent), StyleGestures et STARGATE. Nous avons présenté 30 vidéos pour chaque système, chacune durant 9 secondes et nous avions un total de 25 participants issus de la plateforme Prolific (12 femmes et 13 hommes) parlant un anglais natif. Chaque vidéo est générée en utilisant le modèle 3D du GENEA Challenge 2020 (Kucherenko *et al.*, 2021) sur lequel les poses prédites (rotations) ont été transférées. Les résultats sont présentés dans le tableau 2.

Comme on peut le voir, StyleGestures a obtenu un score légèrement supérieur à celui de notre modèle STARGATE en ce qui concerne l'aspect "naturel" des gestes, mais notre modèle est légèrement meilleur en termes de cohérence et de crédibilité lorsque l'audio est disponible.

4.3 Discussions

Les résultats quantitatifs présentés dans le tableau 1 démontrent que notre modèle surpasse l'état de l'art en termes de FGD. Cela correspond à notre évaluation MOS, où les scores de cohérence et de crédibilité sont légèrement supérieurs à ceux du modèle StyleGestures. Il est intéressant de noter que la variante "Audio uniquement" de notre système présente des valeurs FGD inférieures à celles de notre modèle de base. Ce comportement peut être attribué au fait que notre modèle est capable de générer des gestes non rythmiques convaincants (tels que ceux décrits dans la figure 2), bien qu'en petit nombre. Ces gestes sont nettement plus difficiles à maîtriser et s'écartent souvent de manière significative des gestes de référence, ce qui contribue à des scores FGD plus élevés. Ce n'est pas le cas du modèle StyleGestures où la production de gestes iconiques est absente. Cela peut suggérer que la structure de graphe améliore la compréhension des mouvements et permet au réseau d'établir des liens plus forts entre le texte et le mouvement.

Les performances de notre modèle démontrent des capacités temps réel, l'incorporation du texte dans

le modèle nécessite cependant la génération de séquence supérieure à 10s pour être temps réel. Dans les deux cas, cela implique que l'intégration dans des ECAs peut être effectuée, permettant ainsi une interaction naturelle avec des interfaces homme-machine (les avatars pouvant produire des gestes en temps réel).

Dans le tableau 2, nous avons observé que le modèle StyleGestures recevait des notes plus élevées pour l'aspect "naturel" des gestes lorsqu'il était évalué sans audio. Nous attribuons cette différence à la présence de gestes non rythmiques dans notre modèle, qui ne sont pas toujours produits clairement (comme le montre la figure 2, où le geste "walking in" est "avorté"). Cette incohérence se traduit parfois par un mélange de gestes iconiques et de gestes rythmiques, ce qui donne l'impression d'un manque de naturel lorsque les gestes ne sont pas accompagnés de la parole. Le tableau 2 corrobore également les résultats de recherches antérieures (Kucherenko *et al.*, 2021), où le modèle Permuté présente des évaluations plus élevées que StyleGestures et notre modèle. Nous attribuons cette observation à la forte prévalence des gestes de battement dans l'ensemble de données et dans les gestes générés. Les gestes de battement sont intrinsèquement cohérents et crédibles lorsqu'ils s'alignent sur le rythme de parole. Cela est vrai pour les mouvements avec ou sans permutations de l'audio, car ils proviennent tous deux du même modèle, qui est capable de s'aligner sur le rythme global de l'ensemble de données. Par conséquent, les gestes dans les deux scénarios ont pu convaincre les utilisateurs. En revanche, le mouvement de référence présente des gestes plus cohérents et plus crédibles en raison de la présence de gestes hautement sémantiques. Les résultats de cette évaluation perceptive sont relativement proches. Ce point est certainement lié à la difficulté de la tâche par des évaluateurs novices. Les évaluateurs sont peu sensibles à l'apparition de gestes non synchronisés ou, à la présence ou l'absence de gestes iconiques lors de la parole. Nous comptons évaluer notre modèle grâce à des experts en linguistiques et gestualité, notamment à travers des annotations précises des gestes prédits, attestant la qualité de la synthèse de façon plus objective.

5 Conclusion

Nous avons développé STARGATE, une nouvelle architecture autorégressive par blocs qui utilise trois modalités d'entrée pour construire une représentation latente unifiée de la parole et des gestes, et synthétise des gestes liés à la parole. À notre connaissance, cette architecture est la première à utiliser des convolutions de graphe au lieu de convolutions traditionnelles, en incorporant explicitement une connaissance de la structure du squelette humain. Nous avons évalué cette architecture à l'aide de mesures quantitatives et d'études subjectives, démontrant sa capacité à générer des gestes convaincants, non seulement des gestes de battement, mais aussi des gestes plus complexes tels que des gestes iconiques et métaphoriques. Ces gestes exigent une compréhension approfondie du contenu linguistique, de la parole et de la coordination des mouvements du corps. En outre, nos tests de performance indiquent que notre architecture peut générer des gestes en temps réel, même lorsqu'un calcul d'*embeddings* BERT est incorporé.

Dans nos travaux futurs, nous visons à améliorer l'explicabilité et l'interprétabilité de nos résultats. Nous avons l'intention d'analyser en profondeur le comportement du réseau pour comprendre comment la génération de gestes complexes est déclenchée au sein du réseau et comment la structure du graphe favorise les connexions entre les gestes et la parole. Les matrices d'adjacence dynamiques peuvent nous aider à interpréter visuellement le comportement du réseau. Une telle analyse pourrait fournir une perspective nouvelle et unique sur la corrélation entre la parole et la production de gestes.

Références

ALEXANDERSON S., HENTER G. E., KUCHERENKO T. & BESKOW J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, p. 487–496 : Wiley Online Library.

ALEXANDERSON S., NAGY R., BESKOW J. & HENTER G. E. (2023). Listen, denoise, action ! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, **42**(4), 1–20.

AO T., GAO Q., LOU Y., CHEN B. & LIU L. (2022). Rhythmic gesticulator : Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, **41**(6), 1–19.

BOUTET D. (2008). Une morphologie de la gestualité : structuration articulaire. *Cahiers de linguistique analogique*, (5), 81–115. HAL : hal-00607593.

CASSELL J. (2001). Embodied conversational agents : representation and intelligence in user interfaces. *AI magazine*, **22**(4), 67–67.

CHO K., MERRIENBOER B., GULCEHRE C., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

DE RUITER J. P., BANGERTER A. & DINGS P. (2012). The interplay between gesture and speech in the production of referring expressions : Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, **4**(2), 232–248. DOI : 10.1111/j.1756-8765.2012.01183.x.

DEICHLER A., MEHTA S., ALEXANDERSON S. & BESKOW J. (2023). Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, p. 755–762.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

FERRÉ G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous french. In *Language Resources and Evaluation, Workshop on Multimodal Corpora*, volume 6, p. 86–91.

GRASSIA F. S. (1998). Practical parameterization of rotations using the exponential map. *Journal of graphics tools*, **3**(3), 29–48.

HENTER G. E., ALEXANDERSON S. & BESKOW J. (2020). Moglow : Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, **39**(6), 1–14.

HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B. & HOCHREITER S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, **30**.

HUBER P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics : Methodology and distribution*, p. 492–518. Springer.

KENDON A. (2004). *Gesture : Visible Action as Utterance*. Cambridge ; New York : Cambridge University Press.

KIPF T. N. & WELLING M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv :1609.02907*.

KRAUSS R. M. & HADAR U. (1999). The role of speech-related arm/hand gestures in word retrieval. In *Gesture, Speech, and Sign*, p. 93–116. Oxford University Press. DOI : 10.1093/acprof :oso/9780198524519.003.0006.

KUCHERENKO T., HASEGAWA D., HENTER G. E., KANEKO N. & KJELLSTRÖM H. (2019). Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, p. 97–104.

KUCHERENKO T., JONELL P., YOON Y., WOLFERT P. & HENTER G. E. (2021). A large, crowdsourced evaluation of gesture generation systems on common data : The genea challenge 2020. In *26th international conference on intelligent user interfaces*, p. 11–21.

LI J., KANG D., PEI W., ZHE X., ZHANG Y., HE Z. & BAO L. (2021). Audio2gestures : Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 11293–11302.

LIU H., ZHU Z., IWAMOTO N., PENG Y., LI Z., ZHOU Y., BOZKURT E. & ZHENG B. (2022). Beat : A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, p. 612–630 : Springer.

LU S., YOON Y. & FENG A. (2023). Co-speech gesture synthesis using discrete gesture token learning. *arXiv preprint arXiv :2303.12822*.

MCNEILL D. (1992). *Hand and Mind : What Gestures Reveal about Thought*. Chicago and London : The University of Chicago Press.

SO W. C., KITA S. & GOLDIN-MEADOW S. (2009). Using the hands to identify who does what to whom : Gesture and speech go hand-in-hand. *Cognitive Science*, **33**(1), 115–125. DOI : 10.1111/j.1551-6709.2008.01006.x.

TAKEUCHI K., KUBOTA S., SUZUKI K., HASEGAWA D. & SAKUTA H. (2017). Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*, p. 198–202 : Springer.

YOON Y., CHA B., LEE J.-H., JANG M., LEE J., KIM J. & LEE G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, **39**(6), 1–16.

ZHANG F., JI N., GAO F. & LI Y. (2023). Diffmotion : Speech-driven gesture synthesis using denoising diffusion model. In *International Conference on Multimedia Modeling*, p. 231–242 : Springer.

ZHAO W., HU L. & ZHANG S. (2023). Diffugesture : Generating human gesture from two-person dialogue with diffusion models. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, p. 179–185.

ZHOU K., CHENG Z., SHUM H. P., LI F. W. & LIANG X. (2021). Stgae : Spatial-temporal graph auto-encoder for hand motion denoising. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, p. 41–49 : IEEE.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Un paradigme pour l'interprétation des métriques et pour mesurer la gravité des erreurs de reconnaissance automatique de la parole

Thibault Bañeras-Roux[1] Michael Rouvier[2] Jane Wottawa[3] Richard Dufour[1]
(1) Laboratoire des Sciences du Numérique de Nantes (LS2N), France
(2) Laboratoire Informatique d'Avignon (LIA), France
(3) Laboratoire d'Informatique de l'Université du Mans (LIUM), France
thibault.roux@univ-nantes.fr, jane.wottawa@univ-lemans.fr,
michael.rouvier@univ-avignon.fr, richard.dufour@univ-nantes.fr

RÉSUMÉ

Les mesures couramment employées pour l'évaluation des transcriptions automatiques de la parole, telles que le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER), ont fait l'objet d'importantes critiques en raison de leur corrélation limitée avec la perception humaine et de leur incapacité à prendre en compte les nuances linguistiques et sémantiques. Bien que des métriques fondées sur les plongements sémantiques aient été introduites pour se rapprocher de la perception humaine, leur interprétabilité reste difficile par rapport au WER et CER. Dans cet article, nous surmontons ce problème en introduisant un paradigme qui intègre une métrique choisie pour obtenir un équivalent du taux d'erreur appelé Distance d'Édition Minimale, ou Minimum Edit Distance (minED). Nous proposons également d'utiliser cette approche pour mesurer la gravité des erreurs en fonction d'une métrique, d'un point de vue intrinsèque et extrinsèque.

ABSTRACT

A Paradigm for Interpreting Metrics and Measuring Error Severity in Automatic Speech Recognition

The commonly employed metrics for the evaluation of automatic speech transcriptions, such as Word Error Rate (WER) and Character Error Rate (CER), have faced significant criticism due to their limited correlation with human perception and their inability to account for linguistic and semantic nuances. While metric-based embeddings have been introduced to approximate human perception, their interpretability remains challenging compared to WER and CER. In this article, we overcome this problem by introducing a paradigm that integrates a chosen metric to obtain an equivalent of the error rate called Minimum Edit Distance (minED). We also propose to use this approach to measure the severity of errors according to a metric, from an intrinsic and extrinsic perspective.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Métriques d'évaluation, Interprétabilité, erreurs de transcriptions.

KEYWORDS: Automatic speech recognition, Evaluation metric, Interpretability, Transcription errors..

1 Introduction

Malgré les progrès considérables réalisés dans le domaine de l'apprentissage automatique et l'utilisation intensive de données pour l'entraînement des modèles, les systèmes de Reconnaissance Automatique de la Parole (RAP) présentent encore des erreurs de transcription dans des proportions variables en fonction de leurs conditions d'utilisation.

L'évaluation d'un système de RAP consiste le plus souvent à comparer les transcriptions manuelles (référence) et automatiques (hypothèse) à l'aide d'une mesure choisie, généralement le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER). Tous deux consistent à calculer une distance de Levenshtein entre la référence et l'hypothèse. Cependant, ces métriques sont critiquées pour attribuer le même poids à toutes les erreurs tout en négligeant les nuances linguistiques et sémantiques (Favre *et al.*, 2013; Ruiz & Federico, 2015; Kafle & Huenerfauth, 2017; Gordeeva *et al.*, 2021).

Pour remédier à ces limitations, des mesures fondées sur les plongements (Zhang *et al.*, 2020; Kim *et al.*, 2021; Bañeras-Roux *et al.*, 2022) ont été proposées pour intégrer les aspects sémantiques.

De même, d'un point de vue perceptif, Kafle & Huenerfauth; Kim *et al.*; Gordeeva *et al.*; Bañeras-Roux *et al.* ont utilisé des ensembles de données annotées pour évaluer rigoureusement l'alignement des métriques de reconnaissance vocale avec la perception humaine, révélant la corrélation supérieure des métriques sémantiques avec le jugement humain.

Si les mesures sémantiques offrent une perspective d'évaluation différente, leurs scores, calculés par similarité cosinus, manquent d'interprétabilité, contrairement au WER qui s'appuie simplement sur les mots. Dans cet article, nous proposons d'intégrer une métrique dans un nouveau paradigme, appelé Distance d'Édition Minimale, ou en anglais Minimum Edit Distance (minED), afin de rendre interprétables les scores des métriques basés sur les plongements. Ce paradigme est également appliqué pour mesurer la gravité des erreurs, ce qui peut être utilisé pour l'analyse des métriques.

Le document est organisé comme suit. La section 2 présente les métriques de RAP et un ensemble de données avec des annotations de perception humaine. La section 3 décrit le paradigme minED proposé pour l'interprétabilité des mesures, tandis que la section 4 examine la capacité du paradigme à mesurer la gravité des erreurs. Enfin, nous concluons le travail et donnons des perspectives dans la section 5.

2 Méthodologie

Dans la section 2.1, nous fournissons des détails sur les métriques de RAP utilisées dans cette étude. Ensuite, dans la section 2.2, nous présentons le jeu de données HATS, utilisé pour évaluer les métriques ainsi que le paradigme proposé.

2.1 Métriques

Comme indiqué précédemment, la communauté a développé diverses métriques s'appuyant sur les plongements. En utilisant BERT (Devlin *et al.*, 2019), nous pouvons extraire des représentations sémantiques des phrases. L'une de ces mesures, **SemDist** (Kim *et al.*, 2021) calcule la similarité cosinus entre la référence et l'hypothèse à l'aide des plongements obtenus au niveau de la phrase.

Une autre mesure, **BERTScore** (Zhang *et al.*, 2020), appliquée dans diverses tâches de traitement automatique du langage (TAL) (Yilmaz *et al.*, 2019; Hanna & Bojar, 2021), calcule un score de similarité pour chaque token de la phrase candidate avec chaque token de la phrase de référence à l'aide de plongements contextuels.

Dans cette étude, SemDist intègre la version Sentence-BERT (Reimers & Gurevych, 2019) de CamemBERT (Martin *et al.*, 2020) [1], une version française de BERT, et BERTScore utilise un BERT (Devlin *et al.*, 2019) multilingue. Pour nos expériences, nous avons normalisé toutes les mesures sur une échelle de [0, 1] en appliquant la règle "le plus faible le meilleur".

2.2 Jeu de données HATS

L'un des moyens d'évaluer correctement les métriques consiste à utiliser un ensemble de données d'annotations humaines. L'ensemble de données HATS est un corpus en libre accès, pour le français, conçu pour évaluer la corrélation entre les mesures d'évaluation de RAP et la perception humaine du point de vue du lecteur. L'ensemble de données HATS a été développé à l'aide d'une expérience côte-à-côte (Gordeeva *et al.*, 2021; Kafle & Huenerfauth, 2017; Kim *et al.*, 2022). Une référence textuelle, ainsi que deux hypothèses erronées produites par des systèmes de RAP (8 systèmes de bout-en-bout (Ravanelli *et al.*, 2021) et deux systèmes basés sur une architecture DNN-HMM [2] (Povey *et al.*, 2011)), ont été présentées à au moins 7 sujets qui ont sélectionné la meilleure hypothèse. L'ensemble des données comprend 1 000 triplets : une référence, chacune accompagnée de deux hypothèses et du nombre de votes associé.

En calculant le nombre de fois qu'une métrique est en accord avec les annotations humaines (la métrique indique le meilleur score pour l'hypothèse choisie par les humains), nous pouvons calculer un ratio correspondant à la corrélation entre cette métrique et l'évaluation humaine.

La métrique SemDist obtient la corrélation la plus forte avec la perception humaine sur HATS. L'ensemble des données sera utilisé pour déterminer si l'utilisation du paradigme minED entraîne une réduction de la corrélation avec la perception humaine par rapport à l'utilisation de la métrique seule.

3 Intégrer des métriques pour l'interprétabilité

Le paradigme minED est conçu pour améliorer l'interprétabilité des mesures produisant des scores difficiles à comprendre. Pour ce faire, nous intégrons une métrique non interprétable telle que SemDist dans minED. Cela consiste à calculer le nombre minimum de modifications à appliquer à l'hypothèse pour qu'elle soit suffisamment proche de la référence en ce qui concerne sa perception humaine. Suivant cette idée, nous appliquons cette méthode aux mots (minWED) et aux caractères (minCED).

Le paradigme est décrit dans la section 3.1, tandis que la section 3.2 traite du paramétrage de la méthode. Nous discutons ensuite de deux types de mesures (cohérentes, incohérentes) influençant le coût de calcul (section 3.3), et explorons la corrélation entre minED et la perception humaine (section 3.4).

1. https://huggingface.co/dangvantuan/sentence-camembert-large
2. https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/

FIGURE 1 – Graphe de chaque modification possible pour obtenir une hypothèse sans erreur avec le paradigme minWED. Chaque arête correspond à une erreur corrigée. Étant donné la référence, nous avons trois erreurs de mots, chacune d'un type différent : 1 substitution, 1 insertion, 1 suppression. La métrique est basée sur la règle "le plus faible le meilleur". Le symbole ϵ correspond à une suppression.

3.1 Distance d'Édition Minimum (minED)

La correction de mots, ou de caractères, consiste à éditer l'hypothèse afin qu'il n'y ait plus de substitutions, d'insertions ou de suppressions. Le paradigme minED calcule le nombre minimum de corrections (mots ou caractères) nécessaires pour rendre une hypothèse "acceptable" sur la base d'une métrique non interprétable. Pour ce faire, nous générons un graphique qui représente toutes les modifications qui peuvent être apportées à l'hypothèse pour qu'elle devienne la référence (voir figure 1). Pour chaque élément corrigé, nous calculons un score entre la référence et la nouvelle hypothèse à l'aide de la métrique incorporée. Si le score est inférieur à un seuil prédéfini, l'hypothèse est jugée "acceptable". Il n'est donc pas nécessaire de calculer le reste du graphique. Le score minED est le nombre de niveau minimum à calculer pour obtenir un score en dessous du seuil.

La définition du seuil est cruciale, et la section 3.2 détaille la manière de l'établir.

La Figure 1 présente le graphe des possibilités pour la référence « *I will book them an appointment* » et l'hypothèse « *will book them an appointment and* ». Dans ce scénario, nous avons trois erreurs : une suppression, une substitution et une insertion. L'erreur de suppression est représentée par un symbole ϵ.

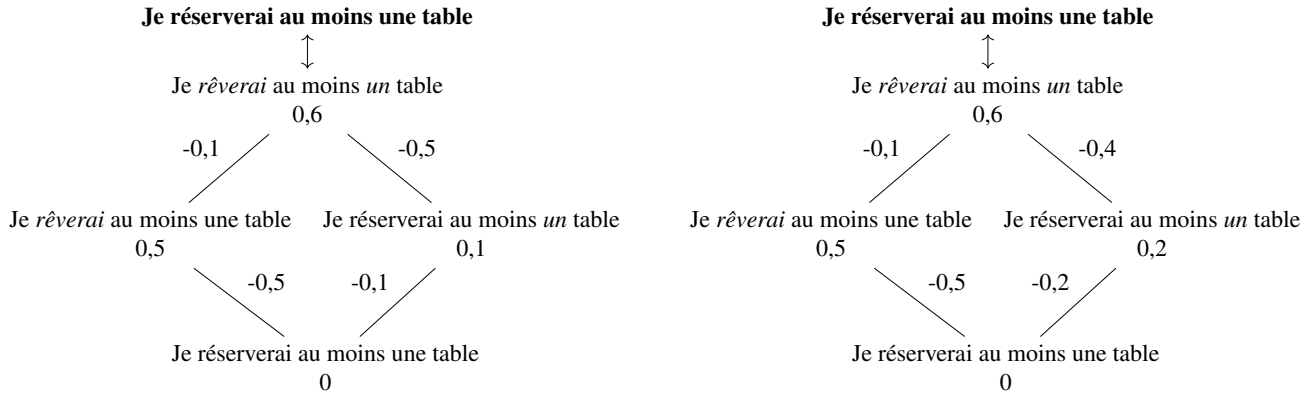
3.2 Fixation du seuil d'acceptabilité

Comme indiqué dans la section 3.1, minED signifie les modifications nécessaires pour une hypothèse acceptable. Ce concept repose sur le fait qu'une métrique puisse donner un score considérée comme acceptable par les humains. Par exemple, lorsqu'un humain lit une hypothèse erronée, si une métrique sémantique indique un score inférieur au seuil (dans un contexte où la valeur la plus basse est la meilleure), le sens de la phrase originale est censé être compris.

Lorsque le seuil est trop bas, les mesures minWED et minCED tendent à se rapprocher des valeurs WER ou CER. Inversement, des valeurs de seuil trop élevées font converger ces mesures vers des

scores nuls, ce qui signifie qu'aucune correction n'est nécessaire.

Une approche pourrait consister à sélectionner un seuil qui maximise la corrélation avec la perception humaine, mais il n'est pas exclu de réfléchir à d'autres méthodes.



(a) Métrique cohérente.

(b) Métrique incohérente.

FIGURE 2 – Comparaison de l'impact des corrections sur des métriques cohérentes et incohérentes. Les métriques respectent une règle de "le plus faible le meilleur".

3.3 Cohérence des métriques

Lors de la correction d'une hypothèse pour la rapprocher de la référence, nous pouvons observer une amélioration du score selon la métrique incorporée.

La correction peut avoir deux effets connus : soit elle améliore le score indépendamment des modifications précédentes (voir figure 2a), soit elle améliore le score en fonction des modifications précédentes (voir figure 2b). Par exemple, dans la figure 2a, la correction de la substitution *réserverai*/*rêverai* améliorera la performance de la métrique de *0,5*, que *une*/*une* ait été corrigé ou non. Dans la figure 2b, la correction de *réserverai*/*rêverai* améliorera la performance de la métrique de *0,5* ou *0,4*, selon que *une*/*un* ait été corrigé ou non.

La propriété de cohérence permet de calculer plus rapidement le nombre minimum de modifications, car il n'est plus nécessaire de calculer l'ensemble du graphique. Une approche pratique consiste plutôt à calculer le deuxième niveau où une seule erreur dans l'hypothèse est corrigée. Ensuite, on soustrait le score de l'hypothèse initiale du nombre minimum d'améliorations rédactionnelles requises pour que le score obtenu soit inférieur au seuil.

WER et CER sont des exemples de métriques cohérentes, tandis que BERTScore et SemDist sont des exemples de métriques incohérentes.

3.4 Corrélation avec la perception humaine

La figure 3 montre la corrélation entre la perception humaine et minED pour différentes valeurs de seuil (θ). Les seuils les plus bas donnent des corrélations plus proches de la métrique associée à l'édition (WER ou CER), tandis que les valeurs trop élevées entraînent une baisse des performances.

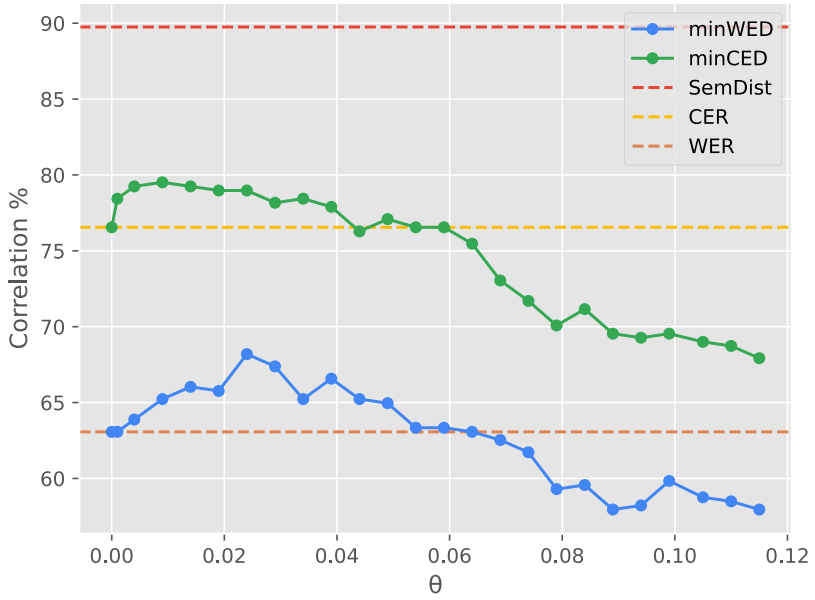


FIGURE 3 – Corrélation de MinED avec le jeu de données HATS en fonction de différentes valeurs de seuil (θ).

Alors que minWED gagne 5,12 % par rapport au WER et améliore l'interprétabilité par rapport à SemDist, il perd 21,56 % de corrélation par rapport à SemDist. De même, minCED est mieux corrélé avec la perception humaine que CER, mais présente une proportion significative de perte par rapport à SemDist. Ces résultats montrent les limites de l'utilisation des taux d'erreur pour évaluer les transcriptions de RAP d'un point de vue humain.

4 Mesurer la gravité des erreurs

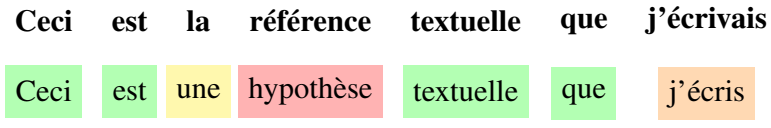


FIGURE 4 – Visualisation de la gravité des erreurs selon notre paradigme MinED intégrant une métrique sémantique.

Dans cette section, nous étudions la capacité de notre paradigme à identifier les erreurs et à mesurer leur gravité, comme l'illustre la figure 4. La section 4.1 présente notre méthode d'évaluation de la gravité des erreurs, tandis que la section 4.2 se penche sur les résultats et l'analyse.

4.1 Protocole d'évaluation

Pour évaluer correctement la capacité de notre paradigme à identifier les erreurs critiques, nous partons du principe que la correction d'une erreur grave devrait avoir un impact plus important sur une tâche en aval que la correction d'une erreur mineure.

Dans notre étude, nous avons choisi une tâche de traduction du français vers l'anglais à partir de données vocales. Cette tâche comprend d'abord une transcription automatique, considérée comme l'évaluation intrinsèque des sorties de RAP à l'aide des métriques SemDist et CER. La transcription résultante est ensuite transmise à un traducteur automatique pour générer l'hypothèse finale, qui est considérée comme l'évaluation extrinsèque du système de reconnaissance de la parole à l'aide des métriques BLEU et BERTScore.

	Transcriptions	**Traductions**	**SemDist**	**BERTScore**
Référence	à nos résultats	to our results		
Hypothèses	un non résultat	a no result	57,8	28,1
Hypothèses Corrigées	à non résultat	to no result	50,1 (+7,7)	23,6 (+4,5)
	un nos résultat	a our result	20,2 (+37,6)	21,4 (+6,7)
	un non résultats	a no results	52,7 (+5,1)	28,0 (+0,1)

TABLE 1 – Exemple des améliorations de SemDist et de BERTScore obtenues par la correction de l'hypothèse « *à nos résultats* ». Les scores sont projetés dans une règle "le plus faible le meilleur" et une échelle de [0, 100] pour une meilleure lisibilité.

Comme le montre le tableau 1, nous générons autant de corrections à une hypothèse erronée qu'il y a d'erreurs de transcription. Cette approche nous permet d'obtenir, pour chaque correction, le score d'amélioration pour nos mesures intrinsèques et extrinsèques. Si nous observons une corrélation entre ces deux valeurs, cela signifie que le paradigme est effectivement capable de mesurer la sévérité des erreurs.

Notre dispositif expérimental utilise le jeu de données HATS pour obtenir les références et les hypothèses erronées associées, les traductions étant générées à l'aide de Google Traduction.

4.2 Résultats et analyse

Le tableau 2 présente la corrélation de Spearman entre l'amélioration intrinsèque et extrinsèque de la transcription automatique pour la tâche de traduction. Une corrélation notable entre SemDist et BERTScore est observée, démontrant la capacité du paradigme à mesurer la gravité des erreurs. Des corrélations différentes apparaissent pour les mesures intrinsèques et extrinsèques, suggérant des variations potentielles dans les résultats pour des tâches autres que la traduction.

Intrins./Extrins.	**BERTScore**	**BLEU**
SemDist	0,39	0,26
CER	0,22	0,23

TABLE 2 – Moyenne de la corrélation de Spearman entre les améliorations intrinsèques et extrinsèques en fonction de différentes métriques pour la tâche de traduction.

5 Conclusions et perspectives

Nous avons proposé un paradigme qui non seulement rend les mesures de RAP interprétables, mais qui permet également de mesurer la gravité des erreurs. L'approche minED fournit un cadre plus transparent pour l'évaluation des systèmes de RAP. Alors que notre étude a révélé une diminution notable de la corrélation avec la perception humaine lors de l'intégration d'une métrique dans minWED (*i.e.* sur les mots), nos résultats démontrent que minCED (sur les caractères) maintient une performance relativement forte dans la capture de la perception de l'erreur comparé à d'autres métriques évalué par Bañeras-Roux *et al.*.

L'étude montre également, par la perte significative de corrélation avec l'interprétabilité, qu'une mesure du nombre d'erreurs ne correspond pas à la manière dont les humains se comportent. Il semble que les humains donnent la priorité à la prise en compte de la gravité des erreurs plutôt qu'à la simple proportion d'erreurs graves.

Une autre stratégie pour développer des mesures interprétables plus étroitement liées à la perception humaine consisterait à développer des métriques qualitatives plutôt que quantitatives. Par exemple, des ensembles de données comme HypRatings (Kim *et al.*, 2022) intègrent des annotations qualitatives telles que "`exact`", "`hyp utile`", "`hyp fausse`", et "`hyp incohérente`". L'étude du développement de métriques prédisant ces caractéristiques qualitatives pourrait constituer une perspective intéressante pour de futures recherches.

6 Considérations éthiques

Lors de la mise en production d'un système de reconnaissance de la parole, si minED est utilisé pour évaluer le système, il convient d'être prudent dans le choix du seuil : l'acceptabilité des erreurs est subjective et peut ne pas s'appliquer à tous les humains.

Tous les modèles et données utilisés dans cet article sont publics et librement accessibles à des fins de reproductibilité. Notre code est disponible sur un dépôt GitHub public [3].

Références

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2022). Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In *Interspeech 2022*.

BAÑERAS-ROUX T., WOTTAWA J., ROUVIER M., MERLIN T. & DUFOUR R. (2023). Hats : An open data set integrating human perception applied to the evaluation of automatic speech recognition metrics. In *Text, Speech and Dialogue 2023 - Interspeech Satellite*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

3. https://anonymous.4open.science/r/mined

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of ASR systems : Does WER really predict performance ? In *INTERSPEECH*, p. 3463–3467.

GORDEEVA L., ERSHOV V., GULYAEV O. & KURALENOK I. (2021). Meaning Error Rate : ASR domain-specific metric framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, p. 458–466.

HANNA M. & BOJAR O. (2021). A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, p. 507–517.

KAFLE S. & HUENERFAUTH M. (2017). Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 165–174.

KIM S., ARORA A., LE D., YEH C.-F., FUEGEN C., KALINLI O. & SELTZER M. L. (2021). Semantic Distance : A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In *Proc. Interspeech 2021*, p. 1977–1981. DOI : 10.21437/Interspeech.2021-1929.

KIM S., LE D., ZHENG W., SINGH T., ARORA A., ZHAI X., FUEGEN C., KALINLI O. & SELTZER M. (2022). Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In *Proc. Interspeech 2022*, p. 3978–3982. DOI : 10.21437/Interspeech.2022-11144.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, volume CONF : IEEE Signal Processing Society.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992.

RUIZ N. & FEDERICO M. (2015). Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, p. 296–302 : IEEE.

YILMAZ Z. A., WANG S., YANG W., ZHANG H. & LIN J. (2019). Applying bert to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) : System Demonstrations*, p. 19–24.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Un système d'annotation automatique de la structure prosodique

Philippe Martin
LLF, UFRL, Université Paris Cité
Place Paul Ricoeur, 75013 Paris, France
philippe.martin@utoronto.ca

RESUME

On présente un système d'annotation prosodique permettant de visualiser les structures prosodiques générées par des règles de dépendance appliquées sur des événements prosodiques annotés automatiquement.

Les événements prosodiques sont définis par des cibles tonales dans la notation ToBI, ou par des contours mélodiques, montants ou descendants, atteignant la hauteur la plus basse ou la plus haute de la phrase, et au-dessus ou au-dessous du seuil de glissando (c'est-à-dire perçus comme un changement mélodique ou un ton statique), mais d'autres définitions peuvent être utilisées au gré de l'utilisateur.

À partir de ces définitions, les contours ou les cibles tonales alignés sur les voyelles des syllabes accentuées localisées sont affichés automatiquement. Des règles de dépendance définies par l'utilisateur opèrent sur ces évènements prosodiques pour déterminer et afficher la structure prosodique correspondante, permettant une comparaison visuelle avec la structure morphosyntaxique et conduisant à une meilleure compréhension de la manière dont les structures prosodiques peuvent amorcer le décodage syntaxique par l'auditeur.

Abstract

An automatic prosodic annotation system

A prosodic annotation system is presented that enables the prosodic structure generated by dependency rules applied to automatically annotated prosodic events to be visualized.

Prosodic events are defined by tonal targets in ToBI notation, or by melodic contours, rising or falling, reaching the lowest or highest pitch in the sentence, and above or below the glissando threshold (i.e. perceived as a melodic change or static tone), but other definitions can be used at the user's discretion.

Based on these definitions, vowel-aligned contours or tonal targets of localized stressed syllables are displayed automatically. User-defined dependency rules operate on these prosodic events to determine and display the corresponding prosodic structure, enabling visual comparison with the morphosyntactic structure and leading to a better understanding of how prosodic structures can initiate syntactic bootstrapping by the listener.

MOTS-CLES : Structure prosodique, cible tonale, contour mélodique, annotation prosodique.

KEYWORDS: Prosodic structure, tonal target, melodic contour, dependency rules, prosodic annotation.

1 Introduction

Il est généralement admis que la structure prosodique d'une phrase résulte de la fusion des groupes accentuels AP (*accent phrases*) en syntagmes intonatifs intermédiaires ip (*intermediate intonation*

phrases), des ip en syntagmes intonatifs IP (*Intonation Phrases*) et, finalement, des IP en structure prosodique PS (*Prosodic Structure*). À ce jour, plusieurs systèmes d'annotation prosodique automatique ont été développés, basés sur le modèle autosegmental-métrique et l'annotation ToBI (Wightman and Ostendorf 1994, Syrdal et al. 2001, Rosenberg 2010). Des implémentations récentes utilisent des algorithmes d'apprentissage profond (Zhai and Hasegawa-Johnson 2023), dont l'efficacité repose sur la fiabilité de l'annotation ToBI d'un corpus d'apprentissage.

On décrit ici un système d'annotation automatique basé sur un modèle sensiblement différent, impliquant l'autonomie de la structure prosodique par rapport à la morphosyntaxe, et autorisant la transcription des évènements prosodiques par contours aussi bien que par cibles tonales. Pour décrire les évènements prosodiques, par hypothèse localisés essentiellement sur les voyelles des syllabes accentuées, ce système intègre la valeur de glissando pour mieux lier la perception de l'auditeur à la réalité acoustique. Ce choix contraste avec les systèmes dominants dans lesquels la transcription des événements prosodiques est plus étroitement liée à la courbe de la fréquence fondamentale et pour lesquels, pour les langues à accent lexical comme l'italien, seuls les tons de frontière et non les accents de hauteur jouent un rôle dans l'indication de la structure prosodique (Selkirk 1984).

Sur la base de l'autonomie de la structure prosodique, dans les unités desquelles viendraient s'insérer les unités syntaxiques, on est conduit à abandonner, voire à inverser, la vision selon laquelle la structure prosodique résulterait d'une mise en correspondance avec la structure morphosyntaxique de la phrase. Dès lors, on est conduit à admettre que 1) il doit exister des marqueurs prosodiques qui indiquent les fusions successives des AP en groupes intonatifs plus grands indépendamment des autres structures morphosyntaxiques ou informationnelles, et 2) il doit exister des règles qui régissent la distribution de ces marqueurs prosodiques pour indiquer sans ambiguïté une structure prosodique donnée. Un choix théorique porte alors sur 1) la sélection de caractéristiques dans le matériel prosodique qui instancient de manière adéquate les marqueurs prosodiques, et 2) le type de grammaire prosodique approprié.

2 Sélection des paramètres acoustiques

Un système d'annotation prosodique automatique doit s'appuyer sur des caractéristiques acoustiques et donc sur la fiabilité de l'algorithme d'analyse de la parole pour obtenir l'intensité syllabique, la durée des voyelles et, en particulier, la hauteur mélodique. Excepté pour les enregistrements effectués dans des salles insonorisées ou dans des conditions équivalentes, l'analyse de la parole lue et spontanée nécessite des algorithmes fiables de suivi mélodique, qui en pratique peuvent se révéler déficients pour beaucoup d'enregistrements, en particulier spontanés.

Ces systèmes peuvent être complétés par l'affichage superposé, sur la même échelle de fréquence, d'un spectrogramme à bande étroite. L'utilisateur peut alors être guidé par la première ou la deuxième harmonique affichée sur le spectrogramme pour saisir manuellement à l'écran un segment de courbe de hauteur corrigée, grâce à une fonction de dessin intégrée au logiciel venant au secours d'une mesure de F0 défaillante. On dispose ainsi d'un outil permettant l'annotation prosodique (et aussi des formants) quelle que soit la qualité de l'enregistrement, même ceux très dégradés du début du 19ème siècle.

Le choix des caractéristiques acoustiques devrait aussi refléter la perception humaine mieux que la courbe de fréquence fondamentale souvent prise telle quelle comme source de données. Au lieu de décrire les caractéristiques prosodiques des voyelles par les changements de durée, d'intensité et de fréquence fondamentale évalués séparément, on adopte le paramètre de glissando, un paramètre combinant ces trois entités acoustiques en tenant en compte avec une approximation acceptable la

perception d'un événement prosodique en termes de changement mélodique, perçu comme tel ou perçu comme un ton statique, selon sa valeur par rapport à un seuil (Rossi 1971).

La valeur du glissando est obtenue en rapportant la variation de hauteur en demi-tons à leur durée :

Glissando = $(DT_2- DT_1) / (t_2-t_1)$ avec DT (Demi-ton) = 12 * (log($F0_t$/100.0)) / log (2.0) ($F0_t$ étant la valeur de la fréquence fondamentale à l'instant t).

Le seuil de glissando en demi-tons est donné par coeff / $(t2-t1)^2$, coeff étant un paramètre ajustable variant de 0,16 à 0,32 de manière à prendre en compte le degré éventuel de non-linéarité de la courbe de fréquence fondamentale de la voyelle ainsi que la sensibilité des auditeurs aux changements de hauteur.

Le choix des contours mélodiques à l'endroit des voyelles accentuées comme marqueurs prosodiques est renforcé par des recherches neuro-perceptives récentes, qui suggèrent que les accents de hauteur sont encodés comme des catégories discrètes et contrastives dans le cerveau (Llanos et al. 2021). Le positionnement des syllabes accentuées est semi-automatique pour le français, langue à accent rythmique, et ne peut être totalement déterminée à partir du texte, car dépendant entre autres du débit de parole (Martin 2018). Par défaut, ce sont les mots grammaticaux, dont la catégorie est établie à partir d'un lexique (Lexique 3), suivi d'un correcteur trigramme, qui reçoivent l'accent. Pour l'italien par contre, langue à accent lexical, la position des syllabes accentuées relève directement d'un lexique, mais peut être obtenue également à partir d'une analyse morphologique des mots (Martin 1990).

Pour l'annotation prosodique de la phrase française, les événements prosodiques situés sur les voyelles des syllabes accentuées sont classés automatiquement par le logiciel à partir de la mesure de trois paramètres : a) F0 montant – F0 descendant ; b) en dessous ou au-dessus du seuil de glissando et c) atteignant la valeur de F0 la plus basse ou la plus élevée parmi toutes les autres voyelles accentuées de la phrase.

Les classes de contours sont alors :

a. **Cdec↓,** (L*L%) Contour terminal déclaratif atteignant la valeur mélodique (F0) la plus basse dans la phrase parmi les autres fins de voyelles accentuées

b. **Cint↑,** (H*H%) Contour terminal interrogatif atteignant la valeur mélodique (F0) la plus haute dans la phrase parmi les autres fins de voyelles accentuées

c. **Cris↗,** (LH*) Variation mélodique montante supérieure au seuil de glissando (cf. continuation majeure en français)

d. **Cfap#↘,** (H*L#) Variation mélodique descendante supérieure au seuil de glissando suivie d'une pause d'au moins 250 ms (contour de la dictée)

e. **Cfal↘,** (HL*) Variation mélodique descendante supérieure au seuil de glissando (cf. continuation mineure en français)

f. **Cneu–,** (H*) Montée ou descente mélodique inférieure au seuil de glissando (contour neutralisé)

Pour l'italien, il existe un contour complexe g) **Ccom √**, (LL*H), descendant en dessous du seuil de glissando sur la voyelle accentuée, et généralement montant au-dessus du seuil de glissando sur la voyelle de la syllabe finale (et éventuellement sur la consonne voisée finale). Ce contour mélodique complexe est réalisé sur une même voyelle d'une syllabe finale et accentuée.

En plus de ces 7 catégories adaptées au français et à l'italien, le logiciel permet de définir un total de 14 catégories distinctes s'adaptant à la langue et aux conditions de l'analyse (par exemple en ajoutant un contour d'insistance, d'hésitation, etc.).

À partir de ces définitions et de l'analyse acoustique (durées vocalique, courbes de F0 et d'intensité), les contours ou les cibles tonales ToBI sont automatiquement affichés avec leurs caractéristiques graphiques sélectionnées par l'utilisateur (couleur, largeur du trait...), comme le montre la figure 1. Cette classification est effectuée en fonction des limites des voyelles accentuées sur l'échelle temporelle, définies soit manuellement par l'opérateur, soit à l'aide d'un moteur de segmentation automatique intégré au logiciel (disponible dans plus de 42 langues...).

FIGURE 1. Définitions programmables des évènements prosodiques en contours (à gauche) ou en notation ToBI (à droite).

3 Choix d'une grammaire

En utilisant le système d'annotation ToBI, une grammaire prosodique dans le cadre autosegmental-métrique inclut un ensemble de règles décrivant toutes les séquences bien formées d'événements prosodiques dans une phrase. En pratique, ces règles sont découvertes et énoncées à partir de l'analyse de diverses configurations syntaxiques qui sous-tendent la primauté de la morphosyntaxe sur la prosodie de la phrase (Delais, Post et Yoo 2020).

Cependant, une autre vision peut être envisagée dans laquelle la structure prosodique est considérée comme autonome par rapport à la morphosyntaxe et à toute autre structure de la phrase. On peut alors considérer que la structure prosodique, c'est-à-dire les regroupements successifs d'AP en ip, d'ip en IP et d'IP en SP, sont indiqués par des marqueurs prosodiques spécifiques perceptivement saillants tels que décrits plus haut.

Afin d'intégrer les marqueurs prosodiques dans une grammaire, le choix s'est porté sur une grammaire de dépendance, basée sur une réinterprétation du concept de continuité, présent par exemple chez (Delattre 1966). En effet, en tant que marqueur prosodique, la continuité mineure indique à l'auditeur que la phrase se poursuivra jusqu'à l'apparition d'une continuité majeure, et la continuité majeure indique que la phrase se poursuivra jusqu'à l'apparition d'un contour terminal.

En termes de relations de dépendance sur l'échelle temporelle, l'occurrence de la continuité mineure dépend de l'avènement attendu de la continuité majeure plus loin dans la phrase. De même, la continuité majeure dépend de l'occurrence attendue d'un contour terminal dans le futur de la phrase. En termes de regroupements des groupes accentuels, les séquences de groupes accentuels dont le dernier se termine par une continuité mineure sont fusionnées avec tous les groupes accentuels déjà regroupés avec une continuité majeure, et tous les groupes accentuels se terminant par une continuité majeure sont fusionnées avec tous les groupes accentuels déjà regroupés se terminant par un contour terminal.

Une grammaire de dépendance prosodique consiste en un ensemble de règles de dépendance qui relient la dépendance des évènements prosodiques (contours ou cibles tonales) à d'autres évènements prosodiques situés plus tôt (dépendance "à gauche") ou plus tard (dépendance "à droite") dans la phrase.

4 Grammaire de dépendance prosodique

Les règles de la grammaire de dépendance pour le français, langue à accent rythmique, portant sur les voyelles accentuées finales des groupes accentuels (en termes de contours) sont proposée ci-dessous pour des phrases déclaratives. {A, B} ⇒ {X, Y} s'interprète l'occurrence de A ou de B dépendent de l'occurrence de X ou de Y apparaissant plus tard dans la phrase, et {A, B} ⇦ {X, Y} s'interprète " l'occurrence de X ou de Y dépendent de l'occurrence de A ou de B apparus plus tôt dans la phrase.

Cneu─ ⇒ {**Cfal↘**, **Cris↗**, **Cfap#↘**, **Cdec↓**, **Cint↑**} (dépendance « à droite »)

Cfal↘ ⇒ {**Cris↗**, **Cfap#↘**, **Cint↑**} (dépendance « à droite »)

{**Cris↗**, **Cfap#↘**,} ⇒ **Cdec↓** (dépendance « à droite »)

Cdec↓ ⇦ **Cneu─** (thème dans une configuration rhème-thème, dépendance « à gauche »)

Pour l'italien, langue à accent lexical, comportant des voyelles accentuées non finales dans les groupes accentuels :

Cneu─ ⇒ {**Cfal↘**, **Cris↗**, **Cfap#↘**, **Cdec↓**, **Cint↑**} (dépendance « à droite »)

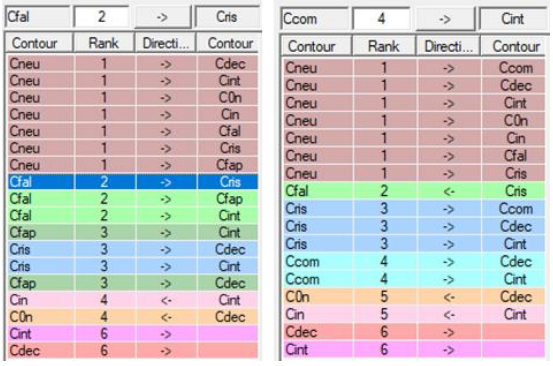
Cris↗ ⇦ **Cfal↘** (dépendance « à gauche »)

Cris↗ ⇒ **Ccom √** (dépendance « à droite »)

Ccom √ ⇒ **Cdec↓** (dépendance « à droite »)

Cdec↓ ⇦ **Cneu─** (thème dans une configuration rhème-thème)

Une fonction de calcul de la structure prosodique vérifie la cohérence de la distribution des contours relativement à ces règles par rapport à leurs réalisations acoustiques et met automatiquement à jour leur catégorie si nécessaire. Par exemple, aucun contour **Cfal↘** ne peut précéder immédiatement un contour déclaratif terminal en français (sauf s'il est suivi d'une pause, auquel cas il est étiqueté comme un contour de dictée **Cfap#↘**).



Cfal	2	->	Cris
Contour	Rank	Directi...	Contour
Cneu	1	->	Cdec
Cneu	1	->	Cint
Cneu	1	->	C0n
Cneu	1	->	Cin
Cneu	1	->	Cfal
Cneu	1	->	Cris
Cneu	1	->	Cfap
Cfal	2	->	Cris
Cfal	2	->	Cfap
Cfal	2	->	Cint
Cfap	3	->	Cint
Cris	3	->	Cdec
Cris	3	->	Cint
Cfap	3	->	Cdec
Cin	4	<-	Cint
C0n	4	<-	Cdec
Cint	6	->	
Cdec	6	->	

Ccom	4	->	Cint
Contour	Rank	Directi...	Contour
Cneu	1	->	Ccom
Cneu	1	->	Cdec
Cneu	1	->	Cint
Cneu	1	->	C0n
Cneu	1	->	Cin
Cneu	1	->	Cfal
Cneu	1	->	Cris
Cfal	2	<-	Cris
Cris	3	->	Ccom
Cris	3	->	Cdec
Cris	3	->	Cint
Ccom	4	->	Cdec
Ccom	4	->	Cint
C0n	5	<-	Cdec
Cin	5	<-	Cint
Cdec	6	->	
Cint	6	->	

FIGURE 2 Règles de dépendance définies par l'utilisateur (français, à gauche - italien, à droite).

De même, une phrase en italien ne peut pas commencer par un contour **Cfal↘**, au-dessus du seuil de glissando (mais elle peut commencer par un **Cneu─** descendant neutralisé, en dessous du seuil de glissando). Si aucune règle de dépendance ne s'applique à un contour, celui-ci n'est pas intégré dans

la représentation graphique de la structure prosodique, il reste isolé, ce qui se traduit par une structure prosodique incomplète. Un exemple est donné figure 5.

La structure résultante est automatiquement affichée, avec une représentation arborescente utilisant des branches orthogonales orientées selon la direction des dépendances (Figure 3 et 4).

Chaque règle de dépendance est modifiable par l'utilisateur avec le rang de l'événement prosodique et la direction gauche ou droite de la dépendance. Par exemple, en français, **Cfal↘** dépend de **Cris↗** à droite, alors qu'en italien, **Cfal↘** dépend de **Cris↗** à gauche. Le code couleur correspond aux catégories d'événements prosodiques définies par l'utilisateur dans le tableau de la figure 2.

5 Exemples en français et en italien

Le français et l'italien permettent une comparaison entre une langue à accent rythmique et une langue à accent lexical, en particulier en ce qui concerne l'apparition de contours mélodiques "complexes" placés sur la voyelle accentuée du groupe accentuel, mais en dessous du seuil de glissando, et remontant au-dessus de ce seuil sur la voyelle de la syllabe finale (et éventuellement sur la consonne voisée finale). Ce contour mélodique, qui n'est pas normalement présent en français, est automatiquement identifié par le système, étiqueté et colorié selon les propriétés graphiques sélectionnées par l'utilisateur. L'analyse acoustique dispose de plusieurs algorithmes de détection de F0 (peigne spectral, AMDF, autocorrélation), complété par une estimation linéaire aux moindres carrés pour les sections des voyelles accentuées.

5.1 Exemple en italien

Un premier exemple de parole lue en italien (corpus SIWIS 2016) montre une structure prosodique reflétant partiellement la syntaxe (les voyelles accentuées sont en majuscules grasses, figure 3)

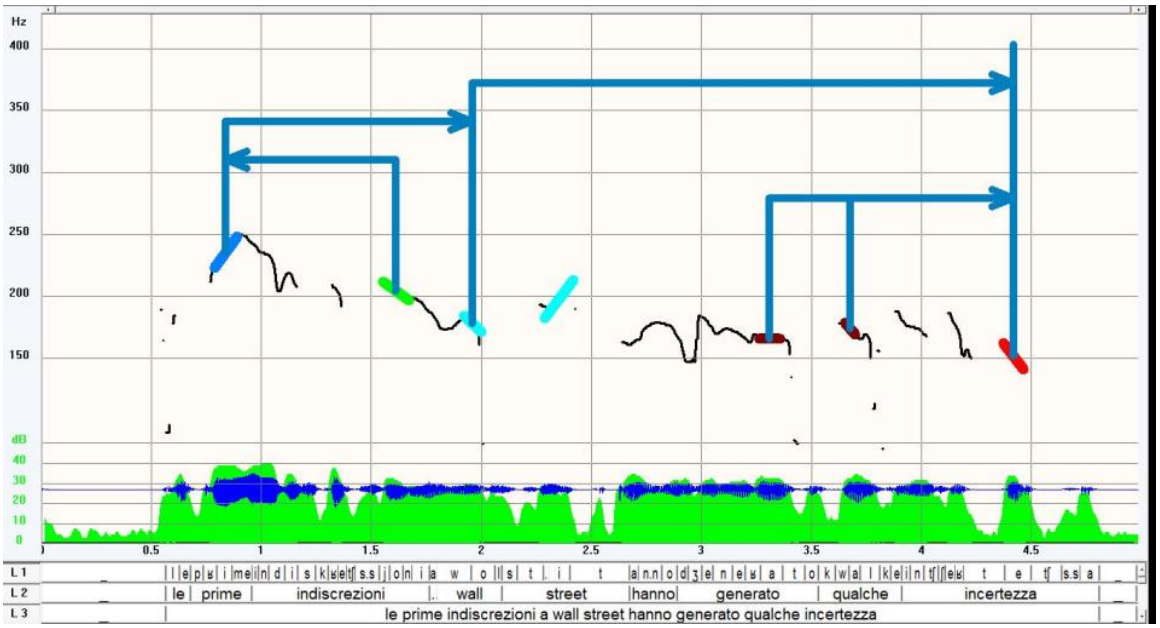


FIGURE 3. Contours colorés des voyelles de syllabes accentuées successives : **Cris↗** au-dessus du seuil de glissando (GT), **Cfal↘** au-dessus de GT, **Ccom √** complexe non final, **Cneu—** en dessous de GT, **Cneu—** en dessous de GT, et **Cdec↓** terminal. (**Cris↗** est représenté en bleu, **Cfal↘** en vert, **Ccom √** en turquoise, **Cneu—** en brun et **Cdec↓** en rouge). [*Le pr**I**me*] **Cris↗** [*indiscrezi**O**ni*] **Cfal↘** [*a W**A**ll Str**EE**t*] **Ccom √** [*hanno gener**A**to*] **Cneu—** [*qu**A**lche*] **Cneu—** [*incert**E**zza*] **Cdec↓** (SIWIS it_a1_08_123).

{[(*Le prIme*) (*indiscreziOni*)] [(*a WAll StrEEt*)]} {(*hanno generAto*) (*quAlche*) (*incertEzza*)}
"Les premières rumeurs à Wall Street ont généré une certaine incertitude" regroupant les deux premiers groupes accentuels avec la séquence **Cris↗** - **Cfal↘** avec une dépendance "à gauche", suivie de la fusion du groupe avec le premier AP présentant un contour complexe **Ccom √**.
Les sections surlignées de la courbe mélodique de cet exemple montrent les variations mélodiques successives sur les voyelles accentuées, annotées par des variations linéaires classées automatiquement par le logiciel selon les catégories définies par l'utilisateur et décrites ci-dessus.

5.2 Exemple en français

La figure 4 affiche la séquence de contours prosodiques et la structure prosodique correspondante d'un segment de parole spontanée "*et du coup tu sais j'aurai au moins j'aurai de peau mieux ça sera quoi* " (corpus ORFEO 2017).

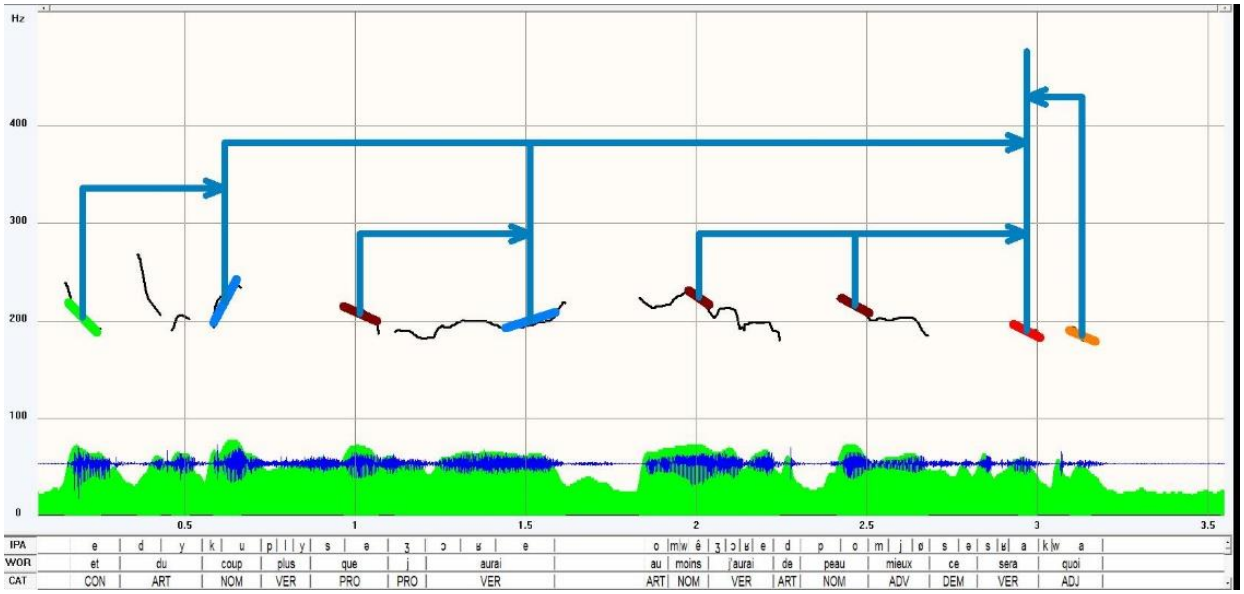


FIGURE 4. Représentation de la structure prosodique de l'exemple (***Et***) **Cfal↘** (*du c**OU**p*) **Cris↗** [*plus qu**E***] **Cneu─** (*j'aur**AI***) **Cris↗** (*au mou**IN**s*) **Cneu─** (*j'aurai de p**AEU***) **Cneu─** (*mieux ça ser**A***) **Cdec↓** (*qu**OI***) **Cneu←** (ORFEO 07madmc110912).

Le syntagme accentuel [*tu s**AI**s*] est aligné sur une parenthèse prosodique terminée par un contour conclusif terminal **Cdec↓** et constitue une structure prosodique indépendante à un seul groupe accentuel. Le dernier groupe accentuel (*qu**OI***) porte un contour neutralisé **Cneu─** indiquant une dépendance « à gauche » vers le contour terminal (cf. rhème dans une configuration rhème-thème).

Le calculateur de structure prosodique fournit une représentation graphique prenant en compte a) les événements prosodiques et b) les règles de dépendance opérant sur les événements prosodiques, évènements et règles programmables et définies par l'utilisateur. Le système est interactif, et constitue un outil efficace permettant d'expérimenter l'adéquation de la description des événements prosodiques et la validité de la grammaire prosodique sur de nombreux exemples facilement analysables, quelle que soit la qualité des enregistrements. Il est également possible d'écouter l'effet d'un changement de contour effectué par une commande graphique de l'utilisateur grâce au morphing prosodique, qui utilise un algorithme Psola intégré.

On peut par exemple vérifier sur la figure 5 l'absence de grammaticalité en français de **Cfal↘** chute du contour mélodique au-dessus du seuil de glissando suivi de **Cdec↓**, ou l'absence de grammaticalité du même contour **Cfal↘** en position initiale dans une structure prosodique en italien. La figure 5 affiche une séquence de contours invalide pour le groupe accentuel [*pour cette min**U**te*]

annotée avec un **Cfal↘** non grammatical au lieu du **Cneu─** correct. Le groupe accentuel n'est donc pas intégré à la structure prosodique de la phrase.

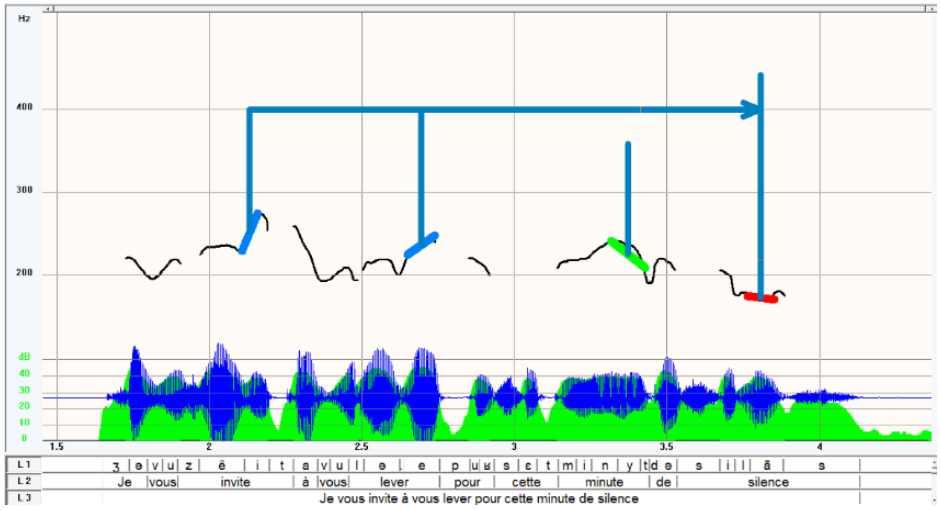


FIGURE 5. *[*Je vous inv**I**te*] **Cris↗** [*à vous lev**ER***] **Cris↗** [*pour cette min**U**te*] **Cfal↘** [*de sil**EN**ce*] **Cdec↓** (SIWIS fr_b_29_001) avec [*pour cette min**U**te*] annoté **Cfal↘** au lieu de **Cneu─**, menant à la non intégration dans la structure prosodique.

Les caractéristiques de chaque contour annoté, la gamme de fréquence fondamentale, la durée et l'intensité sont affichées dans un tableau et peuvent être sauvegardées au format Excel ou équivalent d'un simple clic de souris (figure 6).

Name	Width	T 1 [s]	T 2 [s]	F0 1 [Hz]	F0 2 [Hz]	Duration [s]	Range [Hz]	F0 1 [ST]	F0 2 [ST]	Int 1 [dB]	Int 2 [dB]	Diff Int [dB]	Glissando [st/s]	Glissando ratio	Text
Cris	12	0.781	0.904	219	248	0.123	29	13	15	36	36	1	17//10	1.700	
Cfal	12	1.554	1.685	216	198	0.131	-18	13	11	27	27	2	13//9	1.444	
Ccom	12	1.929	1.983	183	175	0.054	-8	10	9	30	30	-2	14//54	0.259	
Ccom	12	1.983	2.308	175	186	0.324	11	9	10	28	28	1	4//1	4.000	
Ccom	12	2.308	2.385	186	209	0.077	23	10	12	29	29	-3	28//26	1.076	
Cneu	12	3.247	3.363	166	165	0.116	-1	8	8	31	31	1	1//11	0.090	
Cneu	12	3.657	3.715	175	165	0.058	-10	9	8	32	32	2	19//47	0.404	
Cdec	12	4.400	4.450	158	144	0.050	-14	7	6	34	34	-5	36//63	0.571	

Name	Width	T 1 [s]	T 2 [s]	F0 1 [Hz]	F0 2 [Hz]	Duration [s]	Range [Hz]	F0 1 [ST]	F0 2 [ST]	Int 1 [dB]	Int 2 [dB]	Diff Int [dB]	Glissando [st/s]	Glissando ratio	Text
Cfal	12	0.157	0.254	206	188	0.096	-18	12	10	27	27	6	21//17	1.235	e
Cris	12	0.593	0.656	219	253	0.063	34	13	16	34	34	6	43//39	1.102	u
Euh	12	0.980	1.059	214	208	0.078	-6	13	12	35	35	-2	7//25	0.280	ø
Cris	12	1.485	1.612	192	214	0.127	22	11	13	32	32	-15	29//9	3.222	ɛ
Cneu	12	1.970	2.027	232	221	0.057	-11	14	13	37	37	0	13//48	0.270	ẽ
Cneu	12	2.436	2.499	221	208	0.063	-13	13	12	37	37	-5	20//39	0.512	o
Cdec	12	2.923	2.983	201	188	0.060	-13	12	10	27	27	-1	19//43	0.441	a
C0n	12	3.104	3.141	193	190	0.036	-3	11	11	22	22	4	9//121	0.074	a

FIGURE 6. Liste des contours annotés au format Excel pour les exemples des figures 3 (en haut) et 4 (en bas), avec leurs valeurs de F0, de durée, d'intensité et de glissando de F0 pour les voyelles accentuées.

6 Conclusion

Un outil d'annotation prosodique efficace est essentiel pour mieux comprendre le rôle de la structure prosodique en relation avec la morphosyntaxe et l'accès au sens par l'auditeur. Il devient également crucial lorsqu'un grand nombre d'exemples annotés est nécessaire pour entraîner des systèmes d'apprentissage profond. L'outil est intégré dans le logiciel d'analyse de la parole WinPitch (2024) et peut être téléchargé gratuitement sur www.winpitch.com.

Références

DELAIS-ROUSSARIE E, POST B., YOU H-Y. (2020) Unités prosodiques et grammaire intonative du français : vers une nouvelle approche, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole*, Nancy, France, 08-19 juin 2020.

DELATTRE P. (1966) Les dix intonations de base du français, *French Review* (40) 1-14.

LLANOS F., SNEED GERMAN J., NIKE GNANATEJA G., CHANDRASEKARAN B. (2021) The neural processing of pitch accents in continuous speech, *Neuropsychologia*, 2021, 158.

LEXIQUE 3 http://www.lexique.org/shiny/lexique

MARTIN PH. (1990) Positionnement automatique de l'accent lexical de l'Italien, *Actes des XVIIIèmes Journées d'Études sur la parole*, Montréal 1990, 149-152.

MARTIN PH. (2018) *Intonation, structure prosodique et ondes cérébrales*, London : ISTE, 322 p.

ORFEO (2017) Outils et Recherches sur le Français Écrit et Oral. http://www.projet-orfeo.fr/

ROSENBERG R. (2010) AuToBI - A tool for automatic ToBI annotation, *Proc. Interspeech September 2010*, doi: 10.21437/.2010-71.

SELKIRK E. O. (1984) *Phonology and syntax: the relation between sound and structure*. Cambridge, Mass., MIT Press.

ROSSI M. (1971) Le seuil de glissando ou seuil de perception des variations tonales pour la parole, *Phonetica* (23) 1-33.

SYRDAL A, HIRSCHBERG J, MCGORY J. and BECKMAN M. (2001) Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1–2), 135–151.

SIWIS Corpus (2016) Yamagishi, J. et al. The SIWIS French Speech Synthesis Database, https://doi.org/10.7488/ds/1705.

WIGHTMAN C. and OSTENDORF M. (1994) Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4), pp. 469–481.

WINPITCH (2024) www.winpitch.com

ZHAI W., HASEGAWA-JOHNSON M. (2023) Wav2ToBI: a new approach to automatic ToBI transcription. *Proc. INTERSPEECH 2023*, 2748-2752, doi: 10.21437/Interspeech.2023-477.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une comparaison de l'intonation ironique en français et en mandarin

Ziqi Zhou[1, 2] Jalal Al-Tamimi[1] Hiyon Yoo[1]
(1) Laboratoire de Linguistique Formelle, UMR 7110, CNRS, Université Paris Cité, 5 Rue Thomas Mann, 75013, Paris, France
(2) CLILLAC-Arp, Université Paris Cité, 5 Rue Thomas Mann, 75013, Paris, France
ziqi.zhou@etu.u-paris.fr, jalal.al-tamimi@u-paris.fr, hi-yon.yoo@u-paris.fr

RESUME

L'utilisation de corrélats acoustiques dans la production de l'ironie a été bien documentée. Cependant, dans quelle mesure les résultats sont comparables dans différentes langues reste une question inexplorée. Cette étude vise à réaliser une comparaison des caractéristiques de l'intonation ironique entre le français et le mandarin, en utilisant un protocole expérimental unifié. Une expérience de production a été menée pour susciter l'énoncé ironique. Les résultats ont d'abord été analysés par forêts aléatoires pour explorer le poids relatif de huit corrélats acoustiques comme marqueur de l'ironie. Ensuite, des modèles linéaires à effets mixtes (LMM) ont été utilisés pour explorer davantage les principaux corrélats acoustiques. Nos résultats ont confirmé que la caractéristique de l'intonation ironique est spécifique à chaque langue, révélant des schémas différents de corrélats acoustiques utilisés pour produire l'ironie en français et en mandarin. De plus, un effet de genre sur l'énoncé ironique en français a été identifié.

ABSTRACT

A comparison of ironic tone of voice in French and Mandarin
The use of acoustic correlates in the production of ironic speech has been well-documented. However, the extent to which we can compare the results in different languages remains questionable. This study aims to conduct a comparison of ironic tone of voice in French and Mandarin, using a unified experimental protocol. A production experiment was conducted to elicit ironic speech. The results were first subjected to a Random Forest analysis in which we explored the relative weight of eight acoustic correlates for ironic speech. Secondly, a series of linear mixed-effects models (LMM) were built to further explore the major acoustic correlates. Our results revealed that the ironic tone of voice is language-specific, indicating distinct patterns of acoustic correlates used in French and Mandarin for ironic speech. Additionally, a gender effect on French ironic speech was identified.

MOTS-CLES : production, ironie, corrélats acoustiques, français, mandarin.
KEYWORDS : production, irony, acoustic correlates, French, Mandarin.

Actes des 35èmes Journées d'Études sur la Parole, pages 209–218.

1 Introduction

L'étude des caractéristiques de l'intonation ironique est sujet de controverse, puisque certains auteurs refusent d'associer un patron tonal à cette entité (ex. Bryant & Fox Tree, 2005), alors que d'autres résultats expérimentaux semblent confirmer la présence d'un patron tonal spécifique à l'ironie (cf. Anolli et al., 2002; Cheang & Pell, 2008; González-Fuente et al., 2016; Jansen & Chen, 2020; Li et al., 2020; Li & Gu, 2021; Lœvenbruck et al., 2013; Rockwell, 2000; Scharrer & Christmann, 2011).

Ainsi, plusieurs études ces dernières années ont cherché à déterminer s'il existe un patron tonal spécifique à l'ironie dans différentes langues, telles que l'anglais (Cheang & Pell, 2008; Chen & Boves, 2018; Rockwell, 2000); le français (González-Fuente et al., 2016; Lœvenbruck et al., 2013); le mandarin (Li et al., 2020; Li & Gu, 2021) ; l'italien (Anolli et al., 2002) ; le néerlandais (Jansen & Chen, 2020) ; l'allemand (Scharrer & Christmann, 2011), etc. Les résultats semblent non seulement confirmer l'existence de l'intonation ironique, mais aussi à la fois des similitudes ainsi que des caractéristiques spécifiques à chaque langue dans la réalisation de l'intonation ironique. Un résultat récurent quelle que soit la langue étudiée est que l'on observe généralement un débit de parole réduit pour produire l'ironie.

Cependant, d'autres corrélats acoustiques essentiels à l'encodage de l'ironie, tels que la hauteur de la voix, l'intensité et la qualité de la voix, présentent des variations spécifiques à la langue. Par exemple, Cheang & Pell (2008); Chen & Boves (2018) and Rockwell (2000) ont montré que le ton de la voix ironique est caractérisé par une hauteur réduite et une étendue de la hauteur plus étroite par rapport au discours sans ironie, ce qui va dans le sens des résultats obtenus en mandarin (Li & Gu, 2021) et en allemand (Scharrer & Christmann, 2011). En revanche, les études menées sur le français (González-Fuente et al., 2016; Laval & Bert-Erboul, 2018; Lœvenbruck et al., 2013) ou sur l'italien (Anolli et al., 2002) ont donné des résultats opposés : l'ironie en français et en italien est caractérisée par une hauteur plus élevée.

En ce qui concerne l'intensité, l'ironie en mandarin (Li & Gu, 2021), en français (González-Fuente et al., 2016; Lœvenbruck et al., 2013) et en néerlandais (Jansen & Chen, 2020) est marqué par une intensité plus faible, tandis qu'en anglais (Cheang & Pell, 2008; Chen & Boves, 2018; Rockwell, 2000) et en italien (Anolli et al., 2002), une intensité accrue est utilisée pour marquer les énoncés ironiques. Pour évaluer la qualité de la voix, le rapport harmoniques/bruit (HNR) est couramment utilisé. Les résultats expérimentaux indiquent également un moindre bruit (c'est-à-dire des valeurs de HNR plus élevées) dans l'ironie en néerlandais (Jansen & Chen, 2020) et en mandarin (Li et al., 2020), mais davantage de bruit en anglais (Cheang & Pell, 2008).

Les variations observées dans les études sur différentes langues suggèrent des caractéristiques prosodiques distinctives associées à l'ironie. Cependant, il existe plusieurs problèmes méthodologiques pour évaluer l'impact de différences entre les langues.

Tout d'abord, les variations dans les méthodologies utilisées dans les différentes études pourraient entraver des comparaisons fiables de l'intonation ironique entre les langues. Par exemple, Jansen & Chen (2020) ont exploré l'ironie en néerlandais en utilisant trois types de phrases : déclaratives, questions interrogatives et exclamatives, tandis que Lœvenbruck et al. (2013) se sont concentrés uniquement sur les phrases déclaratives lors de leur étude sur l'ironie en français. De même, Li & Gu (2021), dans leur examen du mandarin, ont spécifiquement étudié les phrases exclamatives comme base pour les énoncés ironiques. Or les types de phrases peuvent influencer la tonalité

ironique (Chen & Boves, 2018), et les différences dans les structures des phrases utilisées par divers chercheurs pourraient expliquer les différences observées dans les résultats expérimentaux.

De plus, la définition de l'objet d'étude semble différer selon les auteurs. Alors que la plupart des études se sont concentrées sur l'énoncé critique en la comparant au discours sincère en termes de prosodie, Li & Gu (2021) ont comparé les différences phonétiques entre les éloges ironiques et les critiques directes. La critique ironique exprime une attitude négative envers l'auditeur en utilisant un discours sincère dont le sens littéral est de faire l'éloge, tandis que les éloges ironiques expriment une attitude positive en utilisant des énoncés dont le sens littéral est de critiquer. Il a été établi et largement reconnu dans la recherche que différentes significations émotionnelles sont véhiculées par des combinaisons spécifiques de caractéristiques prosodiques reflétant les attitudes et les états émotionnels du locuteur à l'égard de l'interprétation du discours (Brown et al., 2014; Cole, 2015). Ainsi, les énoncés ironiques exprimant des attitudes différentes peuvent nécessiter des combinaisons différentes de caractéristiques onomatopéiques pour transmettre le sens de l'expression.

Cette étude vise à apporter des preuves supplémentaires soutenant l'existence d'un patron tonal de l'ironie spécifique à chaque langue en se basant sur la production de l'ironie dans le français et le mandarin à l'aide du même protocole. La critique ironique a été au centre des recherches précédentes, et cette étude perpétue cette convention en choisissant la critique ironique et les éloges littéraux comme sujets de recherche.

Nous émettons l'hypothèse que la caractéristique de l'intonation ironique de chaque langue est unique, reposant sur des corrélats acoustiques présentant des schémas différents. Par ailleurs, cette recherche nous permet de vérifier les corrélats acoustiques déjà révélés dans des recherches antérieurs pour les deux langues ; ainsi si l'expression de l'ironie en chinois mandarin et en français sont marqués par un débit de parole plus lent et une intensité plus basse, la gestion de la f0 est différente, le mandarin étant parqué par une hauteur réduite et une étendue de la hauteur plus étroite, tandis que le français se caractérise par une hauteur élevée et une étendue de la haute plus large.

2 Protocole expérimental

2.1 Participants

Vingt participants ont été recrutés, comprenant 10 locuteurs natifs de français (âge moyen 23,7 ans (18-30) ; ET 3,3 ; 5 femmes) et 10 locuteurs natifs de mandarin (âge moyen 25,6 ans (22-30) ; ET 2,6 ; 5 femmes). Tous les locuteurs ont rapporté qu'ils ne présentaient aucuns troubles émotionnels diagnostiqués ou de conditions physiologiques susceptibles d'affecter potentiellement la vocalisation.

2.2 Tâche et procédure

L'expérience se base sur une structure narrative comprenant 12 scénarios adaptés de Spotorno et al. (2012). Chaque scénario attribuait aux participants le rôle de l'un des interlocuteurs dans une conversation quotidienne simulée avec des connaissances ou amis.

Deux types de scènes ont été créés : les critiques ironiques (CI) et son équivalent, les éloges littéraux (EL). Chaque scénario se compose de quatre lignes de contexte, suivies d'une phrase cible à produire. Les deux premières lignes offrent un contexte général. La troisième ligne contient des informations contextuelles critiques susceptibles de donner à la phrase cible une signification ironique ou littérale. Les phrases cibles peuvent être de deux modalités phrastiques différentes : déclaratives et exclamatives. En français, l'exclamative commence par le mot "quel". En mandarin, l'exclamative contient le mot "duome", signifiant également "quel". Un exemple est donné dans le Tableau 1.

	Crtiques Ironiques	Éloges littéraux
Contexte	Léa et toi chantez dans le même opéra.	
	Le soir de la première, vous vous retrouvez au théâtre.	
	Durant la représentation, vous faites beaucoup de fausses notes.	La représentation est excellente et on vous a longuement applaudis.
	Après le spectacle, tu dis à Léa :	
Phrase cible	Ce soir on a fait une performance magistrale.	

TABLE 1 : Exemple de stimuli utilisés pour susciter la critique ironique (CI) et l'éloge littéral (EL).

Les participants étaient installés dans une cabine insonorisée à l'Université Paris Cité pour garantir un environnement acoustique contrôlé. L'expérience a été menée à l'aide d'une présentation Microsoft affichée sur un écran d'ordinateur. Ils ont été informés qu'ils liraient une série de scripts de quatre lignes, endossant le rôle d'un personnage et prononçant la ligne à laquelle ils estimaient approprié pour la scène. À chaque essai, les participants ont reçu des consignes pour se concentrer sur le contenu des trois lignes du script, en considérant sa pertinence pour la ligne qu'ils devaient prononcer. Une fois préparés, les participants étaient libres de lire l'énoncé cible à leur propre rythme, sans limite de temps. Les participants pouvaient répéter l'énoncé plusieurs fois jusqu'à satisfaction, mais seule de leur dernière tentative était prise en compte et utilisée.

2.3 Enregistrement

Les enregistrements ont été réalisés à l'aide d'un microphone casque Shure WH20XLR connecté à un dispositif sonore USB (Komplete Audio 2) et numérisés à une fréquence d'échantillonnage de 44 kHz, en mono-canal et avec une quantification sur 16 bits. Un total de 480 énoncés (20 participants × 2 types de scènes × 12 histoires) ont été enregistrés. Après avoir éliminé les énoncés mal prononcés (19 sur 480), 228 énoncés en français (CI = 116, EL = 112) et 233 énoncés en mandarin (CI = 116, EL = 117) ont été conservés.

2.4 Traitement acoustique des données

Les mots et les phonèmes ont été automatiquement segmentés et alignés à l'aide du Montreal Forced Aligner (McAuliffe et al., 2017). Pour garantir la précision, toutes les frontières entre les mots ont été vérifiées et corrigées manuellement. L'annotation et l'analyse acoustique ont été réalisées à l'aide de Praat (version 6.3.10) (Boersma & Weenink, 2001).

Les caractéristiques acoustiques de l'ironie ont été mesurées selon cinq dimensions. La modification de la hauteur impliquait l'analyse des variations de la fréquence fondamentale moyenne (F0) et de l'étendue de la F0. Le débit de parole a été calculé en divisant la durée totale de la phrase par le nombre total de syllabes. La modulation d'intensité examinait l'intensité moyenne et l'étendue de l'intensité. L'évaluation de la qualité de la voix comprenait le jitter (ddb) et le shimmer (local en dB). De plus, le bruit a été mesuré à travers le rapport harmoniques-bruit (HNR) moyen sur l'ensemble de la plage de fréquences. Un total de huit corrélats acoustiques ont été pris en compte pour l'analyse ultérieure. Les mesures ont été effectuées à l'aide de ProsodyPro (Xu, 2013) pour la hauteur, le débit de parole et l'intensité, tandis que la qualité de la voix a été mesurée sur noyau de syllabe à l'aide d'un script Praat adapté d'Al-Tamimi (2022).

2.5 Analyse statistique

Tous les paramètres acoustiques extraits ont été analysés dans le langage R, version 4.1.2 (R Core Team, 2021), à l'aide du logiciel Rstudio, version 2023.03.1.446 (RStudio Team, 2023).

Nous avons initialement employé l'algorithme de classification forêts aléatoires (Breiman, 2001) pour étudier les schémas acoustiques pour produire l'ironie en calculant la contribution relative de chaque paramètre acoustique à la production des tons ironiques.

Nous avons utilisé la fonction *cforest()* du package party (Hothorn et al., 2006; Strobl et al., 2007) pour construire des modèles de forêts aléatoires basés sur des Arbres d'Inférence Conditionnelle en utilisant les 8 corrélats acoustiques. Dans notre analyse, nous avons fixé *mtry* à 3, correspondant à la racine carrée arrondie des 8 prédicteurs. Le nombre optimal d'arbres (*ntree*) a été déterminé en utilisant une méthode basée sur des métriques de densité[1] employé par des études précédentes (Al-Tamimi, 2017; Al-Tamimi & Khattab, 2018). Pour estimer le nombre optimal d'arbres, nous avons généré 20 forêts aléatoires en utilisant le package party avec 100 itérations par forêt. Nous avons ensuite réalisé une comparaison AUC (Zone Sous la Courbe) en utilisant le package pROC (Robin et al., 2011), en générant une courbe ROC. Les valeurs AUC obtenues ont été comparées, et le groupe avec la valeur AUC la plus élevée a déterminé le nombre optimal d'arbres. Après calcul, il a été constaté que les données en mandarin nécessitaient 1200 arbres pour le meilleur ajustement du modèle, tandis que les données en français nécessitaient 1300 arbres. Par conséquent, nous avons exécuté des forêts aléatoires avec ces valeurs respectives de *ntree* pour assurer les performances optimales du modèle pour chaque ensemble de données. L'importance des variables par permutation conditionnelle a ensuite été calculée pour évaluer le poids relatif de chaque corrélat acoustique dans la présentation des schémas acoustiques pour marquer l'ironie en mandarin et en français.

Deuxièmement, basé sur les résultats de l'analyse de classification Forêt aléatoire, nous avons construit une série de modèles linéaires à effets mixtes (LMM) sur les corrélats acoustiques identifiés comme relativement plus importants. Chaque paramètre acoustique a été analysé avec la fonction *lmer()* du package lme4 (Bates et al., 2015). Dans le modèle, l'Attitude (EL = 0, CI = 1) et le Sexe (Femme = 0, Homme = 1) ont été inclus en tant qu'effets fixes. Pour tenir compte des effets aléatoires, le Participant et le Type de Phrase ont été spécifiés comme interceptes aléatoires. De plus, le modèle comprenait des pentes aléatoires pour l'Attitude par Participant et le Sexe par Type de Phrase. Par conséquent, le modèle final pour chaque paramètre acoustique peut être représenté par la formule suivante :

[1] La méthode est présentée sur le site web suivant : https://jalalal-tamimi.github.io/R-Estimating-Number-Of-Trees-RF/

Paramètre Acoustique ~ 1 + Attitude + Sex + (1 + Attitude | Participant) + (1 + Sex | Type de Phrase)

3 Résultats

3.1 Résultats des forêts aléatoires

Les Figures 1 et 2 présentent les scores d'importance relative des huit paramètres acoustiques pour l'ironie en mandarin et en français.

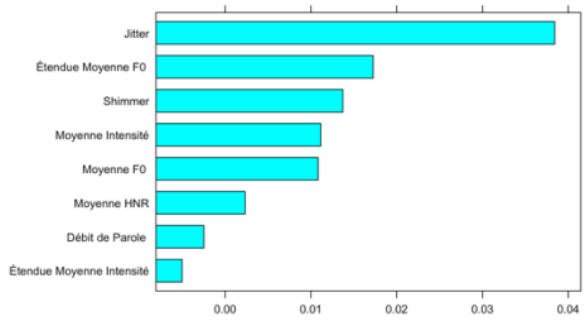


FIGURE 1 : Scores d'importance relative des huit paramètres acoustiques pour l'ironie en français.

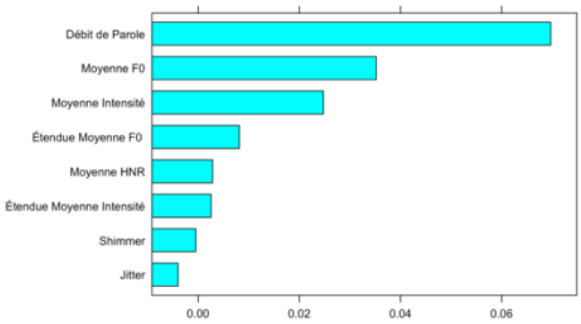


FIGURE 2 : Scores d'importance relative des huit paramètres acoustiques pour l'ironie en mandarin.

Les résultats de notre analyse révèlent des schémas distincts dans l'utilisation des corrélats acoustiques pour distinguer entre les énoncés en CI et EL pour les locuteurs français et mandarins. Pour les locuteurs français, le jitter émerge comme le corrélat acoustique le plus influent, suivi par l'étendue moyenne de la F0, le shimmer, la moyenne de l'intensité et la moyenne de HNR. En revanche, l'étendue moyenne de l'intensité, le débit de parole et la moyenne de F0 présentent un pouvoir prédictif relativement plus faible pour la distinction entre les deux types de discours chez les locuteurs français.

En revanche, les locuteurs mandarins se basent sur des indices acoustiques différents pour produire l'ironie. Comme le montre la FIGURE 2, le débit de parole joue un rôle pour l'énoncé ironique produit par les locuteurs natifs de mandarin. De plus, la moyenne de F0 et la moyenne de l'intensité contribuent également à la différenciation entre les CI et les EL en mandarin. Notamment, l'étendue de l'intensité, le shimmer et le jitter présentent une contribution minimale pour expliquer les différences entre les énoncés CI et EL en mandarin.

3.2 Résultats de LMM

Sur la base de ces résultats de la forêt aléatoire, notre deuxième analyse avec des LMM a été réalisée sur les trois premiers corrélats acoustiques qui ont été identifiés comme des indices principaux de l'ironie. Pour le français, nous avons sélectionné le jitter, l'étendue moyenne de la F0, le shimmer, la moyenne de l'intensité et la moyenne de HNR. Pour le mandarin, nous avons sélectionné le débit

de parole, la moyenne de F0 et la moyenne de l'intensité. Les résultats sont présentés dans le Tableau 2 et le Tableau 3.

Comme indiqué dans le Tableau 2, notre analyse a révélé une diminution statistiquement significative du jitter lors de la production de la CI par rapport à la EL ($p < 0,050$). En ce qui concerne l'étendue de la F0, les données indiquent une diminution de 0,236 dans l'étendue de la F0 pour marquer l'ironie en français. Cependant, les changements ne présentent pas de différences statistiquement significatives ($p = 0,127$). Pour ce qui est du résultat du shimmer, nous avons observé une tendance dans le shimmer entre les énoncés CI et EL en français : une diminution statistiquement significative du shimmer de 0,257 lors de la production de CI par rapport à EL en français ($p = 0,060$). Nos résultats indiquent également qu'en produisant de l'ironie, les locuteurs français présentent une diminution statistiquement significative de l'intensité moyenne ($p = 0,050$). Cependant, en ce qui concerne la moyenne de F0, les changements ne montrent pas de différences statistiquement significatives ($p = 0,280$).

Pour le mandarin, il y a eu une diminution statistiquement significative de 0,553 dans le débit de parole pour la CI par rapport au EL ($p < 0,005$). En ce qui concerne la moyenne de F0, CI présente une diminution statistiquement significative de 0,268 par rapport au EL ($p<0,050$). Une diminution de 0,423 dans l'intensité moyenne était également statistiquement significative pour produire l'ironie en mandarin ($p < 0,005$).

Corrélats acoustiques	Estimate	Std. Error	df	t value	P value
Jitter	**-0,366**	**0,123**	**8,661**	**-2,968**	**0,016**
L'étendue moyenne de la F0	-0,236	0,147	17,987	-1,598	0,127
Shimmer	**-0,257**	**0,119**	**8,598**	**-2,160**	**0,060**
Moyenne de l'intensité	**-0,160**	**0,071**	**9,052**	**-2,256**	**0,050**
Moyenne de la F0	-0,100	0,087	9,097	-1,149	0.280

TABLE 2 : Résultats des modèles linéaires à effets mixtes (LMM) pour l'ironie en français

Corrélats acoustiques	Estimate	Std. Error	df	t value	P value
Débit de parole	**-0,553**	**0,129**	**8,991**	**-4,275**	**0,002**
Moyenne de la F0	**-0,268**	**0,084**	**9,071**	**-3,196**	**0,011**
Moyenne de l'intensité	**-0,423**	**0,157**	**9,081**	**-2,694**	**0,024**

TABLE 3 : Résultats des modèles linéaires à effets mixtes (LMM) pour l'ironie en mandarin

4 Discussion et conclusion

Les résultats de l'analyse forêts aléatoires ont soutenu notre hypothèse selon laquelle la caractéristique de l'intonation ironique est spécifique à chaque langue. Nous avons constaté que les locuteurs natifs français ont tendance à manipuler la qualité de leur voix (ex. F0, jitter et shimmer) et l'intensité pour produire l'ironie, tandis que les locuteurs natifs mandarins dépendent fortement de la réduction de leur débit de parole, de leur hauteur et de leur intensité. De plus, nos résultats suggèrent que les schémas des caractéristiques de l'intonation ironique en mandarin et en français sont inverses. Le jitter et le shimmer, deux des principaux corrélats acoustiques pour marquer l'ironie en français, sont les corrélats acoustiques les moins importants pour les locuteurs mandarins natifs. De même, le débit de parole, qui est le corrélat acoustique le plus important pour marquer l'ironie en mandarin, est l'avant-dernier corrélat acoustique pour le français.

Nos résultats des LMM ont montré que l'ironie en mandarin est caractérisée par un débit de parole réduit, une fréquence fondamentale moyenne plus basse et une intensité plus faible. Tous les résultats sont conformes aux conclusions des recherches antérieures (Li et al., 2020; Li & Gu, 2021). Cependant, Li & Gu (2020) comparent le compliment ironique avec le blâme direct, tandis que dans notre étude, nous comparons la critique ironique avec les éloges littéraux. Cela pourrait suggérer qu'en mandarin, il n'existe qu'un seul schéma acoustique pour l'ironie, indépendamment de l'émotion et de l'attitude réelles. Pour explorer cette hypothèse, nous avons déjà mené une étude comparant les quatre types de paroles en mandarin, et les résultats sont en cours d'analyse.

Nos résultats des LMM ont également que l'ironie en français est marquée par une diminution du jitter et une diminution de la moyenne de l'intensité. Nous avons constaté que la modulation de la F0 (ex. la moyenne de la F0 et l'étendue moyenne de la F0) et une diminution du shimmer ne sont pas des corrélats importants pour marquer l'ironie en français. Nous avançons que le sexe des participants pourrait avoir un effet sur les résultats. En effet, des différences de genre dans les caractéristiques acoustiques dans la production de l'ironie ont déjà été rapportées dans la littérature (Chen & Boves, 2018; Li et al., 2020). Nos résultats des LMM suggèrent effectivement une différence statistiquement significative de genre dans l'utilisation de l'étendue de la F0 ($p < 0,050$), de la moyenne de la F0 ($p < 0,005$) et du shimmer ($p < 0,005$) pour produire l'ironie en français. La différence de genre dans l'utilisation des corrélats acoustiques et le mécanisme sous-jacent méritent une étude approfondie.

En conclusion, notre approche, qui utilise un protocole expérimental uniforme dans différentes langues, établit une base plus fiable pour les études inter-langues sur la caractéristique de l'intonation ironique. En utilisant des outils d'analyse de données tels que forêts aléatoires et LMM, notre étude intègre les corrélats acoustiques avec le sens des phrases, offrant ainsi une analyse approfondie des schémas du patron tonal utilisés pour marquer l'ironie dans le français et le mandarin.

Remerciements

Nous exprimons notre gratitude à Prof. Ioana Chitoran pour ses précieux commentaires et son feedback sur cet article. Nos remerciements vont également à tous les participants ayant pris part à cette étude. Ce travail a bénéficié partiellement d'une aide de l'IdEx Université Paris Cité (ANR-18-IDEX-0001) au titre du Labex Empirical Foundations of Linguistics - EFL et l'opération ProCue au sein de l'axe 2 du Labex. Le premier auteur bénéficie d'une bourse doctorale financée par le China Scholarship Council (CSC).

Références

AL-TAMIMI J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic : Implications for formal representations. Laboratory Phonology, 8(1), Article 1. https://doi.org/10.5334/labphon.19

AL-TAMIMI J. (2022). JalalAl-Tamimi/Praat-VQ-Measurements : Praat VQ measurements (Version v2) [Logiciel]. Zenodo. https://doi.org/10.5281/zenodo.7270191

AL-TAMIMI J., & KHATTAB G. (2018). Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops. Journal of Phonetics, 71, 306-325. https://doi.org/10.1016/j.wocn.2018.09.010

ANOLLI L., CICERI R., & INFANTINO M. G. (2002). From « blame by praise » to « praise by blame » : Analysis of vocal patterns in ironic communication. International Journal of Psychology, 37, 266-276. https://doi.org/10.1080/00207590244000106

BATES D., MÄCHLER M., BOLKER B., & WALKER S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1-48. https://doi.org/10.18637/jss.v067.i01

BOERSMA P., & WEENINK D. (2001). PRAAT, a system for doing phonetics by computer. Glot international, 5, 341-345.

BREIMAN L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

BROWN L., WINTER B., IDEMARU K., & GRAWUNDER S. (2014). Phonetics and politeness : Perceiving Korean honorific and non-honorific speech through phonetic cues. Journal of Pragmatics, 66, 45-60. https://doi.org/10.1016/j.pragma.2014.02.011

BRYANT G. A., & FOX TREE J. E. (2005). Is there an Ironic Tone of Voice? Language and Speech, 48(3), 257-277. https://doi.org/10.1177/00238309050480030101

CHEANG H. S., & PELL M. D. (2008). The sound of sarcasm. Speech Communication, 50(5), 366-381. https://doi.org/10.1016/j.specom.2007.11.003

CHEN A., & BOVES L. (2018). What's in a word : Sounding sarcastic in British English. Journal of the International Phonetic Association, 48(1), 57-76. https://doi.org/10.1017/S0025100318000038

COLE J. (2015). Prosody in context : A review. Language, Cognition and Neuroscience, 30(1-2), 1-31. https://doi.org/10.1080/23273798.2014.963130

GONZALEZ-FUENTE S., PRIETO P., & NOVECK I. (2016). A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French. Speech Prosody 2016, 902-906. https://doi.org/10.21437/SpeechProsody.2016-185

HOTHORN T., HORNIK K., & ZEILEIS A. (2006). Unbiased Recursive Partitioning : A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651-674. https://doi.org/10.1198/106186006X133933

JANSEN N., & CHEN A. (2020). Prosodic encoding of sarcasm at the sentence level in Dutch. Speech Prosody 2020, 409-413. https://doi.org/10.21437/SpeechProsody.2020-84

LAVAL V., & BERT-ERBOUL A. (2018, 49-12 10:49:46). French-Speaking Children's Understanding of Sarcasm. ASHA Wire. https://pubs.asha.org/doi/epdf/10.1044/1092-4388%282005/042%29

LI S., & GU W. (2021). Prosodic Profiles of the Mandarin Speech Conveying Ironic Compliment. 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 1-5. https://doi.org/10.1109/ISCSLP49672.2021.9362092

LI S., GU W., LIU L., & TANG P. (2020). The Role of Voice Quality in Mandarin Sarcastic Speech : An Acoustic and Electroglottographic Study. Journal of Speech, Language, and Hearing Research, 63(8), 2578-2588. https://doi.org/10.1044/2020_JSLHR-19-00166

LŒVENBRUCK H., JANNET M. A. B., D'IMPERIO M., SPINI M., & CHAMPAGNE-LAVAU M. (2013). Prosodic cues of sarcastic speech in French : Slower, higher, wider. Interspeech 2013, 3537-3541. https://doi.org/10.21437/Interspeech.2013-761

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M., & SONDEREGGER M. (2017). Montreal Forced Aligner : Trainable Text-Speech Alignment Using Kaldi. Interspeech 2017, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

R CORE TEAM. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

ROBIN X., TURCK N., HAINARD A., TIBERTI N., LISACEK F., SANCHEZ J.-C., & MÜLLER M. (2011). pROC : An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12(1), 77. https://doi.org/10.1186/1471-2105-12-77

ROCKWELL P. (2000). Lower, slower, louder : Vocal cues of sarcasm. Journal of Psycholinguistic Research, 29, 483-495. https://doi.org/10.1023/A:1005120109296

RSTUDIO TEAM. (2023). RStudio : Integrated Development Environment for R. RStudio, PBC. http://www.rstudio.com/

SCHARRER L., & CHRISTMANN U. (2011). Voice modulations in German ironic speech. Language and Speech, 54, 435-465. https://doi.org/10.1177/0023830911402608

SPOTORNO N., KOUN E., PRADO J., VAN DER HENST J.-B., & NOVECK I. A. (2012). Neural evidence that utterance-processing entails mentalizing : The case of irony. NeuroImage, 63(1), 25-39. https://doi.org/10.1016/j.neuroimage.2012.06.046

STROBL C., BOULESTEIX A.-L., ZEILEIS A., & HOTHORN T. (2007). Bias in Random Forest Variable Importance Measures : Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). https://doi.org/10.1186/1471-2105-8-25

XU Y. (2013). ProsodyPro—A Tool for Large-scale Systematic Prosody Analysis. https://api.semanticscholar.org/CorpusID:33738277

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Utilisation de wav2vec 2.0 pour des tâches de classifications phonétiques : aspects méthodologiques

Lila Kim[1] Cédric Gendroit[1]
(1) **Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle)**, 4 rue des Irlandais, 75005 Paris, France
lila.kim@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr

RÉSUMÉ

L'apprentissage auto-supervisé, particulièrement dans le contexte de la parole, a démontré son efficacité dans diverses tâches telles que la reconnaissance du locuteur et la reconnaissance de la parole. Notre question de recherche se concentre sur l'efficacité des représentations vectorielles - extraites de phonèmes - plus courtes par rapport à des séquences plus longues dans la détection de la nasalité. Deux approches distinctes ont été étudiées : extraire des vecteurs sur la durée du phonème et prendre des séquences plus longues avec une seconde ajoutée de chaque côté du phonème, puis récupérer la partie centrale a posteriori. Les résultats révèlent que les modèles réagissent différemment selon les phones et les locuteurs, avec une variabilité observée à ces niveaux. Le modèle à séquences longues surpasse le modèle à séquences courtes en assurant une corrélation plus robuste avec le débit d'air nasal.

ABSTRACT

Using wav2vec 2.0 for phonetic classification tasks : methodological aspects

Self-supervised learning, particularly in the context of speech, has been shown to be effective in a variety of tasks such as speaker recognition and automatic speech recognition. Our research question focuses on the effectiveness of vector representations extracted from shorter versus longer phoneme sequences in detecting nasality. Two distinct approaches were studied : extracting vectors over the duration of the phoneme and taking longer sequences with a second added on each side of the phoneme, then recovering the central part a posteriori. The results show that the models react differently depending on the phone and the speaker, with variability observed at both levels. The long sequence model outperformed the short sequence model by correlating more robustly with nasal airflow.

MOTS-CLÉS : parole, wav2vec 2.0, nasalité, physiologie.

KEYWORDS: speech, wav2vec 2.0, nasality, physiology.

1 Introduction

Depuis l'utilisation récurrente de l'apprentissage auto-supervisé dans les tâches de reconnaissance automatique de la parole, plusieurs études ont appliqué ces modèles de Transformers à des domaines tels que la reconnaissance du locuteur, la détection de code-switching ou d'émotions, etc. (Fan et al., 2021; Pepino et al., 2021; Tseng et al., 2021; Cormac English et al., 2022). De plus, Pasad et al. ont montré notamment que les informations diffèrent dans les représentations vectorielles selon les

couches de Transformers (Pasad et al., 2022, 2023). Il est à noter qu'un modèle de Transformers tel que wav2vec 2.0 prend en compte les informations contextuelles dans une séquence et travaille généralement sur des séquences de plusieurs secondes (Baevski et al., 2020). C'est dans ce contexte que s'inscrit notre question de recherche : les représentations vectorielles extraites sur une séquence de phonèmes permettraient-elles une meilleure performance dans la détection de la nasalité par rapport à une séquence plus longue ? Notre travail consiste donc à explorer la longueur de la séquence pour la prise de vecteurs : la première consiste à prendre des vecteurs sur la durée du phonème, tandis que la seconde consiste à prendre une séquence plus longue en ajoutant une seconde dans les deux côtés du phonème et à récupérer la partie centrale a posteriori.

En premier lieu, nous décrirons les ressources utilisées lors de l'entraînement et du test, ainsi que les méthodes d'extraction utilisées avec le modèle auto-supervisé wav2vec 2.0. En second lieu, nous nous pencherons sur deux approches différentes pour l'extraction des vecteurs et la détection de la nasalité à l'aide d'une régression logistique. Nous évaluerons ensuite le modèle entraîné sur des données acoustiques, en comparant les résultats avec des données physiologiques obtenues simultanément avec l'acoustique, servant ainsi de référence.

2 État de l'art

2.1 Modélisation acoustique

La parole est un phénomène complexe influencé par divers éléments tels que l'articulation, l'origine géographique ou sociale du locuteur, son état émotionnel et des aspects pragmatiques comme l'auditeur. Pour la transcription automatique de la parole et la modélisation du signal acoustique, des approches ont évolué des systèmes experts vers des approches neuronales. Les premières approches se concentraient sur l'aspect linguistique, alors que les méthodes probabilistes, notamment les modèles de Markov cachés, ont commencé à dominer à partir des années 90. (Juang and Rabiner, 1991; Patel and Srinivas Rao, 2010).

Avec l'avènement des modèles connexionnistes, la phonétisation, l'utilisation de connaissances psycho-acoustiques et le traitement du signal se sont transformés vers des méthodes entièrement neuronales, combinant les représentations spectrales telles que les MFCC avec des perceptrons multicouches. Pour surmonter les défis du Deep Learning tels que le besoin de grandes quantités de données et le manque d'annotations manuelles, des approches d'apprentissage légèrement supervisées ou auto-supervisées ont été entreprises. Ces modèles sont préalablement entraînés sur de grands nombres d'heures d'audio non annoté, puis ajustés sur des ensembles de données annotées de plus moindre taille pour des tâches spécifiques (Baevski et al., 2020). En utilisant la méthode du "probing", Pasad et al. analysent les informations contenues dans les différentes couches de ces modèles en cherchant à mieux comprendre la nature des données à ces niveaux (Pasad et al., 2023, 2022).

2.2 Nasalité

La nasalité, fréquemment considérée comme une composante de la qualité de la voix, est une caractéristique omniprésente dans les langues du monde, se produit lorsque le voile du palais s'abaisse. Ce phénomène crée des effets acoustiques distincts sur les sons nasals (Maeda, 1982). Elle est

essentielle dans la production de la parole pour distinguer phonologiquement les sons nasals des sons oraux, que ce soit dans le cas des voyelles (comme /a/ et /ã/, par exemple) ou des consonnes (comme /b/ et /m/, par exemple). La nasalité d'un son peut être propagée à son voisin oral en raison de réalisations articulatoires, telles qu'un abaissement prématuré, un relèvement tardif du velum (Amelot et al., 2008; Brkan, 2018). Cette coarticulation nasale, influencée par le contexte phonémique, peut se produire dans des langues où la nasalité est un trait phonologique distinctif (comme le français, où /a/ dans "maman" est nasalisé), mais aussi dans des systèmes de langue où cette distinction n'est pas présente (comme en anglais, par exemple dans "can't").

La qualité de voix a de grandes implications dans la caractérisation du locuteur (Gold and French, 2019). Elle peut être un élément permanent de la voix d'un locuteur due à des facteurs physiologiques, mais aussi sujette à la variabilité intra-locuteur, notamment dans le style de discours ou l'émotion (Nolan, 2014). Les nasales offrent une caractéristique fiable pour la reconnaissance des locuteurs (Kahn, 2011) en raison de la morphologie de la cavité nasale stable et variable entre locuteurs (Dang et al., 1994; Serrurier, 2006). Cependant, l'analyse acoustique de la nasalité est complexe car le couplage de deux cavités provoquent des modifications acoustiques en engendrant des pôles et zéros sur le spectre acoustique. Bien que les méthodes d'analyse aient été entreprises pour la nasalité (Chen, 1997; Styler, 2017), elles sont très influencées par les caractéristiques articulatoires propres à chaque son, et à chaque locuteur.

3 Protocole expérimental

3.1 Données pour l'entraînement

Pour l'entraînement et la validation, nous avons extrait les différents types de phones à partir de quatre corpus distincts, chacun représentant un type de parole spécifique. Les corpus de données utilisés dans cette étude comprennent :

1. NCCFr (The Nijmegen Corpus of Casual French) : conversations amicales, impliquant 46 locuteurs français. (Torreira et al., 2010) ;
2. ESTER (Evaluation de Systèmes de Transcription enrichie d'Emissions Radiophoniques) : conversations radiophoniques en français, parole préparée et lue (Gravier et al., 2004; Galliano et al., 2006). Seule une partie de 30 heures a été retenue pour cet entraînement.
3. PTSVOX : créé pour évaluer les variations intra- et inter-locuteurs. (Chanclu et al., 2020). Nous n'avons retenu qu'une petite partie de ce corpus avec des alignements vérifiés, pour les productions de seulement 24 locuteurs ;
4. BREF : développé dans le but du développement et de l'évaluation des systèmes de reconnaissance de la parole, parole continue. (Lamel et al., 1991). Là encore, tous les alignements en phonèmes ne nous ayant pas été communiqués, seule la moitié du corpus BREF a été utilisée pour les entraînements.

Dans le cadre de ce travail, nous avons décidé d'extraire 8 voyelles et 7 consonnes nasales et orales confondues. Les voyelles sujettes à l'extraction sont 3 paires de voyelles /a,ɛ,o,ɑ̃,ɛ̃,ɔ̃/ qui peuvent se distinguer par le trait de nasalité [± nasal]. Nous sommes conscients de la distinction articulatoire entre une voyelle orale et sa contrepartie nasale, cependant, dans le contexte de cette étude, nous avons choisi de concentrer notre attention sur la nasalité en particulier. Deux voyelles /e,ɔ/ ont été ajoutées car la phonétisation des voyelles moyennes en français n'est pas toujours systématique. En

ce qui concerne les consonnes, nous avons retenu quatre consonnes orales et trois consonnes nasales : /b,d,v,l,m,n,ɲ/. Elles présentent différentes manières et lieux d'articulation (bilabiale, labio-dentale, dentale, alvéolaire, et occlusive ou fricative). La même liste de phonèmes a été utilisée pour les deux approches mentionnées ultérieurement dans la section 3.3.3, sans prendre en compte le contexte phonétique des phonèmes examinés.

3.2 Données acoustiques et physiologiques pour l'évaluation du modèle

Les données de test se composent de deux parties : acoustique et physiologique. Elles ont été recueillies simultanément à l'aide d'un masque "Aeromask" développé au Laboratoire de Phonétique et Phonologie (Elmerich et al., 2023). Ce masque enregistre la voix ainsi que les débits d'air nasal et buccal sans perturber la propagation sonore, ce qui permet d'utiliser l'acoustique pour évaluer le réseau de neurones et les débits d'air pour vérifier la présence de nasalité dans les phones évalués. Les enregistrements des phrases ont été réalisés avec six locuteurs masculins, tous natifs du français. Les stimuli étaient insérés dans des mots sans signification littérale (i.e., logatomes) sous forme de VCV ou VNV, où C représente [p,b,t,d,v,s,z], N représente [m,n], et V représente [i,a,y,u,o,e,ɑ̃,ɛ̃,ɔ̃]. (Elmerich et al., 2020, 2023). Ces séquences de stimuli ont été intégrées dans une phrase de cadre : « Non tu n'as pas dit XXX quatre fois, mais tu as dit YYY et ZZZ quatre fois ». Ainsi, les mots XXX, YYY et ZZZ correspondent à des structures VCV ou VNV. La segmentation manuelle a été effectuée pour ces données. À partir de ces listes de phones, nous avons sélectionné les mêmes phones que pour l'entraînement, à l'exception de /l/ qui n'est pas présent dans la liste, ce qui donne /a,ɛ,o,ɑ̃,ɛ̃,ɔ̃,b,d,v,m,n/. En résumé, 269 sons de chaque classe ont été extraits au total. Les mesures aérodynamiques des phones ont été consignées dans un fichier au format CSV en vue d'une comparaison ultérieure avec les résultats du réseau de neurones profonds. Les données utilisées pour l'entraînement, la validation et le test sont récapitulées dans le tableau 1. Les phonèmes des données de test ont été répartis selon le nombre suivant :ɑ̃ (66), ɛ̃ (66), ɔ̃ (66), m (36), n (35), a (66), E (66), o (66), b (25), d (29), v (17).

Jeu de données	**Phone [+ nasal]**	**Phone [- nasal]**
Entraînement	60 000	60 000
Validation	15 000	15 000
Test	269	269

TABLE 1 – statistiques des données utilisées pour l'entraînement, la validation et le test

Les données aérodynamiques que nous utiliserons comme référence consistent en trois valeurs : le débit d'air nasal (DAN), le débit d'air buccal (DAB) et le débit d'air nasal proportionnel. Le calcul de ces valeurs est expliqué dans (Kim et al., 2023).

3.3 Méthodologie

3.3.1 Wav2vec 2.0

Notre recherche s'inscrit dans le cadre de l'exploration de la manière dont le modèle wav2vec 2.0 encode l'information de nasalité dans ses représentations vectorielles. Nous nous concentrons

particulièrement sur le modèle "wav2vec 2.0-FR-3K-large-LeBenchmark", pré-entraîné sur 2 900 heures de divers types de discours en français (spontané, lu et diffusé). Ce modèle a été spécifiquement conçu pour optimiser ses performances dans des tâches liées au français (Parcollet et al., 2023).

Le fonctionnement du modèle wav2vec 2.0 consiste à prendre le signal brut comme données d'entrée, traitées par l'encodeur convolutionnel. Toutes les 25 millisecondes d'audio sont transformées en une séquence de vecteurs, avec un chevauchement de 5 millisecondes entre chaque paire d'échantillons. Ces séquences subissent une normalisation et une fonction d'activation GELU avant d'être acheminées vers les transformers. Pendant la phase de pré-entraînement, le module de quantification est utilisé pour discrétiser les valeurs de sortie de l'encodeur. Les représentations latentes obtenues de l'encodeur subissent ensuite une analyse et une contextualisation par les couches de Transformers, qui capturent l'information sur l'ensemble de la séquence. Le modèle large, en particulier, comporte 24 couches de transformation, chacune produisant un vecteur de 1 024 dimensions en représentations latentes. Les dimensions de la feed-forward sont de 4 096, avec 16 mécanismes d'attention (Baevski et al., 2020).

3.3.2 Génération des représentations vectorielles

L'approche d'extraction des embeddings s'appuie sur la méthodologie présentée par (Guillaume et al., 2023), dont l'étude se focalise sur une analyse linguistique d'une langue à partir de la parole dans un extrait audio de 5 secondes où la stratégie de max pooling a été utilisée pour agréger les différentes représentations latentes d'un enregistrement en un seul vecteur, qui représente l'ensemble du signal. Avec cette méthode, deux longueurs d'extrait audio ont été étudiées pour obtenir les représentations vectorielles de nos données. La première, inspirée de l'approche phonétique, implique l'extraction des représentations vectorielles directement sur les phonèmes découpées à leurs frontières. Dans notre cas d'étude, la petite taille des fenêtres d'analyse que nous utilisons rend l'affinage (fine-tuning) impossible. La deuxième approche consiste à utiliser des séquences plus longues, en ajoutant une seconde au début et à la fin du phonème. Une fois que le wav2vec 2.0 prend des caractéristiques contextuelles sur toute la séquence, nous récupérons le vecteur du milieu a posteriori. Par exemple, si la voyelle dure 200 ms, nous avons extrait une séquence de 2,2 secondes, puis effectué un max pooling sur les 200 ms du milieu lors de la récupération des caractéristiques vectorielles. De cette manière, l'information contextuelle sur l'ensemble de la séquence de 2,2 secondes peut être capturée par les blocs de transformations et être présente dans le vecteur du milieu qui représenterait le phonème en question. La récupération du milieu a été effectuée en retirant les secondes ajoutées. Nous reconnaissons que le vecteur du milieu ne correspondrait pas parfaitement à l'ensemble du phonème en question, rendant ainsi la comparaison imparfaite. Les représentations vectorielles ainsi obtenues ont été labellisées à l'aide des labels phonologiques des sons [+ nasal] et [- nasal].

3.3.3 Feature probing

Un modèle de régression logistique a été mis en place pour déterminer si le phone prononcé est réalisé avec nasalité (=1) ou sans nasalité (=0). Pour cela, la bibliothèque d'apprentissage automatique en python "scikit-learn" a été utilisée avec les hyperparamètres définis par défaut. La probabilité d'appartenance à la classe nasale est considérée comme une probabilité de nasalité dans les analyses (voir 4.2). La procédure de notre méthodologie est décrite dans la figure 1.

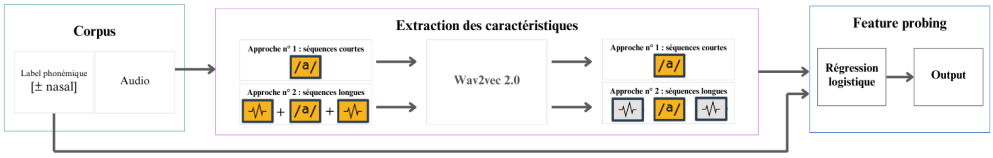


FIGURE 1 – Aperçu de la méthodologie expérimentale comprenant l'architecture du modèle de régression logistique

4 Résultats

L'analyse des résultats obtenus avec les réseaux de neurones profonds par la mesure physiologique aide à vérifier l'indice de nasalité dans la réalisation des phones. Dans la section 4.1, notre objectif est d'établir si la nasalité est détectable lorsqu'un classifieur est basé sur les caractéristiques extraites par le modèle auto-supervisé wav2vec 2.0, et si les erreurs produites par les réseaux peuvent être expliquées par les débits d'air nasal et buccal. Enfin, dans la section 4.2, nous chercherons à déterminer si les classifieurs ont appris à séparer les phonèmes plutôt qu'à détecter la nasalité, en utilisant les mêmes représentations vectorielles.

4.1 Performance du système selon l'approche d'extraction

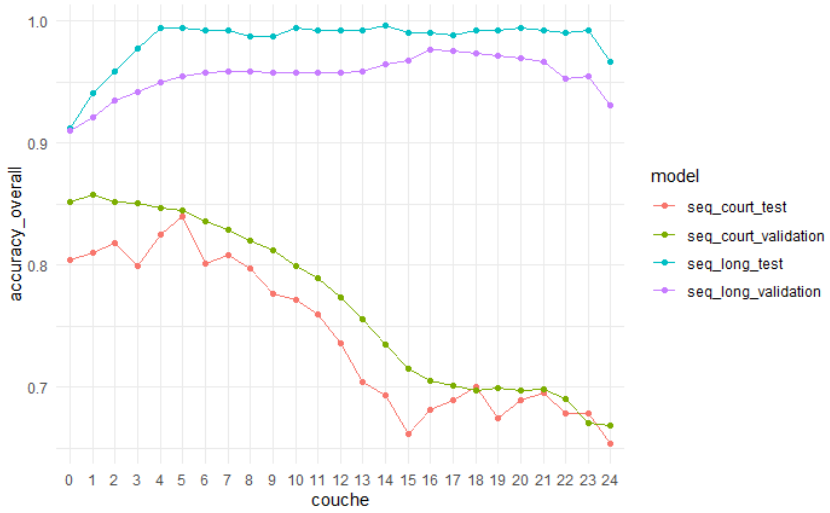


FIGURE 2 – Distribution de l'exactitude globale en fonction des couches du wav2vec 2.0 selon la longueur d'extrait audio

Les taux d'exactitude globale pour la caractéristique de nasalité [± nasal] à travers les différentes couches de Transformers sont illustrés dans la figure 2. Il convient de noter que le seuil sur les probabilités de sortie est fixé à 0,5 pour le choix d'une classe. Afin de déterminer la couche optimale à exploiter, nous avons examiné l'évolution des performances des différentes couches en ce qui concerne la nasalité, allant de l'encodeur CNN à la dernière couche de Transformers. La figure met en évidence la présence d'informations liées à la nasalité dans pratiquement toutes les couches lorsque l'extrait audio est long. Sur les séquences courtes, la nasalité est particulièrement marquée dans la

sortie de l'encodeur CNN et dans les premières couches de Transformers.

Selon Pasad et al., les premières couches du modèle wav2vec 2.0, y compris l'encodeur CNN, sont associées à l'identité acoustique et aux caractéristiques du spectrogramme (Pasad et al., 2022, 2023). À la lumière de ces observations, nous avons décidé de nous focaliser sur la première couche de Transformer afin d'améliorer l'identification de la nasalité en utilisant les caractéristiques acoustiques plutôt que phonémiques. Ainsi, dans le cadre de la classification de la nasalité avec les caractéristiques extraites de la première couche, les performances ont été meilleures pour les séquences longues, avec une exactitude globale de 94.05%, par rapport à 81.04% pour les séquences courtes.

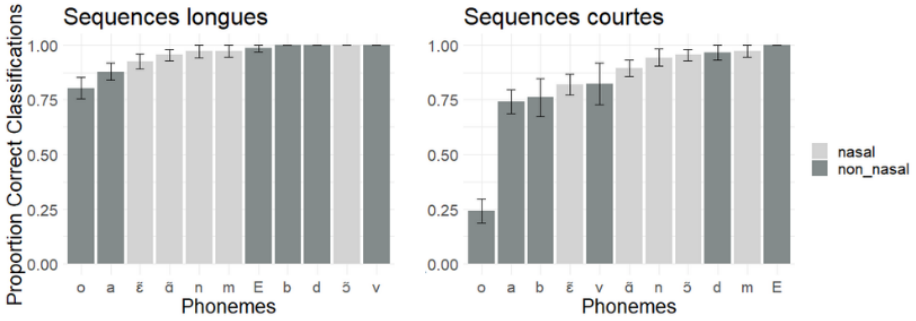


FIGURE 3 – Taux de classification correcte pour chaque phonème (séquences longues à gauche et séquences courte à droite)

Dans la figure 3, la proportion de bonnes attributions de classe pour chaque phonème est représentée. Il convient de noter que dans cette visualisation, /E/ représente /e,ɛ/. Les performances varient selon les phonèmes et les modèles. Par exemple, les phonèmes /ɔ̃,E,m,n,d/ présentent un taux élevé de bonnes classifications, tandis que les voyelles orales /o,a/ sont moins bien classées par les deux modèles. De plus, les voyelles nasales présentent des niveaux de difficulté similaires : /ɛ̃/ est considérée comme la plus difficile à détecter en termes de nasalité, tandis que /ɔ̃/ est identifiée comme la plus facile. En ce qui concerne le modèle à séquences longues, la consonne la plus difficile à prédire est la nasale /n,m/, tandis que pour le modèle à séquences courtes, c'est la consonne orale /b/.

4.2 Comparaison des résultats de classifieurs avec les données physiologiques

Dans cette étude, le coefficient de corrélation de Pearson est utilisé pour examiner la relation linéaire entre la probabilité d'appartenance à la catégorie nasale et le débit d'air nasal. Les corrélations ont été mesurées de trois façons distinctes : (i) nous avons utilisé le débit d'air nasal tel qu'obtenu par l'aeromask (ii) le débit d'air nasal pour chaque paire minimale de phones nasal et oral, par exemple /a/-/ɑ̃/. (iii) la normalisation a été réalisée pour chaque paire nasal-oral et pour chaque locuteur. Pour la consonne /v/ sans correspondant nasal, la valeur a été normalisée par rapport à son ensemble.

Que ce soit avec le débit d'air brut ou normalisé, le modèle à séquences longues présente une corrélation plus forte que le modèle à séquences courtes. Nous remarquons deux observations en commun pour les deux modèles. Dans l'ensemble, les probabilités de nasalité sont les plus fortement corrélées avec les valeurs normalisées par phonème et par locuteur. Ceci montre que le débit d'air nasal est propre aux phonèmes et aux locuteurs. Ensuite, la corrélation est la plus forte pour le locuteur MT04 et cette observation est commune dans les deux modèles. Cependant, le locuteur ayant la corrélation la plus faible diffère selon la longueur d'extrait audio et les mesures de débit d'air nasal.

Débit d'air nasal	Locuteur	MT03	MT04	MT05	MT06	MT07	MT08	Tous
moyenne	Séquences longues	0,75	0,73	0,68	0,70	0,76	0,77	0,70
phonèmes		0,66	0,76	0,68	0,68	0,72	0,72	0,68
phonèmes+locuteurs		0,70	0,79	0,69	0,68	0,73	0,68	0,71
moyenne	Séquences courtes	0,61	0,65	0,59	0,59	0,46	0,69	0,55
phonèmes		0,52	0,69	0,58	0,51	0,45	0,64	0,53
phonèmes+locuteurs		0,55	0,70	0,61	0,52	0,48	0,60	0,57

TABLE 2 – Comparaison des résultats obtenus avec les modèles de classification avec le débit d'air nasal (DAN) à l'aide du coefficient de corrélation de Pearson.

5 Discussion et conclusion

Notre objectif était d'étudier la longueur des séquences pour l'extraction de vecteurs afin de faciliter une tâche de classification phonétique, en particulier celle de la nasalité. Deux longueurs ont été examinées : une séquence d'un phonème et une séquence plus étendue avec une seconde ajoutée de chaque côté du phonème. Ces deux approches ont donné des performances satisfaisantes dans la tâche proposée. Les séquences plus longues ont atteint une exactitude globale de 94,05 %, tandis que les séquences plus courtes ont obtenu 81,04 %.

Nos deux modèles ont réussi à se spécialiser à la nasalité dans la parole, mais avec un comportement variant selon les phonèmes et locuteurs. Dans la section 4.1, il a été démontré que le comportement des modèles diffère selon les phonèmes, ce phénomène peut s'expliquer par la variation des positions des articulateurs pendant la réalisation d'un phone et par le fait que le voile du palais se positionne différemment selon les voyelles (Delvaux and Metens, 2002; Amelot et al., 2008). Par exemple, dans le cas de /ɑ̃/, qui est le plus correctement identifié parmi les voyelles nasales, le voile du palais s'abaisse davantage et la position de la langue devient plus postérieure jusqu'à ce qu'elle atteigne le velum. Comme cette voyelle induit une ouverture de la bouche ainsi qu'une ouverture du port vélopharyngé, l'air peut circuler dans la cavité nasale (Delvaux and Metens, 2002; Amelot et al., 2008).

La comparaison entre les probabilités de nasalité et les données physiologiques révèle une corrélation entre le débit d'air nasal et les probabilités obtenues avec nos modèles. Cette relation de corrélation varie en fonction des phonèmes et des locuteurs. Par exemple, la corrélation est plus forte lorsque le débit d'air nasal est normalisé par phonème et par locuteur que pour le DAN brut. Le locuteur MT04 présente une corrélation particulièrement forte. Ce locuteur peut être considéré comme ayant une bonne distinction entre la production orale et nasale de la voix.

En conclusion, notre étude a mis en évidence l'utilisation de deux longueurs de séquences pour extraire des informations vectorielles dans le cadre d'une tâche spécifique liée à la nasalité. En comparant nos classifieurs avec des mesures aérodynamiques, une corrélation significative a été observée entre les débits d'air nasal et les probabilités de nasalité. Les résultats révèlent le comportement différencié des modèles selon les phonèmes et les locuteurs, avec une variabilité interlocuteur constatée. Les performances restent constantes chez les locuteurs ayant une bonne distinction entre la production orale et nasale dans la voix, ainsi que chez ceux possédant une voix distinctive. Cependant, il convient de noter que la spécification de l'entraînement sur la nasalité permet une bonne corrélation avec le débit d'air nasal, facilitant ainsi la mise en évidence de phénomènes tels que la quantité de nasalité.

Références

Angélique Amelot, Patricia Basset, Shinji Maeda, Kiyoshi Honda, and Lise Crevier-Buchman. Etude simultanée des mouvements du voile du palais et de l'ouverture du port vélopharyngé. *XXVIIè JEP*, pages 65–68, 2008.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations, October 2020. URL http://arxiv.org/abs/2006.11477. arXiv :2006.11477 [cs, eess].

Altijana Brkan. *Etude comparative des phénomènes de coarticulation nasale en anglais américain, bosnien, français, norvégien et ourdou*. PhD thesis, Université Sorbonne Paris Cité, 2018.

Anaïs Chanclu, Laurianne Georgeton, and Corinne Fredouille. PTSVOX : une base de données pour la comparaison de voix dans le cadre judiciaire. 2020.

Marilyn Y. Chen. Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, 102(4) :2360–2370, October 1997. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.419620. URL https://pubs.aip.org/jasa/article/102/4/2360/562446/Acoustic-correlates-of-English-and-French.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington, 2022. Association for Computational Linguistics. DOI : 10.18653/v1/2022.sigmorphon-1.9. URL https://aclanthology.org/2022.sigmorphon-1.9.

Jianwu Dang, Kiyoshi Honda, and Hisayoshi Suzuki. Morphological and acoustical analysis of the nasal and the paranasal cavities. *The Journal of the Acoustical Society of America*, 96(4) : 2088–2100, 1994. Publisher : Acoustical Society of America.

Véronique Delvaux and Thierry Metens. Propriétés acoustiques et articulatoires des voyelles nasales du français. 2002.

Amélie Elmerich, Angelique Amelot, Shinji Maeda, Yves Laprie, Jean Francois Papon, and Lise Crevier-Buchman. F1 and f2 measurements for french oral vowel with a new pneumotachograph mask. In *ISSP 2020-12th International Seminar on Speech Production*, 2020.

Amélie Elmerich, Jiayin Gao, Angelique Amelot, Lise Crevier-Buchman, and Shinji Maeda. Combining acoustic and aerodynamic data collection : A perceptual evaluation of acoustic distortions. In *INTERSPEECH 2023*, pages 3078–3082. ISCA, August 2023. DOI : 10.21437/Interspeech.2023-1918. URL https://www.isca-archive.org/interspeech_2023/elmerich23_interspeech.html.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. Exploring wav2vec 2.0 on speaker verification and language identification, January 2021. URL http://arxiv.org/abs/2012.06185. arXiv :2012.06185 [cs, eess].

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 139–142. Citeseer, 2006.

Erica Gold and Peter French. International practices in forensic speaker comparisons : second survey. *International Journal of Speech, Language and the Law*, 26(1) :1–20, 2019.

G Gravier, J-F Bonastre, E Geoffrois, S Galliano, K Mc Tait, and K Choukri. The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. 2004.

Séverine Guillaume, Guillaume Wisniewski, and Alexis Michaud. Fromsnippet-lects' to doculects and dialects : Leveraging neural representations of speech for placing audio signals in a language landscape. *arXiv preprint arXiv :2305.18602*, 2023.

B H Juang and L R Rabiner. Hidden Markov Models for Speech Recognition. 33(3), 1991.

Juliette Kahn. Parole de locuteur : performance et confiance en identification biométrique vocale. Avignon, 2011.

Lila Kim, Cedric Gendrot, Amélie Elmerich, Angelique Amelot, and Shinji Maeda. Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques. 2023.

Lori F Lamel, Jean-Luc Gauvain, Mazcine Eskénazi, et al. Bref, a large vocabulary spoken corpus for french1. *training*, 22(28) :50, 1991.

Shinji Maeda. Acoustic cues for vowel nasalization : A simulation study. *The Journal of the Acoustical Society of America*, 72(S1) :S102–S102, November 1982. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.2019690. URL https://pubs.aip.org/jasa/article/72/S1/S102/733010/Acoustic-cues-for-vowel-nasalization-A-simulation.

Francis Nolan. Forensic Speaker Identification and the Phonetic. *A Figure of Speech : A Festschrift for John Laver*, page 385, 2014. Publisher : Routledge.

Titouan Parcollet, Ha Nguyen, Solene Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Esteve, Mickael Rouvier, Jerome Goulian, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. LeBenchmark 2.0 : a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech, September 2023. URL http://arxiv.org/abs/2309.05472. arXiv :2309.05472 [cs, eess].

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise Analysis of a Self-supervised Speech Representation Model, December 2022. URL http://arxiv.org/abs/2107.04734. arXiv :2107.04734 [cs, eess].

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models, March 2023. URL http://arxiv.org/abs/2211.03929. arXiv :2211.03929 [cs, eess].

Ibrahim Patel and Y Srinivas Rao. Speech Recognition Using HMM with MFCC-An Analysis Using Frequency Specral Decomposion Technique. *Signal & Image Processing : An International Journal*, 1(2) :101–110, December 2010. ISSN 22293922. DOI : 10.5121/sipij.2010.1209. URL http://www.aircconline.com/sipij/V1N2/1210sipij09.pdf.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, April 2021. URL http://arxiv.org/abs/2104.03502. arXiv :2104.03502 [cs, eess].

Antoine Serrurier. Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production de nasales - Caractérisation articulatori-acoustique des mouvements du voile du palais. 2006.

Will Styler. On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, 142(4) :2469–2482, October 2017. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.5008854. URL https://pubs.aip.org/jasa/article/142/4/2469/853233/On-the-acoustical-features-of-vowel-nasality-in.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3) :201–212, March 2010. ISSN 01676393. DOI :

10.1016/j.specom.2009.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0167639309001629.

Liang-Hsuan Tseng, Yu-Kuan Fu, Heng-Jui Chang, and Hung-yi Lee. Mandarin-english code-switching speech recognition with self-supervised speech representation models. *arXiv preprint arXiv :2110.03504*, 2021.

Deuxième partie

Articles présentés en session poster

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Adaptation de modèles auto-supervisés pour la reconnaissance de phonèmes dans la parole d'enfant

Lucas Block Medin[1, 2], Lucile Gelin[1, 2], Thomas Pellegrini[2]
(1) Lalilo by Renaissance Learning, 236 rue du faubourg Saint Martin, 75010 Paris, France
(2) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
lucas.block@renaissance.com, lucile.gelin@renaissance.com,
thomas.pellegrini@irit.fr

RÉSUMÉ

La reconnaissance de parole d'enfant est un domaine de recherche encore peu développé en raison du manque de données et des difficultés caractéristiques de cette tâche. Après avoir exploré diverses architectures pour la RAP d'enfant dans de précédents travaux, nous nous attaquons dans cet article aux nouveaux modèles auto-supervisés. Nous comparons d'abord plusieurs modèles Wav2vec2, HuBERT et WavLM adaptés superficiellement à la reconnaissance de phonèmes sur parole d'enfant, et poursuivons nos expériences avec le meilleur d'entre eux, un WavLM base+. Il est ensuite adapté plus profondément en dégelant ses blocs transformer lors de l'entraînement sur parole d'enfant, ce qui améliore grandement ses performances et le fait surpasser significativement notre modèle de base, un Transformer+CTC. Enfin, nous étudions en détail les comportements de ces deux modèles en conditions réelles de notre application, et montrons que WavLM base+ est plus robuste à diverses tâches de lecture et niveaux de bruit.

ABSTRACT

Adapting self-supervised learning for phoneme recognition in child speech.

Child speech recognition is still an underdeveloped area of research due to the lack of data and the specific difficulties of this task. Having explored various architectures for child speech recognition in previous work, in this article we tackle new self-supervised models. We first compare several Wav2vec2, HuBERT and WavLM models adapted to phoneme recognition in child speech, and continue our experiments with the best of them, a WavLM base+. We then further adapt it by unfreezing its transformer blocks during fine-tuning on child speech, which greatly improves its performance and makes it significantly outperform our base model, a Transformer+CTC. Finally, we study in detail the behaviour of these two models under the real conditions of our application, and show that WavLM base+ is more robust to various reading tasks and noise levels.

MOTS-CLÉS : reconnaissance automatique de la parole ; parole d'enfant ; modèles auto-supervisés.

KEYWORDS: automatic speech recognition, child speech, self-supervised learning.

1 Introduction

Le langage oral des jeunes enfants (5-8 ans) présente des caractéristiques spécifiques liées au développement de leur appareil vocal et à leur contrôle moteur encore en développement : mécanismes articulatoires instables et variabilité spectrale intra- et inter-locuteur·rices (Lee *et al.*, 1999), fréquences fondamentales et formantiques plus élevées que celles des adultes (Mugitani & Hiroya,

2012), ou encore présence d'erreurs phonologiques (Fringi *et al.*, 2015). Ces différences morphologiques et phonologiques constituent les principales raisons des performances limitées des systèmes de reconnaissance automatique de la parole (RAP) lorsqu'ils sont confrontés aux voix d'enfants.

Les outils numériques d'assistance à la lecture ont un impact pédagogique significatif sur les enfants apprenant à lire, et plusieurs initiatives ont été développées au fil du temps (Mostow & Aist, 2001; Bolaños *et al.*, 2011; Godde *et al.*, 2017). Lalilo [1] propose un assistant de lecture destiné aux enfants de 5 à 8 ans, comprenant un exercice de lecture à voix haute qui offre un retour personnalisé grâce au système de reconnaissance automatique de phonèmes présenté dans cet article.

Des recherches antérieures sur la RAP des enfants ont montré des performances inférieures à celles observées chez les adultes (Potamianos & Narayanan, 2003; Shivakumar & Georgiou, 2020; Yeung & Alwan, 2018). Des systèmes hybrides ont démontré des améliorations en combinant données d'adulte et d'enfant (Serizel & Giuliani, 2014), ou en utilisant des techniques d'apprentissage par transfert (Shivakumar & Georgiou, 2020). Les architectures récentes dites bout-à-bout, ou *end-to-end*, ont récemment été adaptées à la RAP d'enfants, et ont atteint ou surpassé les performances des architectures hybrides(Shivakumar & Narayanan, 2021; Gelin *et al.*, 2022).

Récemment, l'apprentissage auto-supervisé (*Self-Supervised Learning*, SSL) a été introduit dans le domaine de la RAP en raison de son grand potentiel pour améliorer les tâches à faibles ressources en exploitant les connaissances préalables acquises à partir de grandes quantités de données non annotées (Mohamed *et al.*, 2022; N *et al.*, 2021). C'est le cas de la RAP d'enfant, où les données sont rares et leur annotation est complexe et coûteuse. De récentes études ont montré que le potentiel d'apprentissage à partir de données non annotées abondantes est élevé pour de la RAP d'enfants (Jain *et al.*, 2023; Fan *et al.*, 2022).

Nous étudions dans ce papier le comportement des modèles auto-supervisés à l'état de l'art sur nos données spécifiques de parole de jeunes enfants apprenant à lire. Après une rapide comparaison de plusieurs modèles (Wav2vec2 (Baevski *et al.*, 2020), HuBERT (Hsu *et al.*, 2021) et WavLM (Chen *et al.*, 2022)) pour la reconnaissance de phonèmes dans la parole d'enfant, nous adapterons plus en profondeur le meilleur d'entre eux, et le comparerons avec notre système actuel, un Transformer+CTC. Nous analyserons en détail les performances de chacun sur différents types de contenu et dans différents niveaux de bruit.

2 Jeux de données

Pour compenser le manque de données disponibles de parole d'enfant, nous utilisons de la parole d'adulte pour entraîner des modèles sources, que nous adaptons ensuite avec de la parole d'enfant.

2.1 Parole d'adulte

Nous utilisons pour notre modèle de base une version du corpus Common Voice [2] français qui contient environ 150 heures de parole lue. Les modèles auto-supervisés sont préentraînés sur de la parole d'adulte non annotée, provenant des corpus suivants :

- Librispeech : 960 heures, annotées, parole lue, anglais (Panayotov *et al.*, 2015) ;
- Libri-Light : 60 000 heures, non annotées, parole lue, anglais (Kahn *et al.*, 2020) ;
- VoxPopuli : 24 000 heures, non annotées, parole multilingue (23 langues) (Wang *et al.*, 2021) ;
- GigaSpeech : 10 000 heures, non annotées, parole lue/spontanée, anglais (Chen *et al.*, 2021).

1. https://www.lalilo.com/
2. Corpus disponible sur : https://voice.mozilla.org/fr

2.2 Parole d'enfant : Lalilo

Le corpus Lalilo contient des enregistrements d'enfants du CP au CE2, âgés de 5 à 8 ans, lisant oralement divers types de contenu. Les données sont transcrites manuellement au niveau du mot et chaque mot est étiqueté « correct » ou « incorrect ». Les mots corrects sont phonétisés automatiquement, tandis que les mots incorrects sont transcrits manuellement au niveau du phonème. L'annotation est faite par deux annotateur·rice·s, et l'enregistrement est écarté en cas de désaccord.

Lors de l'apprentissage de la lecture, les élèves effectuent diverses tâches de lecture, de difficulté croissante et appelant à utiliser différents processus cognitifs. Dans la plateforme Lalilo, nous proposons principalement quatre types de contenu, plus ou moins difficiles : des mots isolés, des phrases courtes, des listes de mots et des listes de pseudo-mots. Les enregistrements sont principalement recueillis dans le cadre de l'exercice de lecture orale de la plateforme Lalilo, qui est le plus souvent utilisé dans des salles de classe sous surveillance réduite : ils contiennent des niveaux variables de bruit de brouhaha. On calcule le niveau de bruit à l'aide d'un rapport signal à bruit (RSB).

Les ensembles d'entraînement et de validation contiennent respectivement 13 heures et 25 minutes de données. Ayant été conçus avant l'ajout de nouveaux types de contenu, ces ensembles ne contiennent que des mots isolés et des phrases. De plus, ils sont uniquement composés d'énoncés correctement prononcés. La transcription correspond au texte demandé à l'élève, phonétisé automatiquement avec un dictionnaire de prononciation. Les ensembles d'entraînement et de validation ont respectivement des RSB moyens de 21,0 dB et 20,6 dB avec des déviations standards de 13,0 dB et 12,6 dB. L'ensemble de test est composé de 3 heures d'énoncés, avec environ 25% des mots qui contiennent une erreur de lecture. Nous utilisons ici les quatre types de contenu, divisant l'ensemble de test en sous-catégories : mots isolés (M, 51 min), phrases (P, 29 min), listes de mots (LM, 56 min) et listes de pseudo-mots (LPM, 50 min). Les valeurs de RSB du test sont identiques à l'ensemble de validation.

3 Description des systèmes

Cette section présente les différents systèmes que nous étudierons dans ce papier. Nous entraînons nos systèmes à la reconnaissance de phonèmes, et non de mots, afin de pouvoir détecter plus efficacement les erreurs de lecture. Tous nos systèmes sont entraînés avec SpeechBrain (Ravanelli *et al.*, 2021).

3.1 Modèle de base : Transformer+CTC

Proposé par (Vaswani *et al.*, 2017) et adapté à la reconnaissance automatique de la parole (RAP) par (Dong *et al.*, 2018), le modèle Transformer suit une architecture *end-to-end* encodeur-décodeur séquence à séquence (*seq2seq*). Il se fonde uniquement sur des mécanismes d'attentions, abandonnant les réseaux de neurones récurrents habituels des systèmes *seq2seq*. La récurrence, essentielle pour extraire l'information de position des trames audio, est remplacée par des encodages positionnels, des modules d'auto-attention multi-tête et d'attention croisée, et des réseaux de neurones à propagation avant tenant compte de la position. Le modèle Transformer+CTC est complété par une fonction CTC (*Connectionnist Temporal Classification*) en sortie de l'encodeur, qui permet d'améliorer les performances grâce à un entraînement multi-objectif (entropie croisée et CTC) et un décodage joint attention/CTC (Watanabe *et al.*, 2017; Karita *et al.*, 2019a).

Le choix de cette architecture repose sur ses excellentes performances dans des tâches de reconnaissance de parole d'adulte (Karita *et al.*, 2019b), que nous avons confirmées sur la parole d'enfants apprenant·es lecteur·rices dans (Gelin *et al.*, 2021, 2022). Notre modèle suit la même architecture que dans les travaux passés cités ci-dessus, mais une nouvelle version a été entraînée avec SpeechBrain

pour faciliter la comparaison avec les autres modèles. Il contient 14,3 millions de paramètres.

Notre modèle Transformer+CTC est entraîné en deux étapes : un premier modèle source est entraîné sur la parole d'adulte provenant du corpus Common Voice, puis ce modèle est adapté par apprentissage par transfert avec le jeu de parole d'enfant Lalilo. Toutes les couches sont ré-entraînées lors de cette seconde étape, comme le conseillent (Shivakumar & Georgiou, 2020) pour de très jeunes enfants.

3.2 Modèles auto-supervisés pré-entraînés

Depuis l'introduction de Wav2vec (Schneider *et al.*, 2019), les modèles auto-supervisés se sont imposés dans le domaine de la RAP. Grâce à l'utilisation de données non annotées pour extraire des représentations latentes, ces modèles peuvent atteindre des résultats à l'état de l'art avec jusqu'à 100 fois moins de données annotées que d'autres modèles supervisés. Ces résultats sont notamment remarquables dans le contexte de la reconnaissance de parole d'enfants, où les modèles de l'architecture Wav2Vec2 (Baevski *et al.*, 2020) atteignent des performances similaires à celles des modèles supervisés de pointe (Jain *et al.*, 2023). Nous avons sélectionné pour notre étude les modèles auto-supervisés préentraînés pour la RAP les plus répandus : Wav2vec2, HuBERT et WavLM.

3.2.1 Wav2vec2

Wav2vec2 (Baevski *et al.*, 2020) est une architecture auto-supervisée dite *end-to-end*, fondée sur des réseaux de neurones convolutifs et transformer. L'architecture peut se diviser en trois grandes parties : un encodeur, un réseau contextuel transformer, et un module de quantification.

L'encodeur se compose de sept blocs contenant une convolution temporelle, suivis d'une couche de normalisation des activations *(Layer Norm)* et d'une fonction d'activation GELU. Le réseau contextuel suit l'architecture Transformer. Il remplace cependant l'encodage positionnel absolu par une couche de convolution, qui agit comme un encodage positionnel relatif. Cet encodage passe par une fonction GELU, est ensuite concaténé aux sorties de l'encodeur, et le tout subit une normalisation *(Layer Norm)*. Le réseau est composé de 12 de ces blocs, de dimension 768, avec une dimension interne de 3082, et 8 têtes d'attention par bloc. Enfin, le module de quantification récupère également la sortie de l'encodeur et la transforme en un ensemble de représentations discrètes via une « quantification » *(product quantization)*.

Le modèle Wav2Vec2 est pré-entraîné pour une tâche de prédiction masquée : il vise à prédire la représentation audio latente quantifiée correcte en contexte d'une utterance malgré l'application d'un masque sur une partie des trames audio. L'objectif global de l'entraînement est de minimiser les fonctions de perte de contraste *(contrastive loss)* et de perte de diversité *(diversity loss)*.

Nous utilisons un modèle acoustique pré-entraîné Wav2vec2.0 Base [3]. Le modèle est entraîné en utilisant le jeu de données LibriSpeech standard 960h (Panayotov *et al.*, 2015).

3.2.2 HuBERT

Le modèle HuBERT (Hsu *et al.*, 2021) reprend l'architecture de Wav2vec2.0, mais remplace le module de quantification par une quantification *K-Means*.

Ce changement implique trois différences fondamentales :

— La représentation discrète est obtenue par la découverte d'unités cachées (*hidden units*), en attribuant à chaque extrait audio un cluster via un algorithme K-Means ;

3. https://huggingface.co/facebook/wav2vec2-base-960h

— L'extraction des représentations est itérative, utilisant d'abord les résultats d'un MFCC, puis les embeddings des couches intermédiaires du modèle pré-entraîné ;
— Les fonctions de perte de contraste et de perte de diversité sont remplacées par une perte d'entropie croisée, ce qui simplifie l'entraînement.

Nous utilisons dans cette étude un modèle acoustique pré-entraîné HuBERT Base [4], entraîné également sur Librispeech standard 960h.

3.2.3 WavLM

L'architecture WavLM (Chen *et al.*, 2022) reprend celle de HuBERT en introduisant un biais de position relative à porte *(gated relative position bias)* dans les mécanismes d'attention. Au lieu de se fier uniquement aux positions absolues des vecteurs clé et requête, le modèle prend ainsi en compte les positions relatives entre ces vecteurs lors du calcul des scores d'attention.

Le modèle WavLM comporte également des modifications dans la phase de pré-entraînement. La tâche de prédiction masquée est remplacée par une tâche de débruitage et prédiction masquée. Ce procédé, qui cherche à rendre le modèle plus robuste, consiste à simuler des entrées bruitées ou de la parole superposée, puis à prédire des pseudo-étiquettes de l'audio original sur la région masquée.

Nous allons étudier dans ce travail deux modèles WavLM :
— Un modèle WavLM Base [5], entraîné avec les mêmes données que les modèles précédents ;
— Un modèle WavLM Base+ [6], qui possède la même architecture que le WavLM Base, mais est entraîné sur un corpus beaucoup plus vaste composé des données LibriLight, GigaSpeech et VoxPopuli, pour un total d'environ 94 000 heures. Ce corpus étendu permet d'améliorer les performances et la robustesse du modèle WavLM tout en gardant un modèle de taille raisonnable (Chen *et al.*, 2022).

4 Adaptation et évaluation des modèles SSL pour la transcription phonémique de parole d'enfant

Nous avons décidé de nous concentrer sur des modèles pré-entraînés SSL de format *Base* plutôt que *Large*. D'une part, la capacité de calcul nécessaire à l'entraînement et au déploiement des modèles *Base* est bien inférieure de par leur plus petit nombre de paramètres (95M contre 317M). D'autre part, nous pouvons constater que, dans le cas de la parole d'enfant, l'amélioration de performance est faible en contrepartie d'une augmentation significative du nombre de paramètres (Jain *et al.*, 2023). Les modèles pré-entraînés en français étaient peu documentés et très hétérogènes en termes de données utilisées, ce qui rendait la comparaison complexe, et nous a poussé à utiliser des modèles anglais.

Pour adapter les systèmes SSL à notre tâche, nous faisons passer les sorties du réseau Transformer (de taille 768) dans une projection linéaire pour faire de la classification de phonèmes. Cette couche comporte 35 classes représentant les phonèmes français et le phonème « vide ». Le modèle est entraîné de façon supervisée avec les données Lalilo, avec pour objectif de minimiser la fonction de perte CTC (*Connectionist Temporal Classification*). Le taux d'erreur phonème (*Phoneme Error Rate*, PER) est utilisé pour mesurer la performance de nos systèmes sur cette tâche.

Dans cette section, nous comparons les différents modèles présentés précédemment et adaptés à la

4. https://huggingface.co/facebook/hubert-base-960h
5. https://huggingface.co/facebook/wavlm-base
6. https://huggingface.co/facebook/wavlm-base-plus

transcription phonétique de parole d'enfant. Nous explorons également deux méthodes d'adaptation en gelant une ou plusieurs parties des modèles, puis étudierons en détail les performances des systèmes en fonction des caractéristiques spécifiques de notre tâche.

4.1 Comparaison des modèles auto-supervisés

Notre première expérience vise à comparer les différents modèles SSL pour notre application. Nous adaptons les modèles en réentraînant uniquement la couche CTC de classification de phonèmes. Les modèles sont entraînés sur maximum 30 epochs, et la sauvegarde obtenant le meilleur PER sur l'ensemble de validation Lalilo est conservée. Pour cette expérience préliminaire, le décodage utilisé est une recherche gloutonne (*greedy search*).

Modèle	**PER**
Wav2vec 2.0	62,9
HuBERT	46,3
WavLM base	46,8
WavLM base+	**41,5**

TABLE 1 – PER (%) des différents modèles entraînés avec *fine-tuning* (avec le corpus Lalilo) de la dernière couche CTC et décodés par *greedy search*

On observe que les performances des modèles HuBERT et WavLM surpassent largement celle du modèle Wav2vec. A quantité de données égale, la différence entre HuBERT et WavLM base n'est pas significative. En revanche, le PER obtenu par le modèle WavLM base+, qui a le même nombre de paramètres mais est entraîné sur 100 fois plus de données, est significativement meilleur. Le reste de l'étude sera en conséquence concentré sur celui-ci.

4.2 Adaptation du modèle WavLM base+

Nous souhaitons aller plus loin dans l'adaptation du modèle WavLM pour améliorer ses performances sur notre application. Plutôt que d'entraîner uniquement la couche CTC avec la parole d'enfant, nous adaptons également une partie du modèle pré-entraîné. Nous suivons pour cela ce qui est fait dans (Jain *et al.*, 2023) pour l'adaptation d'un modèle Wav2vec2 à de la parole d'enfant : pendant les 1000 premières itérations, seule la dernière couche de classification CTC est entraînée, puis le bloc Transformer est également entraîné. L'encodeur CNN, en revanche, reste gelé. Le taux d'apprentissage est fixé à 5e-4 et la taille de batch à 128, suivant les recommandations de (Chen *et al.*, 2022).

Le tableau 2 affiche les valeurs PER obtenues par le modèle de base Transformer+CTC, ainsi que par deux modèles WavLM : celui de la section 4.1 où seule la couche CTC a été adaptée, dit "gelé", et celui adapté plus en profondeur, dit "dégelé". Ici et dans le reste de l'article, le décodage utilisé pour tous les modèles est une recherche par faisceaux (*beam search*) avec une taille de faisceaux de 10.

Modèle	**# params entraînables (# total)**	**PER**
Transformer+CTC	14 M (14 M)	40,5
WavLM base+ "gelé"	28 k (95 M)	39,2
WavLM base+ "dégelé"	90 M (95 M)	**26,1**

TABLE 2 – PER (%) des modèles Transformer+CTC et WavLM base+, décodés par *beam search*

On observe tout d'abord que le modèle WavLM base+ gelé atteint une performance légèrement meilleure que le modèle de base Transformer+CTC, alors que seule sa couche de classification de phonème (moins de 1‰ des poids du modèle) a été entraînée avec la parole d'enfant. Cela montre que les représentations auto-supervisées du modèle adulte WavLM base+, bien que n'ayant pas vu de parole d'enfant lors de leur apprentissage, sont génériques et aisément adaptables à différents types de parole. Ces résultats doivent cependant être nuancés : les deux modèles obtiennent des résultats comparables mais le Transformer+CTC contient près de 7 fois moins de paramètres. En dégelant le bloc Transformer de WavLM base+, on obtient un PER de 26,1%, soit une réduction relative de 33,4% par rapport au modèle gelé. Ce résultat montre que les représentations WavLM peuvent néanmoins être adaptées pour mieux correspondre à une parole spécifique, et que cette adaptation est efficace malgré une petite quantité de données d'adaptation (13 heures).

4.3 Discussion

Dans la section précédente, nous avons vu que les modèles Transformer+CTC et WavLM base+ affichent une différence de PER de 14.4%. Dans cette section, nous souhaitons savoir si cette différence est répartie de façon égale en fonction des différentes tâches de lecture présentées aux systèmes, ainsi qu'en fonction des différentes conditions de bruit en salle de classe.

4.3.1 Application aux tâches de lecture de Lalilo

Nous proposons maintenant d'explorer la performance des systèmes en fonction des différentes tâches de lecture proposées aux élèves, détaillées dans la section 2.2. Les valeurs de PER obtenues sont visibles dans la première partie du tableau 3. On observe aisément que la différence de PER entre les deux modèles dépend effectivement de la tâche de lecture.

Modèle	**Tâche de lecture**				**Niveau de bruit**		
	P	M	LM	LPM	faible	moyen	fort
Transformer+CTC	16,5	34,0	46,5	59,0	14,6	24,6	40,6
WavLM base+ "dégelé"	16,4	25,5	28,3	32,9	13,4	21,7	31,6

TABLE 3 – PER (%) des modèles Transformer+CTC et WavLM base+, décodés par *beam search*, en fonction de la tâche de lecture (P = phrase, M = mot, LM = liste de mots, LPM = liste de pseudo-mots) et/ou du niveau de bruit.

Les phrases courtes représentent la tâche la plus facile pour la RAP, avec un contexte suffisant mais pas trop grand, et la présence de mots de liaisons couramment vus en apprentissage. C'est de plus une tâche classique pour les corpus de parole d'adulte. Sur ce sous-ensemble (P dans le tableau), il n'y a pas de différence significative entre les deux modèles. Les deux modèles sont adaptés avec la même quantité de parole d'enfant. Sur cette tâche facile et connue, l'apprentissage supervisé d'un petit modèle (14 M de paramètres) avec 150 heures de parole d'adulte est donc aussi efficace que l'apprentissage non supervisé d'un gros modèle (95 M) sur près de 100 000 heures de parole.

Sur la reconnaissance de phonèmes dans des mots isolés (M dans le tableau), on observe que le WavLM est significativement meilleur (-8,5% absolu). Les mots pouvant contenir aussi peu que 2 phonèmes, la difficulté du Transformer+CTC peut s'expliquer par le manque de contexte sur lequel les mécanismes d'attention peuvent s'appuyer. Ce phénomène a notamment été observé dans (Chan *et al.*, 2016), où la performance du modèle se dégrade significativement lorsque l'énoncé ne contient qu'un seul mot. Le WavLM contient également un bloc Transformer qui est affecté par ce phénomène, mais il est probablement compensé par l'utilisation d'un encodeur CNN, dont les convolutions permettent de tirer le meilleur parti du manque de contexte.

Les listes de mots (LM) et de pseudo-mots (LPM) n'ont pas été vues pendant l'apprentissage, ce qui en fait des tâches légèrement hors domaine. C'est d'autant plus le cas pour les listes de pseudo-mots car les pseudo-mots n'existent pas et n'ont jamais été vus dans aucun corpus de parole d'adulte ou d'enfant. Sur ces tâches, on observe que le modèle WavLM est bien meilleur que le Transformer+CTC. On observe également que plus la tâche est hors domaine, plus la réduction relative de PER apportée par le WavLM s'accentue : -39% sur les listes de mots, -44% sur les listes de pseudo-mots. On peut déduire de ces observations que le modèle WavLM possède une meilleure capacité de généralisation, qui est sûrement liée à la quantité de données rencontrées, mais également à l'apprentissage auto-supervisé qui est moins contraint et crée ainsi des représentations plus génériques.

4.3.2 Robustesse au bruit de salle de classe

Nous souhaitons également étudier le comportement de nos deux systèmes en conditions réelles de salle de classe, c'est à dire avec différents niveaux de bruit. Nous divisons notre ensemble de test en trois niveaux de bruit :

— Faible : enregistrements avec un RSB supérieur à 25 dB ;
— Moyen : enregistrements avec un RSB compris entre 10 et 25 dB ;
— Fort : enregistrements avec un RSB inférieur à 10 dB.

La seconde partie du tableau 3 affiche les résultats sur ces sous-catégories. On observe évidemment que, pour les deux modèles, la performance se dégrade fortement avec l'augmentation du niveau de bruit. Il est intéressant de noter que la différence de PER entre les deux modèles augmente avec le niveau de bruit : 1,2% sur du bruit faible, 2.9% sur du bruit moyen et 9.0% sur du bruit fort. Le modèle WavLM témoigne ainsi d'une plus grande robustesse au bruit.

Pour confirmer cette observation, nous regardons les performances des modèles en fonction du bruit sur le sous-ensemble de test P, sur lequel les modèles obtiennent un PER comparable.

— Transformer+CTC : 10,6 (faible) - 17,1 (moyen) - 30,0 (fort)
— WavLM base+ : 12,5 (faible) - 17,1 (moyen) - 26,8 (fort)

On voit ainsi que le WavLM est bien plus robuste dans des conditions de bruit fort, au prix d'une moins bonne performance lorsqu'il n'y a qu'un bruit faible. Ces résultats sont en accord avec les changements introduits dans le pré-entraînement de WavLM, ayant pour objectif de rendre le modèle plus robuste à des conditions acoustiques difficiles (Chen *et al.*, 2022).

5 Conclusion

Les systèmes capables de transcrire avec précision la parole d'enfant sont encore rares, notamment en français, en raison d'un manque de données disponibles et d'une difficulté accrue sur ce type de parole. Nous explorons ici l'adaptation des nouveaux modèles auto-supervisés à la reconnaissance de phonèmes sur de la parole de jeunes enfants. Dans un premier temps, nous sélectionnons trois modèles (Wav2vec2, HuBERT et WavLM, en version *base*) et adaptons une couche CTC de classification de phonèmes avec notre corpus de parole d'enfant. Nous observons que les modèles HuBERT et WavLM surpassent Wav2vec2, et que le modèle WavLM base+, entraîné sur 100 fois plus de données tout en conservant le même nombre de paramètres, est significativement plus performant que les autres. Dans un second temps, nous adaptons le modèle WavLM base+ plus en profondeur en dégelant les blocs Transformer du modèle, ce qui améliore ses performances de 33,4% relatifs. Nous montrons qu'il obtient une précision largement meilleure que celle obtenue par notre modèle de base, un Transformer+CTC. Enfin, nous analysons les comportements de nos modèles face à différentes tâches de lecture et conditions de bruit, et montrons que le WavLM base+ est plus efficace sur des enregistrements très courts, généralise mieux à des contenus non vus en apprentissage, et est plus robuste à un bruit de salle de classe.

Références

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations.

BOLAÑOS D., COLE R., WARD W., BORTS E. & SVIRSKY E. (2011). FLORA : Fluent oral reading assessment of children's speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, **7**(4), 16. DOI : 10.1145/1998384.1998390.

CHAN W., JAITLY N., LE Q. & VINYALS O. (2016). Listen, Attend and Spell : A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4960–4964. DOI : 10.1109/ICASSP.2016.7472621.

CHEN G., CHAI S., WANG G., DU J., ZHANG W.-Q., WENG C., SU D., POVEY D., TRMAL J., ZHANG J., JIN M., KHUDANPUR S., WATANABE S., ZHAO S., ZOU W., LI X., YAO X., WANG Y., WANG Y., YOU Z. & YAN Z. (2021). Gigaspeech : An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., ZENG M., YU X. & WEI F. (2022). Wavlm : Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**, 1–14. DOI : 10.1109/JSTSP.2022.3188113.

DONG L., XU S. & XU B. (2018). Speech-transformer : A no-recurrence sequence-to-sequence model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5884–5888. DOI : 10.1109/ICASSP.2018.8462506.

FAN R., ZHU Y., WANG J. & ALWAN A. (2022). Towards better domain adaptation for self-supervised models : A case study of child asr. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1242–1252. DOI : 10.1109/JSTSP.2022.3200910.

FRINGI E., LEHMAN J. F. & RUSSELL M. J. (2015). Evidence of phonological processes in automatic recognition of children's speech. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden*, p. 1621–1624.

GELIN L., DANIEL M., PINQUIER J. & PELLEGRINI T. (2021). End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, **134**, 71–84. DOI : https ://doi.org/10.1016/j.specom.2021.08.003.

GELIN L., PELLEGRINI T., PINQUIER J. & DANIEL M. (2022). Améliorations d'un système Transformer de reconnaissance de phonèmes appliqué à la parole d'enfants apprenants lecteurs. In *34èmes Journées d'Études sur la Parole - Parole, Geste, Musique : des unités à leur organisation (JEP 2022)*, volume Session Posters n° 2, p. à paraître, Noirmoutier, France : Association Francophone de la Communication Parlée. HAL : hal-03898401.

GODDE E., BAILLY G., ESCUDERO D., BOSSE M.-L. & ESTELLE G. (2017). Evaluation of reading performance of primary school children : Objective measurements vs. subjective ratings. In *Proc. of the International Workshop on Child Computer Interaction (WOCCI)*, p. 23–27. DOI : 10.21437/WOCCI.2017-4.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units.

JAIN R., BARCOVSCHI A., YIWERE M. Y., BIGIOI D., CORCORAN P. & CUCU H. (2023). A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, **11**, 46938–46948. DOI : 10.1109/ACCESS.2023.3275106.

KAHN J., RIVIERE M., ZHENG W., KHARITONOV E., XU Q., MAZARE P., KARADAYI J., LIPTCHINSKY V., COLLOBERT R., FUEGEN C., LIKHOMANENKO T., SYNNAEVE G., JOULIN A., MOHAMED A. & DUPOUX E. (2020). Libri-light : A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE. DOI : 10.1109/icassp40776.2020.9052942.

KARITA S., SOPLIN N. E. Y., WATANABE S., DELCROIX M., OGAWA A. & NAKATANI T. (2019a). Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, p. 1408–1412. DOI : 10.21437/Interspeech.2019-1938.

KARITA S., WANG X., WATANABE S., YOSHIMURA T., ZHANG W., CHEN N., HAYASHI T., HORI T., INAGUMA H., JIANG Z., SOMEKI M., SOPLIN N. E. Y. & YAMAMOTO R. (2019b). A Comparative Study on Transformer vs RNN in Speech Applications. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (April 2020), 449–456. DOI : 10.1109/ASRU46091.2019.9003750.

LEE S., POTAMIANOS A. & NARAYANAN S. S. Y. (1999). Acoustics of children's speech : developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, **105**(3), 1455–1468. DOI : 10.1121/1.426686.

MOHAMED A., LEE H.-Y., BORGHOLT L., HAVTORN J. D., EDIN J., IGEL C., KIRCHHOFF K., LI S.-W., LIVESCU K., MAALØE L. *et al.* (2022). Self-supervised speech representation learning : A review. *IEEE Journal of Selected Topics in Signal Processing*.

MOSTOW J. & AIST G. (2001). Evaluating tutors that listen : An overview of Project LISTEN. In *Smart machines in education : The coming revolution in educational technology.*, p. 169–234. The MIT Press. DOI : 10.5555/570950.570957.

MUGITANI R. & HIROYA S. (2012). Development of vocal tract and acoustic features in children. *The Journal of the Acoustical Society of Japan*, **68**(5), 234–240. DOI : 10.1250/ast.33.215.

N K. D., WANG P. & BOZZA B. (2021). Using Large Self-Supervised Models for Low-Resource Speech Recognition. In *Proc. Interspeech 2021*, p. 2436–2440. DOI : 10.21437/Interspeech.2021-631.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

POTAMIANOS A. & NARAYANAN S. (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, **11**(November 2003), 603–616. DOI : 10.1109/TSA.2003.818026.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

SCHNEIDER S., BAEVSKI A., COLLOBERT R. & AULI M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, p. 3465–3469. DOI : 10.21437/Interspeech.2019-1873.

SERIZEL R. & GIULIANI D. (2014). Deep neural network adaptation for children's and adults' speech recognition. In *Proc. of the Italian Computational Linguistics Conference (CLiC-it)*, p. 137–140. HAL : hal-01393975.

SHIVAKUMAR P. G. & GEORGIOU P. (2020). Transfer learning from adult to children for speech recognition : Evaluation, analysis and recommendations. *Computer Speech & Language*, **63**, 101077. DOI : 10.1016/j.csl.2020.101077.

SHIVAKUMAR P. G. & NARAYANAN S. (2021). End-to-end neural systems for automatic children speech recognition : An empirical study. *ArXiv preprint :2102.09918*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER U. & POLOSUKHIN I. (2017). Attention is all you need. In *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, p. 6000–6010, Red Hook, NY, USA : Curran Associates Inc.

WANG C., RIVIÈRE M., LEE A., WU A., TALNIKAR C., HAZIZA D., WILLIAMSON M., PINO J. & DUPOUX E. (2021). Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation.

WATANABE S., HORI T., KIM S., HERSHEY J. R. & HAYASHI T. (2017). Hybrid CTC/Attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, **11**(8), 1240–1253. DOI : 10.1109/JSTSP.2017.2763455.

YEUNG G. & ALWAN A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad*, p. 1661–1665. DOI : 10.21437/Interspeech.2018-2297.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Allongement vocalique en italien L2 et en français L2 : une marque de focalisation ?

Bianca Maria De Paolis[1, 2]
(1) LFSAG - Laboratorio di Fonetica Sperimentale "Arturo Genre", Università di Torino, Italie
(2) UMR 7023 - Structures Formelles du Langage, CNRS/Université Paris 8, France
biancamaria.depaolis@unito.it

RÉSUMÉ

Notre étude explore le rôle de la durée vocalique comme indice de focalisation, à la fois en italien et en français, tant pour les locuteurs natifs que pour les apprenants L2. Nous visons à décrire l'influence potentielle de la L1 sur la L2 concernant cet indice. L'analyse porte sur la parole élicitée de 60 participants, répartis en quatre groupes : 15 italophones natifs, 15 francophones natifs, 15 apprenants francophones d'italien L2 et 15 apprenants italophones de français L2. Les locuteurs ont produit le même constituant cible en quatre conditions informationnelles : *background*, focalisation large, focalisation étroite identificative, focalisation étroite corrective. Les résultats montrent une influence du contexte informationnel sur la durée des voyelles accentuées chez les natifs italophones, mais pas chez les natifs français. Cette divergence se reflète chez les apprenants : les apprenants italophones de français ajustent la durée des voyelles accentuées selon la condition informationnelle, tandis que les francophones apprenant l'italien ne le font pas. Nous discutons ces résultats en lien avec d'autres marqueurs prosodiques et syntaxiques de focus, en tenant compte des différences typologiques entre l'italien et le français et des théories sur l'acquisition de la prosodie en L2.

ABSTRACT

Vowel lengthening in L2 Italian and L2 French : a cue for focus marking ?

Our study investigates the role of vowel duration as a cue for focus marking in both L1 and L2 Italian and French. We aim to compare our data to highlight potential influences of the native language on L2 productions in the use of this cue. The analysis involves task-elicited speech from 60 participants : 15 native Italian speakers, 15 native French speakers, 15 French learners of Italian (L2), and 15 Italian learners of French (L2). Participants produced the same target constituent under four information-structural conditions : background, broad focus, identification focus, and correction focus. Results reveal that the information-structural function significantly influences stressed vowel duration in native Italian, with identification-focus and correction-focus constituents bearing longer duration than background and broad focus. However, the same pattern does not hold in native French. Crucially, this distinction is mirrored in the production of non-native speakers. Italian learners of L2 French, in fact, modulate duration based on the informational role of the constituent ; in contrast, French learners of Italian L2 do not. We discuss these findings in relation to previous findings on other prosodic and syntactic markers of focus. Results are commented in light of typological differences in discourse-prominence marking and theories of L2 prosody acquisition.

MOTS-CLÉS : focus, durée vocalique, parole non-native, acquisition prosodie L2, français, italien.

KEYWORDS: focus, vowel duration, non-native speech, L2 prosody acquisition, French, Italian.

1 Introduction

1.1 Focus, contraste et durée des voyelles

Les catégories informationnelles telles que le *background*, le focus et le contraste peuvent être exprimées à travers divers moyens linguistiques, dont beaucoup sont réductibles à la catégorie générale de "prosodie" (Arvaniti, 2020). Parmi les traits prosodiques, la durée des segments vocaliques a été décrite comme un indice significatif dans l'expression de la structure informationnelle, avec des études explorant largement son utilisation et son impact dans différentes langues (Maekawa, 1997; de Jong & Zawaydeh, 2002; de Jong, 2004; Vander Klok *et al.*, 2018). Les études portant sur la variation de la durée induite par le focus ne sont pas abondantes pour le français et l'italien. En fait, la majorité des travaux menés sur ces deux langues traitent de l'interaction entre le focus et les mouvements tonals, considérant l'allongement des voyelles comme une sorte d'effet secondaire (Farnetani & Zmarich, 1997; D'imperio, 2002; Avesani & Vayra, 2003). Les résultats de ces études suggèrent, en tout cas, que le même mécanisme observé dans d'autres langues s'applique au français et à l'italien : à mesure que la saillance ou le contraste d'un constituant au niveau discursif augmentent, la durée de sa voyelle métriquement forte augmente. En termes plus simples, les constituants en *background* dans une énonciation tendent à avoir des voyelles toniques plus courtes, tandis que les constituants focalisés présentent des durées plus longues, les *foci* contrastifs montrant l'allongement le plus significatif. En raison des différences phonologiques entre l'italien et le français — en particulier la présence ou l'absence d'un accent lexical — le phénomène d'allongement des voyelles n'est pas ancré de la même manière dans les deux langues. En italien, le phénomène est associé au noyau de la syllabe portant l'accent lexical, tandis qu'en français, l'allongement affecte la dernière syllabe des unités en focus, c'est-à-dire celle précédant une frontière prosodique (Michelas & German, 2020).

1.2 Implications pour l'acquisition d'une langue seconde

Cette convergence-divergence du même phénomène, marquée par des ancrages différents, peut avoir des conséquences intéressantes pour les apprenants L2. Des études menées en situations de contacte linguistique proches à la nôtre (Avesani *et al.*, 2015; Gabriel & Kireva, 2014) ont démontré que la gestion de l'accent, de l'intonation, de la durée vocalique en lien avec l'expression de la structure informationnelle s'avère difficile pour les apprenants L2. Cependant, aucune étude n'a exploré ces phénomènes avec cette combinaison exacte de langues, l'italien en tant que L1 et le français en tant que L2, et *vice versa*. L'objectif de notre travail est donc de mettre en lumière ces interactions, avec une méthodologie et un corpus créés *ad hoc*. Les analyses préliminaires effectuées sur ce même corpus que nous utiliserons ici ont montré que les locuteurs français s'appuient davantage sur les mouvements de fréquence fondamentale (f0) que les locuteurs italiens pour marquer la focalisation. Cette tendance des locuteurs natifs se reflète comme une influence de la L1 dans les groupes d'apprenants (De Paolis *et al.*, 2022, in press) : les apprenants italophones de français utilisent moins l'intonation par rapport aux locuteurs de leur langue cible, et les apprenants francophones d'italien L2 se comportent de manière opposée. Avec cette étude, nous visons à intégrer à cela des informations sur la durée, pour tester si cet autre indice prosodique compense la disparité d'usage de l'intonation. De plus, nous voulons explorer l'influence de la L1 sur le plan fonctionnel : dans les études précédentes, en fait, on a pu observer que les apprenants L2 gardent les memes tendances de leur L1 dans la marque différentielle de deux sous-types de focalisation, identification et la correction (De Paolis *et al.*, 2022, in press).

2 Méthodologie

2.1 Échantillon

Pour examiner ces phénomènes d'influence interlinguistique nous avons opté pour un design inter-individuel, multi-groupes et croisé. Nous avons donc recruté un total de 60 participants, avec deux groupes d'apprenants L2 et deux groupes de locuteurs natifs : locuteurs italophones apprenants de français L2 (FRL2), locuteurs francophones apprenants l'italien L2 (ITL2), locuteurs natifs français sans compétences en italien (FRL1) et locuteurs natifs italiens sans compétence en français (ITL1). Les groupes L2 sont composés d'adultes vivant dans le pays étranger cible et ne suivant aucun cours de langue étrangère [1]. L'échantillon est équilibré en termes de genre, et la tranche d'âge (19-40 ans) est homogène dans les quatre groupes. Une attention particulière a été portée à la circonscription des zones d'origine des locuteurs, minimisant l'impact de la variation diatopique. Nos points d'investigation sont la région de Turin (Piémont) pour les groupes ITL1 et ITL2, et Paris et 'Île-de-France pour les groupes FRL1 et FRL2. Les niveaux de compétence en L2 ont été évalués au moyen de trois tests complémentaires : auto-évaluation, test lacunaire écrit (Tremblay & Garrison, 2010; Vedder, 2008), évaluation des productions orales par des enseignants L2 conformément au CECR (de l'Europe, 2020) [2]. Le tableau 1 donne un aperçu de l'échantillon.

TABLE 1 – Échantillon de l'étude.

Groupe	L1/L2	N. participants	Age (moyenne)	Sexe
ITL1	Italien/ -	N=15	25,6	M=3
FRL1	Français/-	N=15	27,5	M=4
ITL2	Français/Italien	N=15	27,4	M=7
FRL2	Italien/ Français	N=15	32,5	M=8

2.2 Protocole de collecte des données

Les études expérimentales analysant les marques de focus font face à des demandes contradictoires : le besoin de matériel de parole ayant une valeur conversationnelle entre en conflit avec les exigences de l'analyse phonétique, qui exige des unités hautement contrôlées et comparables. Pour tenter de concilier ces besoins, nous avons opté pour une tâche de récit d'images, inspirée par (Gabriel, 2010) et (Gabriel & Grünke, 2018) et adaptée et traduite pour répondre à nos besoins, utilisant des réponses semi-spontanées. Les groupes ITL1 et ITL2 ont réalisé la tâche dans la version italienne, et les groupes FRL1 et FRL2 dans la version française. La tâche se déroule comme suit : d'abord, le participant voit une diapositive PowerPoint contenant une courte histoire, accompagnée d'une légende, servant de

1. Les niveaux de compétence des locuteurs en langue étrangère ont été évalués à travers des tests écrits et oraux. Cependant, nous n'allons pas pas largement les discuter dans cette étude. De manière qualitative, en tout cas, il semble que le niveau de compétence ne soit pas un facteur prédominant pour les analyses conduite dans le contexte de cette étude.

2. Le corpus ayant été constitué pour une étude sur les phénomènes acquisitionnels, nous avons pris soin d'évaluer le niveau de compétence des locuteurs, c'est pourquoi nous le signalons dans la description de l'échantillon. Les considérations sur ce facteur restent cependant hors de portée du présent article, dans lequel nous voulons nous concentrer sur l'observation d'un phénomène circonscit, et non sur sa relation avec les niveaux de compétence des locuteurs. En tout état de cause, le niveau de compétence ne semble pas être un paramètre dominant pour les analyses présentées ici. Ppour une étude approfondie dans ce sens, nous renvoyons à (De Paolis, 2024).

phrase de référence. Ensuite, on passe aux diapositives suivantes, qui contiennent les mêmes images, accompagnées de différentes questions écrites. Des exemples de la version italienne sont illustrés à la page suivante, dans la Figure 1.

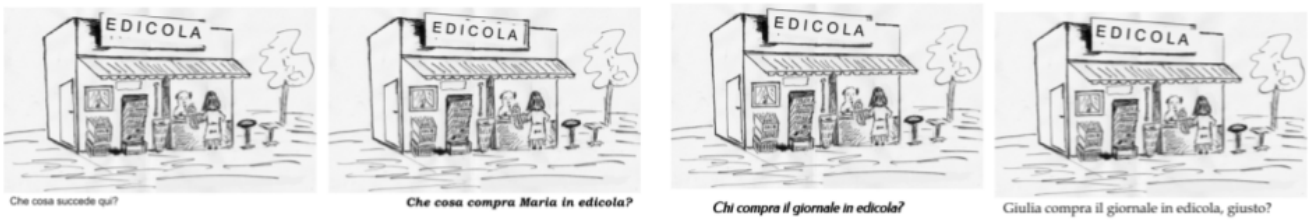


FIGURE 1 – Stimuli de la version italienne de la tâche.

Les participants avancent sont instruits de répondre à chaque question à voix haute, la seule directive étant de ne pas donner des réponses d'un seul mot. Les questions sont conçues pour susciter trois types d'énoncés : focus large (bf), focus étroit d'identification (id),focus étroit de correction (cr). De plus, les questions ciblent divers types de constituants syntaxiques : sujets, verbes, objets, adverbiaux. Pour cette étude, nous nous concentrons sur les énoncés ciblant un type spécifique de constituant syntaxique, à savoir les sujets, dans toutes les conditions possibles : focus large, étroit identificatif, étroit correctif, background (énoncés dans lesquels le constituant focalisé est l'objet). Les sujets ciblés sont les mots "Marie" en français et "Maria" en italien, deux noms propres équivalents dans les deux langues, composés de matériel segmental similaire. Des exemples de questions et de réponses attendues sont fournis ci-dessous (les constituants de focalisation sont soulignés).

— q. Qu'est-ce qu'il se passe ici ?
  a. Marie achète le journal au kiosque. **bf**
— q. Qui achète le journal au kiosque ?
  a. Marie achète le journal au kiosque. **id**
— q. Julie achète le journal au kiosque, n'est-ce pas ?
  a. Non, c'est Marie qui achète le journal au kiosque. **cr**
— q. Qu'est-ce que Marie achète au kiosque ?
  a. Marie achète un journal au kiosque. **bg**

Les enregistrements ont été réalisés dans un environnement insonorisé. Les fichiers audio ont été enregistrés au format .wav, avec une fréquence d'échantillonnage de 44100 Hz.

2.3 Preparation des données et statistiques

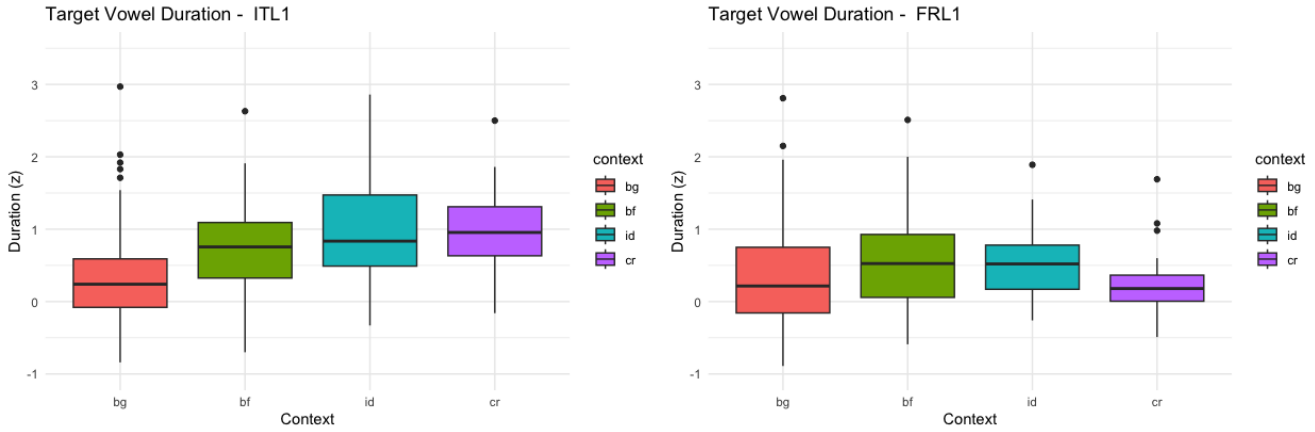
La transcription orthographique a été réalisée manuellement ; la segmentation a été effectuée à l'aide de WebMaus (Kisler *et al.*, 2017) et ajustée manuellement sur Praat (Boersma & Weenink, 2023). La durée moyenne et l'écart type des voyelles ont ensuite été calculés pour chaque locuteur à l'aide du script Polytonia (Mertens, 2014). Chaque participant a produit 30 énoncés ; pour l'analyse, nous n'avons pris en compte que 8 énoncés par locuteur, ceux présentant les sujets cibles dans les quatre conditions de focus : 2 pour le background (bg), 2 pour le focus large (bf), 2 pour l'identification (id) et 2 pour la correction. Nous avons exclu les énoncés dans lesquels le constituant cible était affecté par des hésitations. Dans le cas de l'italien, les voyelles cibles sont [i], le noyau de la syllabe tonique dans "Maria" ; dans le cas du français, la voyelle cible est également [i], située à la frontière droite du constituant focalisé "Marie". Notre ensemble de données comprend 8 observations x 15 locuteurs

x 4 groupes, totalisant 480 unités. Pour l'analyse, la durée des voyelles a été normalisée en scores Z ; les statistiques ont été effectuées sur R (R Core Team, 2022). La normalité de la distribution a été évaluée à l'aide du test de Shapiro-Wilk (Shapiro & Wilk, 1965) et de l'inspection visuelle des graphiques des résidus. Nous avons appliqué un modèle linéaire mixte, estimé en utilisant la méthode REML et l'optimiseur nloptwrap, pour prédire la durée de la voyelle cible en fonction du contexte informationnel (formule : zDur $\sim$ context). Les paramètres normalisés ont été obtenus en ajustant le modèle sur une version normalisée de l'ensemble de données. Les intervalles de confiance à 95% (IC) et les valeurs p ont été calculés en utilisant une approximation de la distribution t de Wald. Le modèle incluait le locuteur en tant qu'effet aléatoire (formule : $1|$speaker).

3 Résultats

3.1 Italien et français L1

Le graphique 2 montre les diagrammes en boîte des durées de la voyelle [i] des mots "Marie" et "Maria" dans les quatres conditions : background, focus large, identification, correction ; à gauche, on voit les résultas du groupe ITL1, et à droite, on voit les résultats du groupe FRL1.



(a) Durée de la voyelle cible (groupe ITL1)

(b) Durée de la voyelle cible (groupe FRL1).

FIGURE 2 – Diagrammes en boîte des durées de la voyelle cible dans les deux groupes L1.

Le graphique 2a montre les résultats des locuteurs italophones natifs. On observe une augmentation de la durée des voyelles accentuées tout au long des quatre conditions, en particulier lors de la transition de background à focus large et de focus large à identification. Dans le cas de l'identification et de la correction, il semble n'y avoir aucune variation significative de la durée. Le modèle statistique confirme ces tendances : en considérant le contexte [bg] comme le niveau de référence, l'effet de [bf] est statistiquement significatif et positif (beta = 0,36, $p < .001$). Les effets de [id] et [cr] sont également significatifs et positifs par rapport à la référence (beta = 0,70, $p < .001$ pour l'identification, beta = 0,75 et $p < .001$ pour la correction), et ils sont également significatifs et positifs par rapport aux conditions [bf]. Cependant, ils ne sont pas séparables l'un un de l'autre.

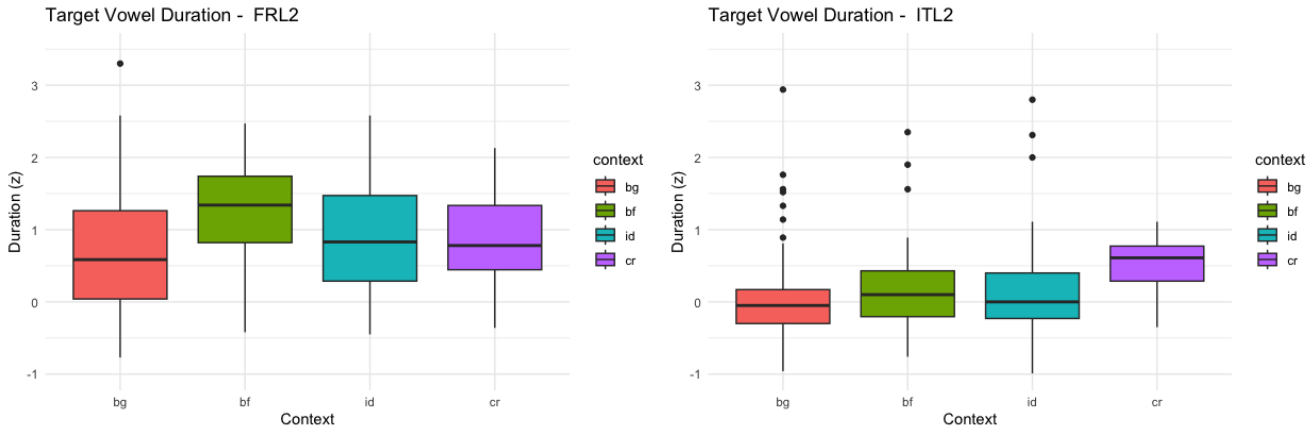
La Figure 2b montre les résultats des locuteurs francophones. Le schéma observé pour ce groupe diffère partiellement de celui observé chez les locuteurs italiens natifs. Notamment, dans les contextes de focalisation large et d'identification, la durée de la voyelle cible est significativement plus élevée

par rapport au background. Contrairement aux locuteurs italiens, cependant, l'effet du contexte [bf] est plus prononcé que celui de [id]. Plus précisément, le contexte de focalisation large présente un score beta positif (beta = 0,42, $p<.001$), qui est plus élevé que celui de la focalisation identificative (beta = 0,23, p = 0,038). Cette tendance à réduire la durée des contextes de focalisation large à focalisation étroite devient encore plus évidente dans le cas de [cr], où l'effet est négatif par rapport à la référence, bien que statistiquement non significatif (beta = -0,04, p = 0,762).

3.2 Comparaison inter-groupes

La comparaison de ces résultats suggère une utilisation différente de l'indice de durée par les deux groupes. Alors que les locuteurs français et italiens marquent la distinction entre background et la focalisation large par un allongement significatif des voyelles, ce mécanisme ne s'applique pas de manière uniforme à la focalisation étroite (id et cr) dans les deux langues. Comme prévu, en italien, la présence d'une focalisation étroite (id ou cr) entraîne des voyelles toniques plus longues pour les constituants focalisés. En français, la tendance semble être inverse : cela pourrait s'expliquer en postulant une relation de *trade-off* de la durée avec d'autres strategies de focalisation, comme par exemple le marquage syntaxique. Nous tenterons d'intégrer et de prendre en compte ce facteur dans le paragraphe de discussion. L'absence d'allongement de la voyelle tonique en contexte contrastif en français L1 pourrait également s'expliquer par le déplacement de la proéminence prosodique de la frontière droite vers la frontière gauche du constituant de focalisation, réalisé sous la forme d'accent initial (German & D'Imperio, 2016).

3.3 Italien et français L2



(a) Durée de la voyelle cible (groupe FRL2)

(b) Durée de la voyelle cible (groupe ITL2)

FIGURE 3 – Diagrammes en boîte des durées de la voyelle cible dans les deux groupes L2.

La Figure 3a représente les résultats des apprenants français en italien L2. Les données montrent une bonne approximation globale vers la langue cible. L'analyse des productions des apprenants francophones suggère que toutes les conditions de focalisation se différencient du niveau de référence (background) par un allongement des voyelles, même si ce n'est pas toujours avec une haute signification statistique. Plus précisément, l'effet du contexte [bf] est positif mais modérément significatif (beta = 0,19, p = 0,016) ; l'effet du contexte [id] est statistiquement significatif et positif (beta = 0,22,

p = 0,006), et l'effet du contexte [cr] est à nouveau positif et statistiquement significatif (beta = 0,44, p = 0,002).

La figure 3b montre les résultats des apprenants italophones de français L2. Dans ce cas, la seule condition de focalisation avec un allongement significatif de la voyelle par rapport au niveau de référence (bg) est la focalisation large (bf). L'effet du contexte [bf], en fait, est statistiquement significatif et positif (beta = 0,53, $p < .001$). L'effet du contexte [id] est positif mais statistiquement non significatif (beta = 0,20, p = 0,138), et l'effet du contexte [cr] est très similaire à celui de [id], positif mais statistiquement non significatif (beta = 0,21, p = 0,240). En comparant ces résultats à ceux des deux groupes L1, nous observons que le comportement des locuteurs italophones en français L2 s'écarte à la fois de la langue cible et de la langue source. Dans l'ensemble, les conditions de focalisation étroite ne présentent pas de différenciation claire par rapport aux deux autres conditions, background et focus large, que ce soit par une durée plus longue (comme chez les locuteurs italiens natifs) ou une durée plus courte (comme observé chez les locuteurs français natifs).

3.4 Comparaison inter-groupes en L2

Les résultats des études sur les groupes L2 mettent en lumière à la fois des différences et des similitudes entre eux ainsi qu'avec les locuteurs des langues sources et cibles. D'un côté, on observe une caractéristique partagée par les locuteurs non natifs des deux groupes : tant chez les apprenants français en L2 que chez les apprenants italiens en L2, la durée des voyelles ne semble pas être utilisée de manière systématique pour distinguer les conditions de focus. Cependant, une différence notable émerge : les apprenants d'italien L2 parviennent à distinguer une condition des trois autres, notamment celle de focalisation corrective, par un allongement significatif des voyelles toniques. De manière intéressante, cette observation contraste avec le groupe FRL1, où l'effet était inverse, montrant un allongement négatif pour la condition [cr].

4 Discussion

4.1 Prosodie et syntaxe : stratégies additives ou alternatives ?

Les résultats tant des locuteurs français natifs que des locuteurs français L2 indiquent un moindre allongement des voyelles dans la condition la plus contrastive, c'est-à-dire le focus correctif. Ce résultat peut être inattendu et ne peut être expliqué qu'avec le recours à l'analyse d'autres niveaux linguistiques, comme celui de la syntaxe. Cela souligne, à notre avis, l'importance d'avoir gardé un protocole flexible pour la collecte de données : cette flexibilité nous a permis de recueillir des informations sur des stratégies de marquage autres que la prosodie. En nous basant sur les résultats de l'analyse syntaxique (De Paolis *et al.*, 2022, in press), en fait, nous savons que le marquage à travers les phrases clivées est particulièrement fréquent dans ces groupes. Cette observation pourrait suggérer une relation alternative et non additive entre la durée et les clivées, ces derniers étant privilégiés dans des contextes plus contrastifs.

4.2 Effets pour l'acquisition en L2

Nos données révèlent que certaines caractéristiques spécifiques des deux langues sources se manifestent dans la production des locuteurs en L2, bien que de manière non uniforme sur tous les aspects. En termes introduits par (Mennen, 2015), nous affirmons que le transfert de la L1 influence principalement le niveau sémantique, plutôt que le niveau de la réalisation phonétique. Les locuteurs en L2 ont tendance à différencier les mêmes conditions que dans leurs langues natives ; cependant, cette différenciation est moins prononcée par rapport à leurs homologues en L1, indiquant une utilisation globalement moindre des indices prosodiques pour marquer la structure de l'information. Nous pensons que l'intégration des résultats de l'analyse syntaxique est cruciale pour expliquer ce phénomène. Plus précisément, nous postulons que le marquage prosodique de la focalisation pose plus de défis aux locuteurs en L2 que le marquage syntaxique, en raison de la similarité syntaxique perçue entre le français et l'italien, qui est majeure par rapport à leur ressemblance au niveau phonologique et phonétique. Par conséquent, les locuteurs français et italiens penchent davantage vers une stratégie reconnue comme similaire à la cible, telle que les phrases clivées ; cette tendance est en fait démontrée par nos données (De Paolis *et al.*, in press). L'utilisation fréquente de phrases clivées en focalisation étroite semble inhiber l'allongement des voyelles, comme si l'indice de durée devenait redondant en présence d'un marquage syntaxique déjà significatif. De plus, la prédominance du marquage syntaxique en L2 concorde avec les résultats d'études sur d'autres combinaisons de langues (au-delà de la similarité perçue). Il a été observé que le marquage prosodique de la focalisation est plus difficile à acquérir que le marquage syntaxique (Zerbian, 2015), et le marquage de la focalisation à travers les clivées s'est avéré plus facile à traiter pour les locuteurs en L2 par rapport à l'encodage prosodique (Yan & Calhoun, 2022).

5 Conclusions et perspectives

L'étude a produit des résultats intrigants, en particulier lorsqu'elle est intégrée aux conclusions de recherches antérieures sur le même matériau. Cependant, nous reconnaissons avoir rencontré certaines difficultés qui pourraient être résolues pour assurer des améliorations dans les travaux futurs. Une difficulté notable est survenue lors de l'analyse de la parole en L2, en raison de la présence de dysfluences, ce qui rendait parfois difficile la distinction entre les hésitations et l'allongement intentionnel des voyelles. L'impact des hésitations dans la planification de la parole a également pu influencer les mesures de durée, telles que la moyenne de la durée des noyaux et l'écart type. Une autre considération importante est que, malgré l'importance reconnue de la validité écologique dans la possibilité d'intégrer la syntaxe dans les études sur la focalisation, l'utilisation de parole semi-spontanée a entraîné une variation considérable dans les réponses des locuteurs. Nous envisageons donc de collecter des données à partir d'un protocole plus contrôlé, pour améliorer l'intégration de ces résultats.

Références

ARVANITI A. (2020). The phonetics of prosody. In M. ARONOFF, Éd., *Oxford Research Encyclopedia of Linguistics*. Oxford Research Encyclopedias.

AVESANI C., BOCCI G., VAYRA M. & ZAPPOLI A. (2015). Prosody and information status in Italian and German L2 intonation. In M. CHINI, Éd., *Il parlato in italiano L2 : aspetti pragmatici e prosodici*, p. 93–116. Milano : Franco Angeli.

AVESANI C. & VAYRA M. (2003). Broad, narrow and contrastive focus in florentine Italian. In *ICPhS-15*, p. 1803–1806. ISCA.

BOERSMA P. & WEENINK D. (2023). Praat : doing phonetics by computer [computer program].

DE JONG K. (2004). Stress, lexical focus, and segmental focus in English : patterns of variation in vowel duration. *Journal of Phonetics*, **32, 4**, 493–516.

DE JONG K. & ZAWAYDEH B. (2002). Comparing stress, lexical focus, and segmental focus : Patterns of variation in Arabic vowel duration. *Journal of Phonetics*, **30**, 53–75.

DE L'EUROPE C. (2020). *Common European Framework of Reference for Languages : Learning, Teaching, Assessment—Companion Volume*. Council of Europe Publishing.

DE PAOLIS B. M. (2024). *Focus induced variations in prosody and word order in native and non-native Italian and French*. Thèse de doctorat, Università di Torino, Université Paris 8.

DE PAOLIS B. M., ROMANO A. & ANDORNO C. (in press). Prosodic and syntactic markers of narrow focus in Italian L1 and L2 (French L1) : an experimental study . In *Il parlato in ambito medico : analisi linguistica, applicazioni tecnologiche e strumenti clinici. Atti del XIX convegno annuale AISV, Lecce 2023*. Officinaventuno.

DE PAOLIS B. M., SANTIAGO F. & ANDORNO C. (2022). Syntaxe ou prosodie ? Une étude préliminaire sur l'expression de la focalisation étroite par les apprenants italophones de français L2. In *Proc. XXXIVe Journées d'Études sur la Parole*, p. 851–860 : ISCA Archives.

D'IMPERIO M. (2002). Italian intonation : An overview and some questions. *Probus*, **14(1)**, 37–69.

FARNETANI E. & ZMARICH C. (1997). Prominence patterns in Italian : an analysis of F0 and duration. In A. BOTINIS, G. KOUROUPETROGLOU & G. CARAYANNIS, Éds., *Intonation : Theory, Models and Applications. Proceedings of ESCA Workshop, Athens, September 18 - 20 september 1997*, p. 115–118.

GABRIEL C. (2010). On Focus, Prosody, and Word Order in Argentinian Spanish. A Minimalist OT Account. *In Revista Virtual de Estudos da Linguagem*, **Special issue 4**, 183–222.

GABRIEL C. & GRÜNKE J. (2018). Focus, prosody, and subject positions in L3 Spanish : analyzing data from German learners with Italian and Portuguese as heritage languages. In M. G. GARCÍA & M. UTH, Éds., *Focus realization in Romance and beyond*, p. 358–386. Amsterdam : John Benjamins.

GABRIEL C. & KIREVA E. (2014). Prosodic transfer in learner and contact variety : Speech Rhythm and Intonation of Buenos Aires Spanish and L2 Castilian Spanish Produced by Italian Native Speakers. *Studies in Second Language Acquisition*, **36**(2), 257–281. DOI : 10.1017/S0272263113000740.

GERMAN J. S. & D'IMPERIO M. (2016). The Status of the Initial Rise as a Marker of Focus in French. *Language and Speech*, **59(2)**, 165–195.

KISLER T., REICHEL U. D. & SCHIEL F. (2017). Multilingual processing of speech via web services. *Computer Speech and Language*, **45**, 326–347.

MAEKAWA K. (1997). Effects of focus on duration and vowel formant frequency in japanese. In Y. SAGISAKA & N. CAMPBELL, N.AND HIGUCHI, Éds., *Computing Prosody*, p. 129–153. Springer, New York.

MENNEN I. (2015). Beyond Segments : Towards a L2 Intonation Learning Theory. In E. DELAIS-ROUSSARIE, M. AVANZI & S. HERMENT, Éds., *Prosody and Language in Contact. Prosody, Phonology and Phonetics.*, p. 171–188. Springer, Berlin, Heidelberg.

MERTENS P. (2014). Polytonia : a system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences*, **4(2)**, 17–57.

MICHELAS A. & GERMAN J. (2020). Focus marking and prosodic boundary strength in French. *Phonetica*, **77**, 244–267.

R CORE TEAM (2022). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SHAPIRO S. S. & WILK M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, p. 591–611.

TREMBLAY A. & GARRISON M. (2010). Cloze tests : A tool for proficiency assessment in research on l2 french. In M. T. PRIOR, Éd., *Selected Proceedings of the 2008 Second Language Research Forum*, p. 73–88. MA : Cascadilla Proceedings Project.

VANDER KLOK J., GOAD H. & WAGNER M. (2018). Prosodic focus in English vs. French : A scope account. *Glossa : a journal of general linguistics*, **3 (1) : 71**.

VEDDER I. (2008). Competenza pragmatica e complessità sintattica in italiano l2 : l'uso dei modificatori nelle richieste. *Linguistica e Filologia*, **25(1)**, 99–123.

YAN M. & CALHOUN S. (2022). Prosodic prominence and clefting in L2 focus interpretation. In *Proc. Speech Prosody*, p. 901–905.

ZERBIAN S. (2015). Markedness considerations in L2 prosodic focus and givenness marking. In E. DELAIS-ROUSSARIE, M. AVANZI & S. HERMENT, Éds., *Prosody and Language in Contact. Prosody, Phonology and Phonetics*, p. 7–27. Berlin, Heidelberg : Springer.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Analyse Factorielle de signaux sonores : développement d'une méthode automatique de détermination des frontières optimales entre canaux de fréquence.

Agnieszka Duniec[1] Elisabeth Delais-Roussarie[1] Olivier Crouzet[1]
(1) Laboratoire de Linguistique de Nantes, LLING – UMR6310, Université de Nantes / CNRS
chemin de la Censive du Tertre, 44312 Nantes Cedex, France
agnieszka.duniec@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Des études récentes supportent l'hypothèse d'une relation entre les propriétés statistiques des signaux de parole et les mécanismes perceptifs : les gammes de fréquence présentant une corrélation dans leurs modulations d'amplitude pourraient être associées à des frontières spectrales relativement stables envisagées comme optimales sur le plan perceptif. Cependant, des limites afférentes à ces études antérieures ressortent : (1) elles se fondent pour la plupart sur des critères subjectifs à travers l'observation visuelle des courbes de résultats statistiques, et (2) elles n'envisagent pas que les résultats puissent varier en fonction des échantillons de données sélectionnés, de la nature des signaux utilisés, ou de la taille des échantillons. Même si cette position peut être argumentée en lien avec l'approche du *codage efficace*, cet aspect afférent au degré de variation potentiel nécessite d'être évalué. Nous avons mis en place une méthode de détermination automatique des frontières qui permet de répliquer les travaux antérieurs en introduisant une évaluation expérimentale de ces limites et discutons de quelques résultats préliminaires en comparaison avec les études précédentes.

ABSTRACT

Factor Analysis of acoustic signals : development of an automatic method for the determination of optimal frequency boundaries.

Recent studies have led to support the hypothesis of a relationship between statistical properties of spectro-temporal modulations in speech and mechanisms of perceptual analysis : frequency channels exhibiting co-modulated amplitude would be associated with frequency boundaries considered as *optimal* in perceptual terms. However, limits pertaining to some of these studies have to be taken into accout : (1) previous results associated with frequency boundary determination were for some part based on visual inspection of statistical curves, (2) studies have generally assumed that these results would hold for any sample of speech signals, and for any sample size. Even though such assumption may hold in relation to the *efficient coding* hypothesis, the degree of variation relating to the observed results requires to be investigated. A method for the automatic determination of frequency boundaries associated with these approaches has been developped and applied on a speech database. This method provides a way to replicate previous data while experimentally investigating variation. Some preliminary results are discussed and compared with previous studies.

MOTS-CLÉS : perception, statistiques des signaux naturels, hypothèse du codage efficace, implants cochléaires.

KEYWORDS: perception, natural signal statistics, efficient coding hypothesis, cochlear implants.

1 Introduction

L'*hypothèse du codage efficace* pour la perception sonore (Smith & Lewicki, 2006) tire ses origines des travaux sur les *statistiques des signaux naturels* (Simoncelli & Olshausen, 2001; McDermott & Simoncelli, 2011). Du point de vue de cette approche, les signaux acoustiques de communication sont caractérisés par des propriétés statistiques régulières, lesquelles seraient au fondement des mécanismes d'analyse perceptive malgré la diversité apparente des réalisations sonores. Si certaines de ces études (Ming & Holt, 2009; Kluender *et al.*, 2013) se fondent sur une modélisation des signaux à travers des trains d'impulsions (*spike trains*) en lien avec la théorie de l'information et les approches contemporaines du codage neuronal (Smith & Lewicki, 2005), d'autres ont adopté une approche statistique plus traditionnelle fondée sur l'Analyse Factorielle des modulations d'amplitude (Ueda & Nakajima, 2017; Grange & Culling, 2018).

Plusieurs de ces études se sont penchées sur cette hypothèse dans le but d'identifier des propriétés statistiques des signaux de parole dans différentes langues (Ueda & Nakajima, 2017) ainsi qu'en lien avec le codage de la parole dans des implants cochléaires (Grange & Culling, 2018; Ming & Holt, 2009).

1.1 Hypothèse du codage efficace et implant cochléaire

Ming & Holt (2009) ont mesuré les performances de reconnaissance de parole vocodée, souvent décrite comme simulant les informations diffusées par les implants cochléaires, chez des auditeurs normo-entendants. Ils ont montré que, sans changer le nombre de canaux spectraux (6 en l'occurrence) les changements de localisation des frontières spectrales en parole vocodée ont des effets sur les taux de reconnaissance de mots et de segments phonétiques. Les performances sont meilleures si les localisations de ces frontières concordent avec des positions spectrales dérivées de modélisations issues de la théorie de l'information et correspondent donc à une « perspective efficace ». Ces frontières seraient en outre nettement plus basses que celles qui découlent d'une organisation tonotopique, lesquelles sont généralement la référence pour la décomposition spectrale dans les implants cochléaires.

Dans une toute autre perspective, Ueda & Nakajima (2017), ont développé une méthode d'analyse inspirée des travaux de Plomp *et al.* (1967) sur les voyelles : ils étendent cette approche à l'étude d'un corpus de phrases et procèdent, sur la base de signaux acoustiques de parole codés sur environ 20 canaux de représentation spectrale à des Analyses en Composantes Principales (ACP) portant sur les enveloppes d'énergie de ces canaux. Il font varier le nombre de facteurs associés à la sortie de l'ACP (2, 3, 4, 5, 6). Leur travail aboutit à la conclusion que 4 facteurs suffiraient à représenter optimalement des signaux de parole, et ce pour chacune des 8 langues de leur échantillon. Ils constatent par ailleurs que les 3 frontières fréquentielles découlant de chacune des ACP à 4 facteurs réalisées sur ces 8 langues seraient parfaitement appariées (env. 540, 1720, 3300 Hz), ce qui les amène à conclure que les langues seraient de manière générale fondées sur des indices qui seraient parfaitement adaptés à un traitement perceptif « parcimonieux » (ou efficace) de la parole.

Grange & Culling (2018) ont répliqué l'étude de Ueda & Nakajima (2017) en modifiant légèrement l'algorithme d'analyse statistique (accroissement du nombre de canaux spectraux entrés dans l'ACP à environ 100 canaux notamment, estimation de la contribution de chacune des 20 premières Composantes Principales issues de l'ACP à travers les valeurs propres –*eigenvalues*). Ils ont ensuite évalué ces données à la lueur des performances observées en perception de parole vocodée (simulations

d'implants cochléaires) et ont abouti à des conclusions assez similaires aux travaux précédents. Leurs résultats suggèrent néanmoins que, pour rendre compte de manière appropriée des propriétés acoustiques de la parole vocodée, il faudrait 6 à 7 canaux spectraux pour représenter optimalement ces signaux. Cette limite correspond, dans le graphique des valeurs propres (*scree plot*) qu'ils présentent, à un point d'inflexion au-delà duquel les valeurs propres semblent augmenter plus lentement. Cette limite est aussi associée dans les courbes de performance en fonction du nombre de canaux vocodés, à une amélioration moins marquée des performances observées à partir de 8 canaux spectraux. Ces deux mesures (l'une statistique issue de signaux naturels, l'autre comportementale issue de signaux vocodés) seraient donc cohérentes et suggèreraient que cette limite de 7/8 canaux pourrait refléter une version optimale de la représentation perceptive des signaux de parole appropriée à la tâche expérimentale utilisée (reconnaissance de mots dans des phrases simples présentées dans le silence).

1.2 Limites des études antérieures

Toutes les études évoquées ont dérivé, de l'analyse acoustique et statistique d'un corpus de parole, des estimations de localisation de frontières entre canaux de fréquence qui sont considérées comme optimales : elles différencieraient des canaux de fréquence maximalement comodulés entre eux et ces frontières sépareraient les canaux qui sont les moins corrélés et qui, par conséquent, seraient maximalement informatifs. Il ressort de ces approches qu'aucun des travaux précédents n'a évalué le degré de variation des estimations réalisées en fonction de la composition du corpus, de sa taille (environ 1h de parole dans les études précécentes) ni de la durée acoustique des items concaténés pour la constitution du corpus. Or il semble essentiel, avant d'envisager un caractère stable de ces *frontières optimales*, de pouvoir évaluer leur variabilité potentielle.

Une limite forte à l'étude de cette variation est liée au fait que dans les études précédentes, les valeurs de ces *fréquences optimales* séparant les canaux de fréquence ont été estimées par le biais d'inspections purement visuelles des graphiques de résultats. Il va de soi que pour être en mesure d'évaluer le degré de variation d'une telle mesure, il convient de mettre en place une méthode de détermination automatique qui ouvrirait la voie au calcul d'un grand nombre d'estimations différentes en étudiant aussi bien les effets liés à la taille globale du corpus que ceux afférents à la dispersion des résultats lorsqu'on sélectionne aléatoirement des sous-ensembles distincts du corpus d'origine. Nous présentons dans cet article la méthode que nous avons développée dans cette perspective et donnons une illustration préliminaire des résultats à partir de la comparaison avec les études antérieures (Ueda & Nakajima, 2017; Grange & Culling, 2018; Ming & Holt, 2009)

2 Méthode

Les analyses acoustiques et statistiques ont été réalisées dans l'environnement Matlab. Les scripts d'analyse sont disponibles sur un dépôt OSF [1].

1. Lien vers le dépôt OSF en lecture : https://page.hn/9ij6kk

2.1 Corpus

La base de données utilisée est la *Clarity Speech Database* (Graetzer *et al.*, 2021) qui contient des enregistrements de parole en anglais en accès libre (échantillonnés à 44.1 kHz sur 32 bits au format WAV). La base de données complète est composée d'environ 10000 phrases issues du *British National Corpus (BNC)*, lues par 40 locuteurs et locutrices de l'anglais Britannique. Dans le cadre de l'étude présentée ici, un échantillon aléatoire de 1600 phrases en a été extrait dans le but de cibler une durée totale équivalente à celle utilisée dans les précédentes études (environ 1 h de parole).

2.2 Paramètrage acoustique des signaux

Préalablement à l'analyse statistique des signaux, nous procédons à une paramétrisation acoustique comparable à celles qui ont été utilisées dans les travaux précédents (Ueda & Nakajima, 2017; Grange & Culling, 2018). Les signaux extraits de chaque fichier sont concaténés les uns aux autres. Le signal concaténé est soumis à un filtrage passe-bas (fréquence de coupure 8 kHz) et sous-échantillonné à 16 kHz. Les enveloppes de modulation temporelle des signaux sont extraites à partir d'un banc de filtres dont la largeur croît de manière logarithmique avec la fréquence centrale (canaux de largeur $\frac{1}{4}$ d'ERB, Moore & Glasberg, 1983, ce qui correspond à 116 canaux spectraux allant jusqu'à la fréquence supérieure maximale de 8 kHz). Ces enveloppes subissent une rectification demi-onde puis un filtrage passe-bas avec une fréquence de coupure de 50 Hz (fréquence d'échantillonnage 100 Hz). Les signaux d'enveloppe résultants sont ensuite élevés au carré et convertis en notes centrées réduites (*z-scores*). Cette chaîne de paramétrage permet de procéder à une analyse des co-modulations d'énergie entre les bandes de fréquence (corrélation entre les enveloppes d'énergie des canaux).

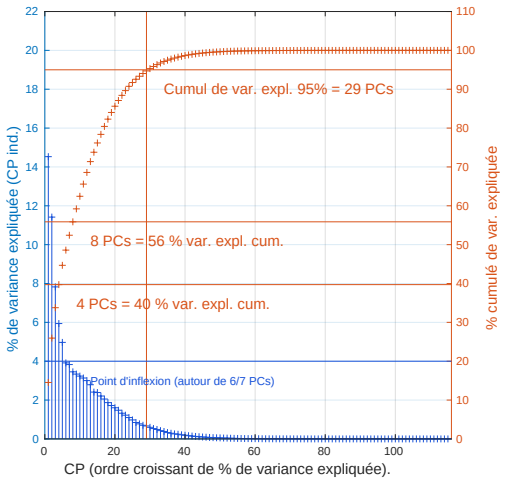


FIGURE 1 – Graphe des valeurs propres (% de variance expliquée) issu de l'Analyse Factorielle (en bleu : % associés à chaque CP individuelle — en rouge : % cumulés). La CP pour laquelle on atteint un cumul égal à 95% est indiquée, ainsi que le % de variance cumulée expliquée pour resp. 4 et 8 Composantes Principales.

La matrice de données résultante, composée de 116 canaux de fréquence, correspond aux modulations temporelles de l'enveloppe de chaque canal spectral au cours du temps. Elle est transférée vers un outil statistique d'Analyse en Composantes Principales afin de procéder à une Analyse Factorielle.

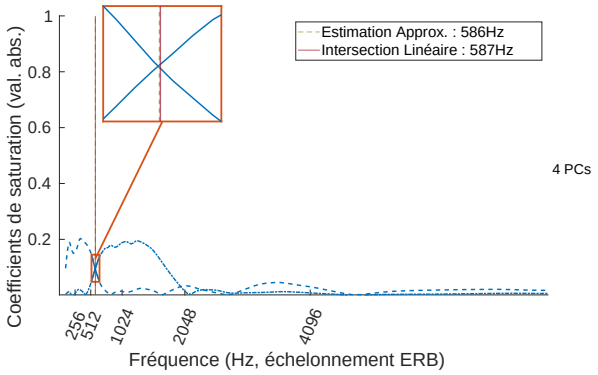
2.3 Analyse Factorielle

L'*Analyse Factorielle* est une méthode descriptive d'analyse de données qui repose sur la technique d'*Analyse en Composantes Principales* (ACP). Elle permet une étude simultanée de plus de 2 dimensions (analyse multivariée). L'objectif est de représenter l'essentiel de l'information contenue dans un tableau de données quantitatif en réduisant le nombre de facteurs explicatifs. Le principe

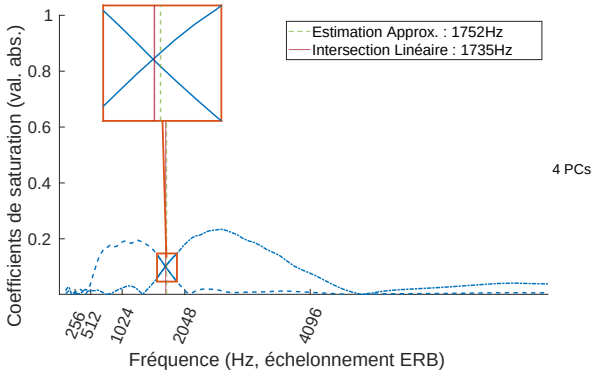
est de transformer des variables liées (statistiquement corrélées entre elles) en nouvelles variables *synthétiques*. Les Composantes Principales (CP) sont donc regroupées en facteurs abstraits.

Concrètement, les variables initiales sont représentées dans un nouvel espace de facteurs définis par les vecteurs propres de la matrice de corrélations. L'hypothèse sous-jacente à l'application de cette méthode sur des signaux sonores est que certains canaux spectraux contiendraient des informations redondantes et qu'il serait alors économique de restreindre l'analyse perceptive à une séparation en zones de fréquences étant maximalement informatives (donc minimalement redondantes). En celà, l'Analyse Factorielle permettrait d'identifier les canaux de fréquence optimaux pour différencier de manière parcimonieuse les propriétés sonores distinctives d'un corpus. Une description plus précise de la procédure d'ACP mise en oeuvre est disponible dans Duniec *et al.* (2022). Pour information, le graphe des valeurs propres indiquant le pourcentage cumulé de variance expliquée en fonction du nombre de CP est représenté dans la Fig. 1.

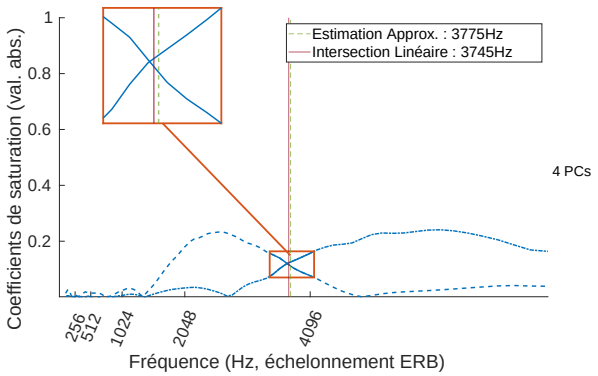
Tout comme dans les travaux antérieurs (Ueda & Nakajima, 2017; Grange & Culling, 2018; Duniec *et al.*, 2022), ces valeurs sont représentées sous forme de courbes de coefficients de saturation dont la valeur absolue fait ressortir quelles sont les gammes de fréquences qui sont maximalement corrélées entre elles (cf. Duniec *et al.*, 2022, pour une illustration). Ce regroupement indique quelles fréquences sont associées / co-modulées. Les gammes de fréquences correspondantes sont, dans le cadre de l'Analyse Factorielle, associées à un facteur synthétique / une Composante Principale. C'est sur la base de ces ensembles de courbes de coefficients de saturation que nous procédons ensuite à une estimation automatique de ces frontières pour chaque condition associée au nombre de Composantes Principales retenues.



(a) 1ère frontière



(b) 2e frontière



(c) 3e frontière

FIGURE 2 – Illustration de la procédure de détermination des intersections entre les courbes adjacentes de coefficients de saturation par prédiction linéaire.

Les intervalles des courbes de coefficients de saturation pour lesquels la valeur absolue est relativement élevée sont interprétées comme représentant des zones spectrales co-modulées en amplitude, lesquelles sont considérées dans cette perspective comme ne fournissant que peu d'in-

formations perceptives distinctes. Déterminer la localisation de leurs frontières fournirait donc une information sur la zone optimale de découpage spectral qui pourrait par exemple être utilisée dans un implant cochléaire dans le but de dissocier de manière optimale les gammes de fréquence maximalement porteuses d'information pour un nombre de canaux déterminé.

2.4 Estimation des frontières entre canaux

La procédure de détermination des frontières entre chaque paire de courbes de coefficients de saturation adjacentes est ensuite appliquée.

On procède d'abord à une première phase de traitements préparatoires. Une interpolation linéaire des courbes de coefficients de saturation dans un rapport de $\frac{10}{1}$ vise à accroître le nombre de points de mesure sur l'échelle des fréquences. On estime ensuite la localisation en fréquence du pic de la courbe de coefficients de saturation associée à chaque Composante Principale. Pour chacune de ces courbes, on repère approximativement deux frontières (respectivement inférieure et supérieure au pic) à partir d'un critère de valeur du coefficient qui descende à 25% de la valeur maximale observée au pic en se déplaçant du pic vers chacun des bords de la courbe. On classe enfin les vecteurs associés aux courbes de coefficient de saturation par valeur croissante de fréquence associée au pic, ce qui permet de déterminer quelles sont les courbes adjacentes en termes de canaux de fréquence.

On procède enfin, pour chaque paire de courbes adjacentes, à la détermination des coordonnées de leur point d'intersection par modélisation linéaire [2] limitée à la zone probable du croisement. La Fig. 2 illustre cette dernière partie de la procédure pour une Analyse Factorielle retenant 4 Composantes Principales (3 frontières). Ceci repose sur la modélisation par une fonction linéaire de chacune des deux portions de courbes. Pour terminer, on estime les coordonnées de leur point d'intersection dans un espace *Valeur du coefficient de saturation en dB* $\sim$ *Fréquence en Hz* (après échelonnement ERB).

3 Résultats préliminaires

Nous proposons ici une première comparaison avec les données de la littérature. Il est important de noter que les divergences constatées pourraient trouver leur source dans différents paramètres. Le contenu du corpus de parole utilisé peut impacter les résultats. Ainsi, ces études reposent sur des corpus de parole comparables en termes de type de contenu (phrases, multi-locuteurs, langue anglaise) et de conditions de collecte mais qui sont néanmoins différents. Par ailleurs, des aspects précis de la méthode appliquée pour procéder à la paramétrisation acoustique puis à la mise en œuvre de l'Analyse Factorielle peuvent également jouer (par ex. Ueda & Nakajima, 2017 procédent à une décomposition spectrale initiale en seulement 20 canaux alors que nous adoptons, conformément à Grange & Culling, 2018, une décomposition spectrale en plus de 100 canaux). Enfin, le recours à une procédure automatisée vs. visuelle pourrait affecter la précision des estimations. Concernant ce dernier point cependant, les algorithmes que nous avons implémentés semblent produire des résultats conformes aux estimations visuelles.

Notre objectif principal dans cet article est de chercher à mettre en évidence deux points importants : (1) la méthode automatique mise en œuvre fournit des résultats cohérents en termes d'estimation de la

2. Des tests ont également été effectués avec des modélisations polynomiales d'ordre 2 et 3 mais les résultats de la modélisation linéaire étaient généralement plus conformes aux observations visuelles que les fonctions d'ordre supérieur.

fréquence à laquelle le croisement des courbes adjacentes se fait et (2) les données recueillies varient parfois fortement entre les études considérées, ce qui justifie pleinement la nécessité de procéder à une exploration de la variation associée à ces mesures.

Nous procédons à une comparaison de nos résultats avec les données de la littérature qui sont disponibles et qui ont eu recours à la même méthode de traitement (Analyse Factorielle), donc respectivement pour 4 et 8 canaux. Les données de Ming & Holt (2009) portent sur un découpage en 6 canaux mais sont fondées sur une méthode de traitement profondément différente, nous ne les considérons pas directement dans cet article pour des raisons d'espace disponible.

3.1 Résultats pour 4 canaux de fréquence

Les résultats pour 4 canaux de fréquence sont disponibles dans les principales études auxquelles nous nous sommes référés (Ueda & Nakajima, 2017; Grange & Culling, 2018), lesquelles sont toutes les deux fondées sur des ACP réalisées sur les enveloppes d'énergie malgré des choix différents en termes de nombre de canaux initiaux pour la décomposition spectrale (20 pour Ueda & Nakajima, 2017 vs. plus de 100 pour l'étude de Grange & Culling, 2018). Les valeurs numériques de localisation des frontières pour chacun des articles ont été déterminées en utilisant un logiciel d'extraction de données quantitatives à partir de graphes (`g3data` [3]) sur la base des figures publiées. Nous avons utilisé cet outil pour estimer / extraire les coordonnées de fréquence des points d'intersection des courbes adjacentes. Les données de comparaison sont présentées dans la Table 1. On peut constater que les résultats sont cohérents les uns par rapport aux autres mais que des divergences émergent et peuvent pour certaines atteindre 1 à 2 demi-tons d'écart.

Ces degrés de divergence peuvent paraître parfaitement acceptables et pourraient ne relever que de variations attendues dans tout travail quantitatif même s'il conviendra de documenter cette variation. La comparaison avec les résultats publiés pour 8 canaux (cf. *infra.* et Table 2) fait par contre ressortir des différences nettement plus marquées.

TABLE 1 – Estimations (en Hertz) de la localisation des frontières optimales entre canaux de fréquence et écarts mesurés par rapport aux données des travaux antérieurs (en demi-tons) pour 4 Composantes Principales (respectivement : Données de Ueda & Nakajima (2017) ; Données de Grange & Culling (2018) ; Notre estimation par modélisation linéaire ; Différences mesurées entre les estimations issues de la littérature et nos données).

	1/2	2/3	3/4
Ueda & Nakajima (2017)	540	1720	3300
Grange & Culling (2018)	573	1570	3827
Nos observations	587	1735	3745
Écart / Ueda & Nakajima (2017, demi-tons)	1.44	0.15	2.19
Écart / Grange & Culling (2018, demi-tons)	0.41	1.73	-0.38

3.2 Résultats pour 8 canaux de fréquence

Les résultats pour 8 canaux de fréquence (cf. Table 2) ne sont disponibles que pour l'étude de Grange & Culling (2018). Ils ont été déterminés en utilisant le même logiciel d'extraction de

3. https://github.com/pn2200/g3data/

données quantitatives à partir de graphes (`g3data`). On peut constater que certaines estimations sont considérablement plus divergentes, notamment dans les basses et moyennes fréquences (avec des écarts de l'ordre de 10 à 15 demi-tons –autour d'une octave–, ainsi que dans les hautes fréquences –3 à 4 demi-tons), ce qui laisse entrevoir des potentialités de variation considérables de ces propriétés statistiques. Il est notable que la procédure de traitement acoustique et statistique mise en œuvre dans notre approche est en tous points comparable à celle implémentée par Grange & Culling (2018), ce qui laisse supposer un potentiel de variation considérable qu'il conviendra d'explorer.

TABLE 2 – Estimation de la localisation des frontières optimales entre canaux de fréquence pour 8 CP (en Hz). Comparaison avec les observations de Grange & Culling (2018).

	1/2	2/3	3/4	4/5	5/6	6/7	7/8
Données de Grange & Culling (2018)	442	652	1159	1518	1916	2749	4104
Nos estimations	197	332	678	1474	2099	3377	5085
Écart (demi-tons)	-14.00	-11.69	-9.29	-0.51	1.58	3.56	3.71

4 Discussion

Nos observations contrastent assez nettement avec les résultats antérieurs en termes de correspondance des frontières de fréquences telles qu'elles peuvent être déduites de l'Analyse Factorielle de signaux de parole. Si les résultats pour 4 canaux semblent acceptables en termes de dispersion naturelle des mesures empiriques, les résultats obtenus avec 8 canaux mettent en évidence un degré de variation qui peut atteindre plus d'une octave dans certaines gammes de fréquence, ce qui aurait de toute évidence des conséquences perceptives notables.

En outre, l'hypothèse de Ueda & Nakajima (2017) concernant la stabilité de ces mesures pour 8 langues distinctes semble assez spéculative si l'on considère le degré de variation que nous avons commencé à documenter ici. En effet, cette comparaison montre que, pour des données reposant sur une méthode d'analyse acoustique et statistique équivalente, l'utilisation d'une base de données différente dans la même langue et pour un corpus de phrases multi-locuteurs comparable en taille, débouche sur des divergences de valeurs des frontières dont l'empan est considérable.

À tout le moins, même si toute variation est prévisible et attendue dans le cadre de la mesure empirique de phénomènes naturels, ces données mettent en évidence un potentiel de variation important que nous pourrons explorer de manière systématique avec la procédure présentée dans cet article.

Remerciements

Ce travail a reçu le soutien du programme Recherche – Formation – Innovation « Ouest Industries Créatives » (RFI-OIC, Région Pays de la Loire) par une allocation doctorale attribuée à AD.

Références

DUNIEC A., CROUZET O. & DELAIS-ROUSSARIE E. (2022). Analyse factorielle de signaux musicaux : comparaison avec les données de parole dans la perspective de l'hypothèse du codage efficace et de l'application aux implants cochléaires. In O. CROUZET, E. DELAIS-ROUSSARIE, A. DUNIEC, L. LEPRIEUR, P. L. ROHRER, M. TAHON, J. WOTTAWA, M. BRABANT, H. RIGUIDEL, N. BARBOT, S. GIBET, D. LOLIVE & A. SINI, Éds., *Actes des 34e Journées d'Études sur la Parole – JEP2022*, Île de Noirmoutier, France : Nantes Université / Le Mans Université / IRISA / CNRS International Speech Communication Association – ISCA archive.

GRAETZER S., AKEROYD M. A., BARKER J., COX T. J., CULLING J. F., NAYLOR G., PORTER E. & MUÑOZ R. V. (2021). Dataset of british english speech recordings for psychoacoustics and speech processing research : The clarity speech corpus. *Data in Brief*, p. 107951.

GRANGE J. & CULLING J. (2018). The factor analysis of speech : Limitations and opportunities for cochlear implants. *Acta Acustica united with Acustica*, **104**, 835–838. DOI : 10.3813/AAA.919253.

KLUENDER K. R., STILP C. E. & KIEFTE M. (2013). Perception of Vowel Sounds Within a Biologically Realistic Model of Efficient Coding. In G. S. MORRISON & P. F. ASSMANN, Éds., *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, p. 117–151. Berlin, Heidelberg : Springer. DOI : 10.1007/978-3-642-14209-3_6.

MCDERMOTT J. H. & SIMONCELLI E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery : Evidence from Sound Synthesis. *Neuron*, **71**(5), 926–940. DOI : 10.1016/j.neuron.2011.06.032.

MING V. L. & HOLT L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, **126**(3), 1312–1320. DOI : 10.1121/1.3158939.

MOORE B. C. J. & GLASBERG B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, **74**, 750–753.

PLOMP R., POLS L. C. W. & VAN DE GEER J. P. (1967). Dimensional Analysis of Vowel Spectra. *The Journal of the Acoustical Society of America*, **41**(3), 707–712. DOI : 10.1121/1.1910398.

SIMONCELLI E. P. & OLSHAUSEN B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, **24**(1), 1193–1216. DOI : 10.1146/annurev.neuro.24.1.1193.

SMITH E. & LEWICKI M. S. (2005). Efficient Coding of Time-Relative Structure Using Spikes. *Neural Computation*, **17**(1), 19–45. DOI : 10.1162/0899766052530839.

SMITH E. C. & LEWICKI M. S. (2006). Efficient auditory coding. *Nature*, **439**(7079), 978–982. DOI : 10.1038/nature04485.

UEDA K. & NAKAJIMA Y. (2017). An acoustic key to eight languages/dialects : Factor analyses of critical-band-filtered speech. *Scientific Reports*, **7**, 42468. DOI : 10.1038/srep42468.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Apprentissage profond pour l'analyse de la parole pathologique : étude comparative entre modèles CNN et à base de transformers

Malo Maisonneuve[1] Corinne Fredouille[1] Muriel Lalain[2] Alain Ghio[2]
Virginie Woisard[3]
(1) LIA, Avignon University, France
(2) Aix-Marseille University, CNRS, LPL, Aix-en-Provence, France
(3) LNPL, Toulouse University and Toulouse Hospital, Toulouse, France
prénom.nom@[univ-avignon ; univ-amu].fr

RÉSUMÉ

Les cancers des voies aérodigestives supérieures (VADS) ont un impact significatif sur la capacité des patients à s'exprimer, ce qui affecte leur qualité de vie. Les évaluations actuelles de la parole pathologique sont subjectives, justifiant le besoin de méthodes automatiques et objectives. Un modèle auto-supervisé basé sur Wav2Vec2 est proposé pour la classification de phonèmes chez les patients atteints de cancer des VADS, visant une amélioration des taux de bonne classification et une meilleure discrimination des caractéristiques phonétiques. Les impacts des paramètres d'affinage, des données de pré-entraînement, de la taille du modèle et des données d'affinage sont explorés. Nos résultats montrent que l'architecture Wav2Vec2 surpasse une approche basée sur un CNN, et montre une corrélation significative avec les mesures perceptives. Ce travail ouvre la voie à une meilleure compréhension de la parole pathologique, via une représentation auto-apprise de la parole, très pertinente pour des approches d'interprétation à destination des cliniciens.

ABSTRACT

Deep learning for speech pathology : a comparative analysis of CNN and transformer-based models

Head and neck cancers significantly impact patients' ability to speak, affecting their quality of life. Commonly used metrics for assessing pathological speech are subjective, prompting the need for automated and unbiased evaluation methods. This study proposes a self-supervised Wav2Vec2-based model for phone classification in HNC patients, aiming to enhance accuracy and improve the discrimination of phonetic features. The impact of fine-tuning parameters, pre-training datasets, model size, and fine-tuning datasets will be explored. Evaluation on diverse corpora reveals the effectiveness of the Wav2Vec2 architecture, outperforming a CNN-based approach, used in previous work. Correlation with perceptual measures also affirms the model's relevance for clinical speech analysis. This work paves the way for a better understanding of pathological speech, by leveraging a complex self-learnt speech representation, relevant for interpretability approaches for clinicians.

MOTS-CLÉS : troubles de la parole, cancer de la tête et du cou, apprentissage profond, classification de phonèmes, intelligibilité, interprétabilité.

KEYWORDS: speech disorders, Head and Neck Cancer, deep learning, phone classification, intelligibility, interpretability.

Traduction de l'article accepté à Interspeech 2024 : "Towards objective and interpretable speech disorder assessment: a comparative analysis of CNN and transformer-based models"

1 Introduction

Les cancers des voies aérodigestives supérieures (VADS) affectent les voies respiratoires et digestives supérieures, notamment la cavité buccale, le pharynx, le larynx, la cavité nasale et les glandes salivaires. Le traitement de ce cancer, que ce soit par radiothérapie, chimiothérapie et/ou chirurgie, peut impacter significativement la parole des patientes et patients qui en souffrent. La difficulté à communiquer avec les autres a ainsi un impact négatif sur leur qualité de vie. Évaluer correctement la parole qu'ils ou elles produisent, en identifiant le niveau d'altération et ce qui la rend atypique est essentiel pour les aider au mieux lors de séances de rééducation. Malheureusement, les métriques couramment utilisées telles que la sévérité ou le niveau d'intelligibilité sont subjectives et susceptibles d'être mal évaluées, même par des experts (Astésano *et al.*, 2018). Par ailleurs, ces métriques n'apportent aucune information, outre un score, sur la nature des dégradations mesurées. Proposer une manière automatique d'évaluer la parole pathologique est essentiel pour pouvoir baser les stratégies de rééducation sur des évaluations objectives et non biaisées.

Récemment, des modèles auto-supervisés ont montré leur succès dans la capture de concepts phonétiques et de diverses caractéristiques de la parole. Dans (tom Dieck *et al.*, 2022), les auteurs ont montré que les modèles Wav2Vec2 sont capables d'apprendre certains concepts phonétiques et qu'ils modélisent correctement le lieu et le mode d'articulation. Certaines recherches se sont concentrées sur l'utilisation de ces modèles pour évaluer automatiquement le niveau de sévérité de la parole (Hernandez *et al.*, 2022; Favaro *et al.*, 2023; Yeo *et al.*, 2023a; Javanmardi *et al.*, 2024). Bien que la détection d'une maladie et son évaluation quantitative soit importante, nous pensons qu'il est tout aussi important d'expliquer les résultats de ces modèles. En effet, leur explicabilité permet de cibler leurs faiblesses, et ainsi d'accroître la confiance que les cliniciens peuvent avoir en ces systèmes. Jusqu'à présent, un nombre limité d'études se sont concentrées sur l'interprétabilité de ces modèles. Dans (Tu *et al.*, 2017), un modèle prédictif de la sévérité de la parole dysarthrique a été entraîné, en incorporant une couche *bottleneck* dans un réseau neuronal à couches entièrement connectées pour améliorer l'interprétabilité. En effet, un apprentissage par transfert a été utilisé pour apprendre des étiquettes interprétables cliniquement issues de la classification de Darley (Darley *et al.*, 1969). Leurs résultats montrent une amélioration de la précision de l'évaluation de la dysarthrie, avec des justifications basées sur des caractéristiques interprétables. Une extension de ce travail (Xu *et al.*, 2023) a évité les étiquettes perceptives (nécessitant une annotation experte et coûteuse en temps des enregistrements utilisés pour l'entraînement et les tests des modèles) en formant la couche interprétable autour de quatre caractéristiques acoustiques liées à la dysarthrie et extraites automatiquement. L'utilisation de SHAP (*SHapley Additive exPlanations* (Lundberg & Lee, 2017)) a permis d'analyser la contribution de chaque caractéristique acoustique à la prédiction finale. Des avancées récentes ont aussi démontré l'efficacité de l'utilisation de la métrique *Goodness of Pronunciation*, en utilisant la partie d'extraction de caractéristiques – figée – du modèle Wav2Vec2, suivi d'un classifieur de phonèmes (Yeo *et al.*, 2023b). Leur approche montre l'impact relatif de chaque phonème sur le score prédit de sévérité. Une autre méthodologie a été utilisée dans (Abderrazek *et al.*, 2023) en utilisant un réseau neuronal convolutif (CNN) pour la classification de phonèmes. La méthode NCD (*Neuro-Concept Detector*) liée à la métrique ANPS (*Artificial Neuron-based Phonological Similarity*) permettent une interprétation des scores de sévérité et d'intelligibilité prédits, en terme d'altérations de la parole produite par des patients, via l'association des neurones du classifieur aux traits phonétiques qu'ils détectent.

Notre travail vise à étendre cette interprétation en remplaçant le CNN par un modèle auto-supervisé

basé sur l'architecture complète de Wav2Vec2. Cette modification a pour objectif de fournir de meilleurs taux de bonne classification, mais aussi d'apporter ultérieurement une description plus nuancée et plus riche des caractéristiques phonétiques de la parole pathologique. En plus du changement d'architecture du modèle, ce travail examinera l'impact du choix des corpus de pré-entraînement des modèles Wav2Vec2 disponibles à la communauté, de la taille de ces modèles, de l'étape d'affinage et des corpus utilisés lors de celle-ci sur le choix d'un modèle auto-supervisé pour la tâche visée de classification de phonèmes. Cette exploration est cruciale pour optimiser les performances du modèle sur des corpus variés, contribuant à la robustesse et à la capacité de généralisation de notre approche. Ainsi, une analyse détaillée des confusions entre phonèmes sera menée pour valider la capacité de généralisation de nos modèles sur d'autres ensembles de données. Ensuite, nous analyserons la corrélation entre les taux de bonne classification de phonèmes et les mesures perceptives obtenues auprès d'experts sur la parole de patients ayant bénéficié de soins suite à un cancer des VADS. Cette approche d'évaluation multidimensionnelle fournira une évaluation complète du modèle auto-supervisé proposé, démontrant son potentiel pour une analyse précise et cliniquement pertinente de la parole pathologique.

2 Corpus

Pour entraîner et évaluer nos modèles pendant l'entraînement, nous nous sommes appuyés sur un sous-ensemble de BREF (Lamel *et al.*, 1991) et de Common Phone (Klumpp *et al.*, 2022). Pour tester nos modèles une fois affinés, nous nous sommes basés sur BREF-Int et C2SI (Astésano *et al.*, 2018).
Le corpus **BREF** est un corpus de 120 locuteurs français lisant des extraits du journal *Le Monde*. Les enregistrements ont eu lieu dans les années 90, avec des personnes recrutées dans la région parisienne. Un sous-ensemble équilibré en termes de phonèmes de ce corpus a été créé pour garantir que les modèles affinés ne présentent pas de biais envers un phonème spécifique. Il est composé de plus de trois millions de trames de 127 ms, chacune centrée sur un phonème français ou un silence. Nous utilisons également l'ensemble de données **BREF-Int**, un sous-ensemble également équilibré en termes de phonèmes que nous utiliserons en phase de test. Ces ensembles sont identiques à ceux utilisés dans (Abderrazek *et al.*, 2023).
Le corpus **Common Phone** est un corpus équilibré en termes de genre, multilingue et aligné phonétiquement, dérivé du projet Common Voice de Mozilla. Seuls les enregistrements en français ont été utilisés dans ce travail. Nous avons également équilibré cet ensemble de données en termes de phonèmes et de genre pour assurer l'impartialité de nos modèles affinés du point de vue de ces derniers.
Le corpus **C2SI** comprend 87 patients traités pour un cancer buccal ou oropharyngé, ainsi que 41 contrôles sains (HC). Les patients et les HC ont été enregistrés à l'IUCT Oncopole, à Toulouse, France. Les patients ont été confrontés à plusieurs tâches : /a/ tenu, lecture de phrases, lecture de passages de texte (LEC), production de pseudo-mots (DAP) ainsi que diverses autres tâches prosodiques. À l'aide des enregistrements de certaines de ces tâches, des experts ont évalué la sévérité (degré d'altération du signal de parole) et l'intelligibilité de la production de parole des patients sur une échelle de 0 - forte altération - à 10 - discours parfait ; l'intelligibilité étant définie ici comme "la capacité d'un auditeur à reconnaître les mots et/ou les sons de la parole produite par le locuteur" (Astésano *et al.*, 2018). Les enregistrements des tâches LEC et DAP sont utilisés pour tester nos modèles. Ils seront désignés ici par C2SI-LEC et C2SI-DAP. Pour correspondre à la sélection effectuée dans un travail précédent, nous testerons nos modèles sur 24 HC (parmi les locuteurs à disposition, tous n'ayant pas

réalisé les deux tâches C2SI-LEC et C2SI-DAP), tous enregistrés dans les mêmes conditions. Nous confronterons nos résultats aux évaluations perceptives incluant 82 patients parmi les 87 – 5 patients n'ont pas été évalués par les experts.
Le tableau 1 détaille le nombre de trames alignées sur les phonèmes utilisés pour chaque corpus, ainsi que le nombre d'heures qu'elles représentent. Étant donné que certaines trames se chevauchent, ce nombre d'heures est supérieur à la somme des durées des enregistrements. Cependant, cette valeur reflète ce que le modèle voit en entrée. Tous les corpus mentionnés ci-dessus sont alignés phonétiquement avec 31 phonèmes et un silence. Ces 31 phonèmes comprennent quatre archi-phonèmes : /Ê/ = {e, ɛ}, /Û/ = {œ, ø}, /Ô/ = {o, ɔ}, et /µ/ = {œ̃, ɛ̃}. L'utilisation de ces archi-phonèmes permet de neutraliser les oppositions de voyelles moyennes.

TABLE 1 – Usage, nombre de trames et durée totale des données audios pour chaque corpus utilisé.

Corpus	Usage	#trames	#heures
BREF	entraînement, validation	3,118k	110h
Common Phone	entraînement, validation	236k	8.3h
BREF-Int	test	85k	3h
C2SI-LEC (HC)	test	43k	1.5h
C2SI-DAP (HC)	test	73k	2.5h

3 Modèles

Le modèle de réseau neuronal convolutif choisi a été précédemment entraîné dans (Abderrazek *et al.*, 2023) sur le corpus BREF décrit dans la section 2 . Il est composé de deux couches de convolution combinées avec des couches de pooling maximales. L'entrée du modèle correspond à une fenêtre glissante de 11 trames acoustiques de 20ms espacées de 10ms, où chaque trame correspond à un vecteur *Mel-Filterbanks* extrait du signal audio, ainsi qu'à ses dérivées première et seconde. Le modèle a ainsi une fenêtre de 120ms centrée sur le phonème à prédire. Une fois le CNN appliqué, la sortie est ensuite aplatie avant d'être donnée en entrée de trois couches denses, détaillées ci-dessous. La sortie de ce modèle est notre *baseline*.

Concernant la partie Wav2Vec2, les modèles de LeBenchmark2.0 (Parcollet *et al.*, 2023) sont utilisés. Comme nous allons appliquer ces modèles sur de la parole pathologique de locuteurs francophones dans la suite de nos travaux, nous avons ciblé les modèles Wav2Vec2 pré-entraînés exclusivement sur du français. En effet, combiner les phonèmes de plusieurs langues pourrait compliquer l'analyse entre les phonèmes sains et pathologiques. Les modèles LeBenchmark existent avec différentes architectures (6, 12, 24 ou 48 couches cachées) ainsi que des tailles de corpus de pré-entraînement différentes. Dans ce travail, nous comparerons les résultats obtenus avec des modèles contenant 6, 12 ou 24 couches cachées, respectivement appelés *light*, *base* et *large*, ainsi que pré-entraînés sur 3k ou 14k heures de parole française. Wav2Vec2 fonctionne avec des fenêtres de 25ms, espacées approximativement toutes les 20ms, ce qui implique un recouvrement de 5ms environ. Afin d'émuler la fenêtre de 120ms du CNN, nous donnons à Wav2Vec2 des fichiers audio correspondant à six fenêtres superposées de 25ms. Une telle architecture nous donne une fenêtre de 127ms environ, qui est très proche de la durée du contexte utilisé pour le CNN.

La sortie des modèles CNN ou Wav2Vec2, une fois aplatie, passe au travers de trois couches denses de dimension 1024, dédiées à la tâche de classification en phonèmes. La taille de la couche d'aplatissement dépend de la taille de la sortie de l'encodeur utilisé (CNN ou Wav2Vec2, et la taille

du modèle Wav2Vec2 choisi). Le phonème de sortie est ensuite sélectionné en appliquant un softmax sur les 32 valeurs de sortie.

Les corpus utilisés pour la phase d'entraînement ont été répartis de manière aléatoire en deux sous-ensembles équilibrés en termes de phonèmes : entraînement (90% des données) et validation (10% restants). Tous nos modèles Wav2Vec2 ont été affinés en utilisant l'outil SpeechBrain durant 15 époques (choix empirique). Le modèle présentant le meilleur taux d'erreur phonétique sur le jeu de validation a été choisi pour l'inférence sur les ensembles de données de test. Le classifieur utilise un optimiseur Adadelta avec un taux d'apprentissage initial de 0.9, pour améliorer une *cross-entropy loss* appliquée sur la classification de phonèmes. L'architecture de Wav2Vec2 - lorsqu'elle est affinée - utilise un optimiseur Adam avec un taux d'apprentissage initial de 1.10^{-4}. Les recettes SpeechBrain sont disponibles publiquement sur un répertoire GitHub [1].

4 Résultats expérimentaux

4.1 Comparaison des modèles Wav2Vec2

Des expériences ont été menées pour étudier l'impact des facteurs suivants : (1) l'affinage de Wav2Vec2, (2) les données de pré-entraînement, (3) la taille du modèle, et (4) les données d'affinage. Le tableau 2 résume les taux de bonne classification obtenus pour chaque modèle entraîné. Etant donné que la distribution des phonèmes n'est pas équilibrée dans les jeux C2SI, les taux ci-dessous sont équilibrés par phonème. Mathématiquement, il s'agit de la moyenne des taux de bonne classification obtenus pour chaque phonème. Ainsi, même lorsque ce n'est pas précisé, les taux de bonne classification sont systématiquement équilibrés.

Pour analyser l'impact de l'affinage de Wav2Vec2 (1), deux modèles *14k-large* ont été utilisés, seulement l'un des deux ayant été affiné. En comparant ainsi les modèles *14k-large Frozen* et *14k-large*, nous pouvons observer que l'affinage améliore les taux de bonne classification sur les trois jeux de test, de manière significative (les intervalles de confiance ne se recouvrent pas). Ainsi, affiner un modèle Wav2Vec2 sur des données du même type est bénéfique au niveau de cette métrique.

Ensuite, l'impact des données de pré-entraînement (2) a été mesuré en utilisant des modèles LeBenchmark ayant différents corpus de pré-entraînement. Nous nous sommes appuyés sur les familles de modèles 3k et 14k, qui sont entraînés respectivement sur 3 000 et 14 000 heures de parole française. Ils incluent de la parole française lue, jouée, spontanée et professionnelle, avec du français neutre ou comportant un accent. Ces deux ensembles de données varient fortement au niveau de la variété d'accents présents. En effet, l'ensemble de pré-entraînement 3k contient principalement des livres audio et des émissions de radios françaises, avec moins de 1% de parole comportant un accent africain dont nous avons connaissance - les livres audio ne fournissent pas d'informations sur les accents de leurs locuteurs. Malheureusement, nous ne pouvons pas être certains que les locuteurs des radios françaises soient des locuteurs francophones. Cependant, l'ensemble de pré-entraînement 14k contient environ 4700 heures de parole issue du Parlement Européen, qui présentent au minimum trois accents supplémentaires - belge, suisse et italien d'Aoste, ainsi que des quantités plus négligeables de radios africaines, qui incluent des accents maliens et nigériens. D'après nos résultats, nous pouvons observer que l'ajout d'une variabilité linguistique au travers de différents accents français, n'apporte pas d'amélioration significative des taux de bonne classification.

1. github.com/MaloMn/wav2vec2-phone-classification

TABLE 2 – Taux de bonne classification équilibrés (en %) obtenus sur chaque jeu de test. Les intervalles de confiances sont calculés avec une approche Bootstrap (Ferrer & Riera, 2024). Les taux obtenus pour le CNN ont été repris des travaux de (Abderrazek *et al.*, 2023), et ne présentaient pas d'intervalle de confiance.

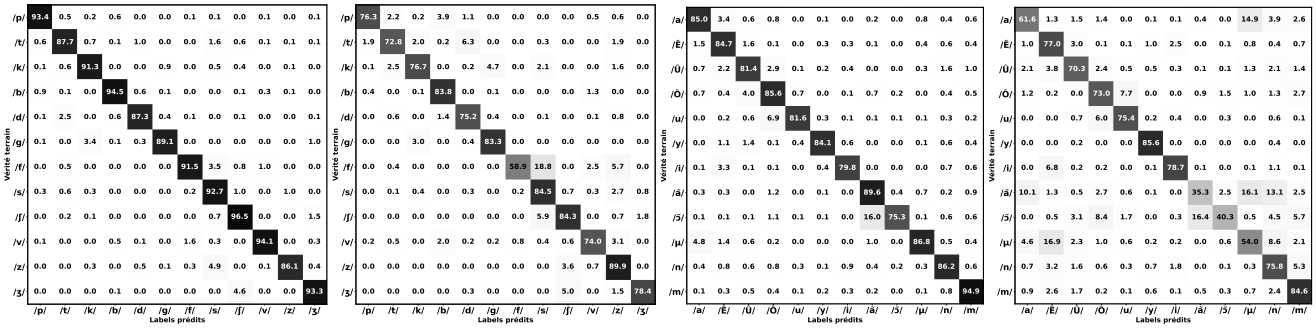
Modèle	Jeu(x) d'affinage	BREF-Int ↑	C2SI-LEC ↑ (locuteurs HC)	C2SI-DAP ↑ (locuteurs HC)
CNN, *Baseline*	BREF	81.4	72.2	69.2
14k-large Frozen	BREF	83.5±0.2	66.9±0.6	66.9±0.4
14k-large	BREF	87.6±0.2	70.2±0.6	70.6±0.4
14k-light	BREF	81.8±0.2	57.0±0.6	57.3±0.4
3k-large	BREF	**88.3±0.2**	70.6±0.6	71.3±0.4
3k-base	BREF	84.9±0.2	48.1±0.6	50.1±0.4
14k-large	BREF, CP	87.4±0.2	72.1±0.5	73.3±0.4
14k-light	BREF, CP	82.9±0.3	64.1±0.6	63.7±0.4
3k-large	BREF, CP	**88.3±0.2**	**72.6±0.6**	**73.9±0.4**
3k-base	BREF, CP	84.9±0.2	61.4±0.6	62.5±0.4

En ce qui concerne la taille du modèle, plusieurs tailles de modèles LeBenchmark (3) ont été testées : *light*, *base* et *large*, avec respectivement 6, 12 et 24 couches cachées. Nos résultats montrent qu'utiliser un modèle *large* (avec 24 couches cachées) offre de meilleurs taux de bonne classification que des modèles plus petits (6 et 12 couches cachées), et ce de manière significative. Malheureusement, comme LeBenchmark ne propose ni de modèles 14k-base et 3k-light, nous ne pouvons pas comparer les performances des modèles 3k-base et 14k-light. Néanmoins, nos résultats semblent montrer que les modèles plus petits ont un pouvoir de généralisation plus faible que les plus gros modèles sur nos jeux de données C2SI non vus des modèles.

Enfin, pour analyser l'impact des jeux d'affinage (4), nous ajoutons un jeu d'entraînement supplémentaire. Le travail antérieur utilisait exclusivement l'ensemble de données BREF pour entraîner l'architecture CNN. Tous les modèles mentionnés ci-dessus ont donc été ré-entraînés en combinant les corpus BREF et Common Phone. D'après les résultats obtenus, l'ajout de discours lu provenant d'autres corpus dans le processus d'affinage améliore la généralisation aux deux ensembles de données C2SI sur tous les modèles que nous avons affinés de manière significative. Les taux de bonne classification sont également similaires sur l'ensemble de données BREF-Int : les intervalles de confiance se chevauchent tous, sauf pour le modèle *14k-light*, qui est significativement meilleur une fois affiné sur la combinaison de BREF et Common Phone. Ainsi, il semble intéressant d'inclure d'autres ensembles de données à l'étape d'affinage, même s'ils ne représentent pas un pourcentage important des données d'entraînement.

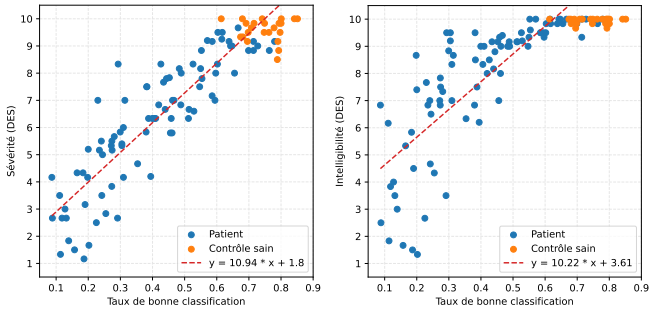
4.2 Wav2Vec2 et CNN

Le meilleur modèle Wav2Vec, par rapport aux résultats évoqués ci-dessus, nous permet d'améliorer de 6.9% les taux de bonne classification sur BREF-Int par rapport au CNN. Sur C2SI-LEC, la différence n'est pas significative. En revanche, on observe une différence significative sur C2SI-DAP, avec une amélioration de 4.7%. En comparant le nombre de paramètres des modèles que nous manipulons, il est intéressant de noter que le meilleur modèle Wav2Vec2 compte 330 millions de paramètres, contre 10 millions pour le CNN. Aussi, les modèles Wav2Vec2 plus petits *base* et *light* – ayant



(a) BREF-Int, obstruants (b) C2SI-LEC, obstruants (c) BREF-Int, vocaliques (d) C2SI-LEC, vocaliques

FIGURE 1 – Matrices de confusion obtenues sur les jeux BREF-Int – figures 1a et 1c – et C2SI-LEC – figures 1b et 1d – (locuteurs HC uniquement).



(a) Sévérité (DES), corrélation r = 0.90 (b) Intelligibilité (DES), corrélation r = 0.80

FIGURE 2 – Nuages de points des patients et HC C2SI suivant les niveaux associés aux évaluations perceptives des experts et les taux de bonne classification équilibrés obtenus.

respectivement 90 et 26 millions de paramètres – ne parviennent pas à généraliser aussi bien que le CNN sur les jeux C2SI. Cet écart est important, mais ces deux types de modèles ne présentent pas non plus la même architecture, et une comparaison basée uniquement sur le nombre de paramètres serait biaisée. Néanmoins, cette différence de taille implique que l'empreinte carbone de nos inférences sera nécessairement plus élevée lors de l'utilisation de modèles Wav2Vec2. A propos des jeux de données utilisés, le meilleur modèle a été entraîné sur davantage de données que le CNN. Nos résultats montrent aussi qu'utiliser uniquement BREF lors de l'affinage ne permet pas de généraliser aussi bien que le CNN sur les jeux C2SI.

Pour nous assurer que nos modèles ne sur-ajustent pas certains phonèmes, et que les confusions restent explicables, des matrices de confusion ont été générées sur BREF-Int et C2SI-LEC, en utilisant le modèle *3k-large* affiné sur BREF et Common Phone. Les figures 1a à 1d présentent des parties spécifiques de ces matrices, dédiées aux phonèmes obstruants et vocaliques. Le choix de restreindre notre analyse à ces phonèmes nous permet de réaliser une comparaison directe avec l'analyse effectuée dans (Abderrazek *et al.*, 2023). La comparaison entre les résultats obtenus et ceux obtenus précédemment montre que Wav2Vec2 diminue dans la plupart des cas les confusions observées. Sur les phonèmes obstruants, /ʒ/ était confondu avec /ʃ/ dans 9% des cas sur BREF-Int sur le précédent travail, contre 4.6% maintenant. La confusion entre ces deux phonèmes est donc toujours présente, mais elle survient dans la moitié des cas par rapport à précédemment. Cette diminution se retrouve aussi sur le jeu C2SI : alors que /p/ était précédemment confondu avec /t/ dans 9.3% des cas, ce n'est le cas que 2.2% du temps avec notre architecture. Nous retrouvons également ici

les mêmes causes pour les confusions importantes : soit la perte du lieu d'articulation (caractère aigu – /f/ → /s/, caractère compact – /ʃ/ → /s/), soit la confusion du voisement – /t/ → /d/. Sur les phonèmes vocaliques, nous retrouvons également de fortes confusions liées aux voyelles orales et aux consonnes nasales sur le jeu C2SI-LEC. Là où /ã/ et /a/ étaient confondus dans 11.5% des cas, ils le sont toujours dans 10.1% des cas. /ã/ est aussi davantage confondu maintenant avec /n/ (13.1%, contre 8.4% précédemment), et est moins confondu avec /m/ (2.5%, contre 6.2% précédemment). Ces confusions, expliquées précédemment par le changement d'accents des locuteurs (Parisien.ne.s pour BREF-Int, et Toulousain.e.s pour C2SI) se retrouvent ici, avec l'utilisation d'un autre modèle. Nos résultats viennent donc appuyer les résultats précédemment obtenus, et montrent que les données utilisées pour l'affinage et les différences de domaine entre jeux de données restent une problématique non négligeable pour ce type d'architecture. Sur BREF-Int, nous retrouvons également une confusion forte entre /ã/ et /c̃/, avec une confusion mutuelle de 16.4%, mais moins intense qu'en utilisant un CNN – 20.6%. Ces résultats sont importants car ils montrent que notre modèle est aussi sensible aux prononciations atypiques (ici, un accent régional), ce qui est souhaitable lors de l'analyse de la parole pathologique.

4.3 Application à la parole pathologique

La robustesse de notre modèle, ainsi que sa capacité à généraliser à d'autres jeux de données ayant été montrées, nous allons observer si les taux de bonne classification peuvent corréler positivement avec les évaluations des 6 experts de C2SI. Les jugements des experts, en termes de scores de sévérité et d'intelligibilité sur la tâche de description d'image (DES) ont été moyennés, puis comparés aux taux de bonne classification. Les figures 2a et 2b comparent sous la forme de nuages de points les scores perceptifs et les taux de bonne classification associés aux enregistrements de parole des patients et HC du corpus C2SI. Les courbes de régression linéaire ont également été tracées. Les coefficients de corrélation de Pearson élevés – 0.90 avec la sévérité et 0.80 avec l'intelligibilité – confirment que les mesures perceptives peuvent être estimées à l'aide des taux de bonne classification équilibrée des phonèmes de notre modèle *3k-large* affiné sur BREF et Common Phone. Ces valeurs sont semblables à celles obtenues précédemment avec le CNN (Abderrazek *et al.*, 2023) dont les taux de bonne classification en phonèmes corrélaient à 0.91 et 0.81 avec la sévérité et l'intelligibilité respectivement. Ces derniers résultats confirment qu'une représentation de la parole basée sur un modèle de type Wav2Vec2 convient bien à une analyse phonétique de la parole pathologique.

5 Conclusion

Dans ce travail, nous avons montré qu'un modèle basé sur Wav2Vec2 surpasse un CNN dans la classification des phonèmes, tout en préservant certaines spécificités linguistiques, telles qu'un accent régional. Nos résultats valident non seulement l'efficacité de l'approche basée sur Wav2Vec2, mais soulignent également l'importance de prendre en compte l'architecture du modèle et la diversité des données d'entraînement pour des performances optimales. Les travaux futurs incluent l'application du concept de NCD développé dans (Abderrazek *et al.*, 2023) pour analyser les couches cachées de notre architecture affinée, et analyser comment ce nouveau modèle influence la détection des traits phonétiques, qui sont cruciaux pour expliquer les phonèmes prédits. À son tour, cette interprétabilité est nécessaire pour avoir une analyse objective de la parole d'un patient, ce qui améliorerait les techniques de rééducation mises en place par les cliniciens.

Références

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2023). Interpreting deep representations of phonetic features via neuro-based concept detector : Application to speech disorders due to head and neck cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**, 200–214. DOI : 10.1109/TASLP.2022.3221039.

ASTÉSANO C., BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., LAARIDH I., LALAIN M., LEPAGE B., MAUCLAIR J., NOCAUDIE O., PINQUIER J., PONT O., POUCHOULIN G., PUECH M., ROBERT D., SICARD E. & WOISARD V. (2018). Carcinologic speech severity index project : A database of speech disorder productions to assess quality of life related to speech after cancer. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, **12**(3), 462–496. DOI : 10.1044/jshr.1203.462.

FAVARO A., TSAI Y.-T., BUTALA A., THEBAUD T., VILLALBA J., DEHAK N. & MORO-VELÁZQUEZ L. (2023). Interpretable speech features vs. dnn embeddings : What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, **166**, 107559. DOI : https ://doi.org/10.1016/j.compbiomed.2023.107559.

FERRER L. & RIERA P. (2024). Confidence Intervals for evaluation in machine learning.

HERNANDEZ A., PÉREZ-TORO P. A., NOETH E., OROZCO-ARROYAVE J. R., MAIER A. & YANG S. H. (2022). Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2022*, p. 51–55. DOI : 10.21437/Interspeech.2022-10674.

JAVANMARDI F., KADIRI S. R. & ALKU P. (2024). Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, p. 103047. DOI : https ://doi.org/10.1016/j.specom.2024.103047.

KLUMPP P., ARIAS T., PÉREZ-TORO P. A., NOETH E. & OROZCO-ARROYAVE J. (2022). Common phone : A multilingual dataset for robust acoustic modelling. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 763–768, Marseille, France : European Language Resources Association.

LAMEL L. F., GAUVAIN J.-L., ESKÉNAZI M. *et al.* (1991). Bref, a large vocabulary spoken corpus for french. *Eurospeech'91, Italy*, **22**(28), 50.

LUNDBERG S. M. & LEE S.-I. (2017). A unified approach to interpreting model predictions. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTEVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2023). Lebenchmark 2.0 : a standardized, replicable and enhanced framework for self-supervised representations of french speech.

TOM DIECK T., PÉREZ-TORO P. A., ARIAS T., NOETH E. & KLUMPP P. (2022). Wav2vec behind the Scenes : How end2end Models learn Phonetics. In *Proc. Interspeech 2022*, p. 5130–5134. DOI : 10.21437/Interspeech.2022-10865.

TU M., BERISHA V. & LISS J. (2017). Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Proc. Interspeech 2017*, p. 1849–1853. DOI : 10.21437/Interspeech.2017-1222.

XU L., LISS J. & BERISHA V. (2023). Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, **3**(1), 015201. DOI : 10.1121/10.0016833.

YEO E. J., CHOI K., KIM S. & CHUNG M. (2023a). Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5. DOI : 10.1109/ICASSP49357.2023.10094605.

YEO E. J., CHOI K., KIM S. & CHUNG M. (2023b). Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification. In *Proc. INTERSPEECH 2023*, p. 166–170. DOI : 10.21437/Interspeech.2023-173.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Audiocite.net: Un grand corpus d'enregistrements vocaux de lecture en Français

Soline Felice* Solène Evain Solange Rossato François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
solene.evain, solange.rossato, francois.portet@univ-grenoble-alpes.fr,
soline.felice@univ-tlse2.fr

RÉSUMÉ

L'arrivée de l'apprentissage auto-supervisé dans le domaine du traitement automatique de la parole a permis l'utilisation de grands corpus non étiquetés pour obtenir des modèles pré-appris utilisés comme encodeurs des signaux de parole pour de nombreuses tâches. Toutefois, l'application de ces méthodes de SSL sur des langues telles que le français s'est montrée difficile due à la quantité limitée de corpus de parole du français publiquement accessible. C'est dans cet objectif que nous présentons le corpus Audiocite.net comprenant 6682 heures d'enregistrements de lecture par 130 locuteurs et locutrices. Ce corpus est construit à partir de livres audio provenant du site *audiocite.net*. En plus de décrire le processus de création et les statistiques obtenues, nous montrons également l'impact de ce corpus sur les modèles du projet LeBenchmark dans leurs versions 14k pour des tâches de traitement automatique de la parole.

ABSTRACT

Audiocite.net : A Large Spoken Read Dataset in French

The advent of self-supervised learning (SSL) in speech processing has allowed the use of large unlabeled datasets to learn pre-trained models, serving as powerful encoders for various downstream tasks. However, the application of these SSL methods to languages such as French has proved difficult due to the scarcity of large French speech datasets. To advance the emergence of pre-trained models for French speech, we present the Audiocite.net corpus composed of 6 682 hours of recordings from 130 readers. This corpus is built from audiobooks from the *audiocite.net* website. In addition to describing the creation process and final statistics, we also show how this dataset impacted the models of LeBenchmark project in its 14k version for speech processing downstream tasks.

MOTS-CLÉS : Ensembles de données vocales, apprentissage auto-supervisé, traitement automatique de la parole.

KEYWORDS: Spoken Datasets, French Speech, Self Supervised Learning, Automatic Speech Processing.

1 Introduction

L'arrivée de l'apprentissage auto-supervisé (*Self-Supervised Learning* – SSL) dans le domaine du traitement automatique de la parole a permis l'utilisation de grands corpus non étiquetés pour obtenir

∗. maintenant à l'IRIT, Univ. Toulouse 2 Jean Jaurès.

des modèles pré-appris.

De nombreux modèles profonds pré-appris modélisant le signal acoustique de la parole ont émergé utilisant l'apprentissage auto-supervisé génératif (PASE+ (Ravanelli *et al.*, 2020), Mockingjay (Liu *et al.*, 2020)) ; une fonction de coût contrastive (CPC (Oord *et al.*, 2019), Speech SimCLR (Jiang *et al.*, 2021), Wav2Vec 2.0 (Baevski *et al.*, 2020)) ; ou prédictive (HuBERT (Hsu *et al.*, 2021), wavLM (Chen *et al.*, 2022), data2vec (Baevski *et al.*, 2022)) (Abdel-Rahman *et al.*, 2022). Ces modèles ont fait avancer les performances du traitement de la parole en les adaptant et utilisant comme encodeurs pour les tâches de traitement automatique de la parole (*downstream tasks*). Par exemple, Wav2Vec 2.0 a pu atteindre des résultats à l'état de l'art avec un modèle pré-appris puis ajusté avec un minimum de données étiquetées pour une tâche de reconnaissance automatique de la parole (RAP) dans un contexte de lecture en anglais.

Les modèles pré-appris par SSL dépendent fortement de la disponibilité d'une grande quantité de données d'apprentissage. Bien que plusieurs grands ensembles de données pour l'anglais et multilingues ont été publiés, des ensembles de données aussi volumineux pour le français sont rares. Jusqu'à récemment, il était difficile de trouver de grands ensembles de données de parole française disponibles publiquement (à l'exception des 1,700 heures de parole transcrites automatiquement d'EPAC par (Estève *et al.*, 2010)). Récemment, de grands corpus multilingues incluant le français ont été rendus disponibles, tels que MLS (1,096 h) (Pratap *et al.*, 2020), ou VoxPopuli (non transcrit +4,500 h) (Wang *et al.*, 2021). Cependant, ces ensembles de données représentent toujours une quantité bien inférieure à ce qui est disponible pour l'anglais. Le multilinguisme a été souligné comme un moyen de traiter les langues sous-dotées, mais l'étude menée dans le projet LeBenchmark (Parcollet *et al.*, 2024) visant à créer des modèles de parole pré-appris pour le français a montré que les modèles entraînés sur des données cibles monolingues sont bien plus efficaces que ceux multilingues.

Dans cet article, nous présentons Audiocite.net, un corpus non transcrit d'environ 6 600 heures d'enregistrements de parole lue en français, disponible librement pour la communauté. Nous décrivons la sélection et l'acquisition des données (voir sec. 2) ainsi que les principales caractéristiques de la publication du jeu de données (sec. 3). Nous montrons également comment ce jeu de données a impacté le modèle de 14k du projet LeBenchmark 2.0 (Parcollet *et al.*, 2024) pour certaines tâches de traitement de la parole, notamment la reconnaissance automatique de la parole lue (sec. 4).

Cet article est l'adaptation vers le français d'un article publié à LREC/COLING 2024 (Felice *et al.*, 2024).

2 Sélection et acquisition des données

Les grands corpus de parole proviennent souvent de l'extraction de contenus libres sur le web, comme Librispeech, extrait du projet LibriVox (Panayotov *et al.*, 2015). Cependant, les oeuvres publiées en France ne deviennent disponibles dans le domaine public que 70 ans après le décès de leurs auteurs et autrices, rendant les livres plus modernes publiés après 1953 inaccessibles. Pour surmonter cette limitation, l'initiative Common Voice (Ardila *et al.*, 2020) a été mise en place par la fondation Mozilla pour capturer la parole lue en utilisant des phrases collectées sur le web. En quatre ans, environ 1,100 heures de segments de phrases ont été collectées en Français. Pour rassembler un plus grand volume de parole lue continue, allant de la littérature classique à la moderne et librement accessible, nous avons décidé de nous concentrer sur le site *audiocite.net*.

Catégorie	# Fichiers				Durée (hh :mm :ss)			
	All	Train	Dev	Test	All	Train	Dev	Test
animaux	**160**	108	31	21	**26 :49 :04**	16 :12 :46	05 :33 :54	05 :02 :23
juniors	**35**	18	7	10	**04 :56 :39**	01 :45 :14	00 :41 :59	02 :29 :24
charme	**166**	166	0	0	**33 :02 :05**	33 :02 :05	00 :00 :00	00 :00 :00
contes	**2430**	1711	415	304	**490 :40 :12**	325 :44 :09	94 :56 :54	69 :59 :08
cuisine	**39**	35	3	1	**02 :44 :47**	02 :19 :18	00 :13 :37	00 :11 :51
documents	**1494**	1191	181	122	**367 :17 :28**	265 :35 :35	58 :22 :12	43 :19 :41
histoire	**1341**	1167	99	75	**397 :33 :01**	333 :51 :58	33 :19 :12	30 :21 :50
nouvelles	**2772**	1721	534	517	**721 :09 :55**	375 :11 :21	167 :15 :11	178 :43 :21
philosophies	**1052**	773	53	226	**181 :04 :51**	117 :50 :07	17 :02 :07	46 :12 :36
planete-actuelle	**145**	145	0	0	**18 :27 :20**	18 :27 :20	00 :00 :00	00 :00 :00
poesies	**2274**	1956	160	158	**116 :09 :42**	84 :37 :27	14 :41 :07	16 :51 :07
religions	**777**	777	0	0	**213 :21 :47**	213 :21 :47	00 :00 :0	00 :00 :0
romans	**14664**	13088	713	863	**3 943 :54 :09**	3 377 :46 :53	267 :58 :14	298 :09 :01
science-fiction	**478**	349	117	12	**122 :03 :39**	87 :48 :10	28 :16 :00	05 :59 :28
theatre	**658**	603	19	36	**42 :45 :29**	36 :58 :06	02 :35 :49	03 :11 :34
Tout	**28 485**	23 808	2 332	2345	**6 682 :00 :18**	5 290 :32 :24	690 :56 :22	700 :31 :31

TABLE 1 – Statistiques (durées et nombre de fichiers) par catégorie de livres avec les détails des partitions (**all**, train / dev / test)

Audiocité est une association à but non lucratif qui met à disposition une plate-forme où les bénévoles peuvent partager leurs lectures. Avant leur première contribution, les bénévoles doivent passer un test de lecture afin d'évaluer leur prononciation, leur rythme de lecture, les conditions d'enregistrement et le format final du fichier audio. Des conseils de post-traitement sont fournis selon les besoins (par exemple, pour réduire les bruits de respiration). Les enregistrements comprennent environ 5 000 livres audio d'œuvres littéraires classiques issues du domaine public en français (Balzac, Hugo, Maupassant, Molière…) et environ 700 livres audio d'auteurs et autrices de l'époque contemporaine qui ont choisi de partager librement leurs oeuvres (Brussolo, Huchon, Del, Martin, Fée …).

2.1 Critères de sélection des livres audio

Pour qu'un corpus soit utile et serve la recherche reproductible, il devrait idéalement être à la fois accessible et gratuit. Sur le site web, tous les livres audio sont distribués sous une licence Creative Commons, car, avant de déposer leur enregistrement, les personnes doivent prendre en compte les droits d'auteur qui confèrent à l'auteur ou autrice du livre les droits exclusifs d'utiliser, de copier, de licencier, d'exécuter et de modifier l'œuvre. Ainsi, les lecteurs et lectrices sont autorisés à lire des livres du domaine public (c'est-à-dire sans droits d'auteur) ou des livres contemporains dont les auteurs et autrices ont donné l'autorisation d'enregistrer leur texte et de le distribuer sur *audiocite.net* sous une licence spécifique.

2.2 Processus de téléchargement

Les données ont été collectées en deux étapes en novembre 2021. L'administrateur du site web a aimablement donné son aval. Dans un premier temps, tout le catalogue du site a été extrait afin de collecter toutes les métadonnées concernant chaque livre audio (œuvre originale, sujet, auteurs/autrices, lecteurs/lectrices, licence. . .). Dans un second temps, tous les fichiers audio ont été téléchargés ce qui a pris environ une semaine. Les fichiers téléchargés étaient soit au format MP3, soit dans des archives zip. Par la suite, les enregistrements audio qui ne répondaient pas aux critères d'une durée supérieure à 5 secondes ou qui étaient fournis sans licence permettant leur utilisation ont été retirés. Quelques livres audio avec des URL défectueuses ont également été exclus.

2.3 Données recueillies

Au total, 6,682 heures de livres audio lus par 130 locuteurs et locutrices, dont 70 hommes (62%), 51 femmes (34%) et 9 personnes dont le genre n'a pu être identifié (4%) ont été collectées. Cela correspond à une taille totale de 340 Go de fichiers audio et de métadonnées. Le tableau 1 indique le nombre de fichiers et la durée des fichiers audio pour chaque catégorie de livre.

Fichiers audio : Il convient de mentionner que sur les 4,378 livres audio, 388 étaient directement hébergés par *audiocite.net*, tandis que les 3,990 restants étaient hébergés sur des instances de *archive.org*. Contrairement aux lignes directrices données, les enregistrements peuvent être mono ou stéréo, et des variations dans les débits binaires ou les taux d'échantillonnage peuvent également être constatés. Certains enregistrements peuvent contenir une succession de plusieurs locuteurs et locutrices et des bruits de fond ou de la musique. En outre, tous les enregistrements n'impliquent pas nécessairement la lecture de livres publiés ; certains sont des articles ou des podcasts.

Métadonnées : Parallèlement aux fichiers audio, des métadonnées ont été téléchargées, telles que la durée de chaque fichier audio, l'identifiant du locuteur ou de la locutrice, le titre du livre lu, l'auteur ou l'autrice du livre, la catégorie du livre et la licence liée à l'audio. Ces informations ont été fournies par les locuteurs et locutrices eux-mêmes sur la page web de chaque livre audio de *audiocite.net*.

3 Organisation du corpus Audiocite.net

Bien que l'un des principaux usages envisagés pour ce jeu de données soit l'apprentissage auto-supervisé, nous anticipons également d'autres utilisations. En effet, même s'il n'est pas transcrit, le jeu de données pourrait être utilisé pour la modélisation thématique, la reconstruction de signaux ou la synthèse vocale. C'est pourquoi nous l'avons publié avec des partitions officielles et des fichiers de métadonnées faciles à interroger.

3.1 Estimation du genre des locuteurs et locutrices

Pour minimiser les biais dans les partitions, nous avons déduit le genre des locuteurs et locutrices en fonction de leurs identifiants ou en écoutant leurs voix. Cette information a été ajoutée aux métadonnées. Cependant, nous ne garantissons pas que l'information soit fiable ni que la méthode

utilisée soit viable pour déduire le genre d'une personne puisqu'elle n'est pas basée sur l'auto-identification de celle-ci. Les métadonnées concernant le genre doivent être traitées avec prudence.

3.2 Partitions des données

Le jeu de données a été divisé en trois partitions : une partition d'apprentissage (train) comprenant 80% des enregistrements, une partition de développement (dev) en comprenant 10%, et une partition de test (test) comprenant les 10% restants. Le tableau 2 fournit les statistiques de durée par genre pour chaque sous-ensemble.

# Personnes	Durée Totale	Durée Moyenne	# Fichiers
TRAIN			
74 T	5290 :32 :24	00 :13 :19	23808
35 F	1577 :23 :53	00 :16 :54	5600
30 H	3431 :01 :21	00 :11 :52	17329
9 I	282 :07 :09	00 :19 :15	879
DEV			
78 T	690 :56 :22	00 :17 :46	2332
44 F	344 :14 :51	00 :15 :40	1317
34 H	346 :41 :31	00 :20 :29	1015
TEST			
61 T	700 :31 :31	00 :17 :55	2345
38 F	350 :39 :38	00 :15 :39	1344
23 H	349 :51 :53	00 :20 :58	1001
ALL			
130 T	6682 :00 :18	00 :14 :04	28485
70 F	2272 :18 :23	00 :16 :30	8261
51 H	4127 :34 :45	00 :12 :48	19345
9 I	282 :07 :09	00 :19 :15	879

TABLE 2 – Statistiques du corpus Audiocite.net - Nombre de fichiers, de personnes (locuteurs/locutrices) et durée par genre (tous, femme, homme et inconnu) par partitions

Les partitions de développement et de test ont été conçues pour ne pas inclure de contenu potentiellement sensible, spécifiquement ceux relevant des catégories *charmes* (érotique), *planete-actuelle* (géopolitique) et *religion*. De plus, pour ces partitions, une représentation égale de la parole masculine et féminine a été assurée et les fichiers dont le genre de la personne était inconnu n'ont pas été inclus. Le tableau 1 indique le nombre de fichiers et la durée pour chaque catégorie de livre pour les partitions d'apprentissage, de développement et de test.

3.3 Organisation de l'ensemble de données

Le jeu de données est organisé comme suit : nous partageons un fichier README et une fiche technique (inspirée par *Datasheet for datasets*, (Gebru *et al.*, 2021)) où l'on peut trouver des statistiques détaillées, ainsi que des précisions sur la composition, l'utilisation et la distribution du corpus, et trois dossiers (`wavs/`, `scripts/` et `metadata/`). Dans le dossier `wavs/`, les fichiers de livres audio sont rangés dans des dossiers selon le titre du livre lu, triés par ordre alphabétique. Nous fournissons également un dossier `scripts/` avec des scripts pour générer des statistiques sur le corpus et les fichiers json fournis. Concernant le dossier `metadata/`, deux types de fichiers de métadonnées sont partagés avec le jeu de données : *.csv* et *.json*.

Fichier download.csv : Chaque livre audio possède une entrée dans le fichier csv. Nous fournissons également des informations tels que l'identifiant du locuteur ou de la locutrice, le titre du livre lu, l'auteur ou l'autrice du livre, le genre du livre, la licence de l'enregistrement, l'adresse URL du livre audio sur *audiocite.net* et le chemin vers le fichier audio dans le dossier *wavs*.

Fichiers json : Nous fournissons quatre fichiers json : *train.json*, *dev.json*, *test.json* et *all.json*, ce dernier étant une concaténation des trois précédents. Dans ces fichiers, une entrée correspond à un fichier audio (un livre audio peut contenir plusieurs fichiers audio). Ces fichiers contiennent l'identifiant du locuteur ou de la locutrice, la durée du fichier audio en secondes, le chemin d'accès vers le fichier dans le dossier *wavs* et le genre du locuteur ou de la locutrice (F/M/U). Une entrée json prend la forme suivante :


```
"Raiponce.mp3": {
 "path": "../wavs/Raiponce/Raiponce.mp3",
 "trans": "",
 "duration": 471.552,
 "spk_id": "Demelza",
 "spk_gender": "F"
},
```


4 Impact de Audiocite.net sur les tâches de traitement automatique de la parole

Le jeu de données collecté a été partagé avec l'équipe LeBenchmark pour entraîner leurs modèles 14k (Parcollet *et al.*, 2024). Dans cette section, nous avons comparé la performance du modèle 14k à celle du modèle 7k, qui n'a pas été entraîné sur le corpus. Nous rapportons les taux d'erreur de mots (WER) pour les systèmes de reconnaissance automatique de la parole (RAP) sur un jeu de données du même type de parole et de situation (livre audio), mais aussi les résultats de reconnaissance automatique de la parole et de vérification de locuteur issus de Parcollet *et al.* (2024).

4.1 Modèles LeBenchmark

Modèle 7k-large : Ce modèle a été entraîné sur 7,000 h de parole, incluant 1,626 h de radio, 1,115 h de parole lue, 127 h de parole spontanée, 38 h de dialogue téléphonique joué et 29 h de parole émotionnelle jouée.

Modèle 14k-large : Ce modèle a été entraîné sur 14,000 h de parole contenant les données utilisées pour le modèle 7k plus l'intégralité des données de Audiocite.net (toutes les partitions), ainsi que 111 h de parole issues de diffusions radiophoniques.

4.2 Expérimentations de Reconnaissance Automatique de la Parole (RAP)

En utilisant la recette CTC (*Connectionist Temporal Classification*) de SpeechBrain pour Common Voice [1], nous avons composé un système de RAP avec LeBenchmark (Wav2Vec2) en encodeur suivi d'une couche de BiLSTM et d'une dernière couche de DNN. Nous avons utilisé les partitions

1. https://github.com/speechbrain/speechbrain/tree/develop/recipes/CommonVoice/ASR/CTC

officielles de la partie française du corpus *Multilingual Librispeech* (MLS), allouant 1,076.58 h pour l'apprentissage, 10.07 h pour le développement et 10.07 h pour les tests. Deux scénarios ont été utilisés dans l'expérience : (1) avec l'encodeur LeBenchmark figé (c'est-à-dire sans ajustement) ou (2) avec ajustement de l'encodeur (c'est-à-dire avec ajusté en même temps que le modèle de RAP). Pour l'apprentissage de la RAP, les taux d'apprentissage ont été initialisés à 0,001 pour le modèle Wav2Vec2 et à 0,1 pour la partie BiLSTM+DNN, et un recuit a été utilisé avec des facteurs de 0,9 et 0,8 respectivement. Dans le cas de l'ajustement complet, le gradient n'est propagé dans l'encodeur LeBenchmark qu'après 500 étapes. La taille du lot d'apprentissage était de 8 et la dimension de la couche de sortie était de 43 (nombre de caractères dans l'ensemble d'apprentissage).

	WER (%) ↓	
Encodeur	**Figé**	**Ajusté**
7K-*large*	31.71	9.56
14K-*large*	9.96	9.98

TABLE 3 – Résultats de RAP (WER %) sur l'ensemble de test de la partie française du corpus MLS pour les systèmes de RAP avec encodeur LeBenchmark figé ou ajusté

Le tableau 3 résume les résultats de l'expérience. L'ajustement de la partie encodeur conduit très rapidement à des performances similaires pour les modèles 14k et 7k, indiquant que Audiocite.net n'a pas beaucoup d'impact lorsque l'encodeur est ajusté. Cependant, pour l'expérience avec l'encodeur figé, il y a une nette supériorité du modèle 14k indiquant que Audiocite.net a joué un rôle important dans la modélisation de la parole lue.

4.3 Expérimentations LeBenchmark

Parmi les différentes expériences réalisées par l'équipe LeBenchmark, nous rapportons dans les tableaux 4 et 5 les résultats des tâches de reconnaissance automatique de la parole (RAP) et de vérification du locuteur issues de (Parcollet *et al.*, 2024).

Comme on peut le voir dans les expériences de reconnaissance automatique de la parole (RAP) avec Common Voice et ETAPE, Audiocite.net (14K-large) n'apporte aucune amélioration par rapport au modèle 7K. Il est même dégradé sur ETAPE qui est composé de discours radiophoniques, un type de parole très différent de Audiocite.net. Cependant, dans une tâche de vérification de locuteur sur le corpus Fabiole (Ajili *et al.*, 2016) constitué de discours d'émissions de radio et de télévision, Audiocite.net (14K-large) apporte une nette amélioration par rapport au modèle 7K. Il semble que l'ajout de locuteurs et de locutrices dans le 14K ait amélioré ce type de modélisation.

	WER (%) ↓	
Encodeur	**Common Voice**	**ETAPE**
7K-*large*	**9.39**	**23.46**
14K-*large*	9.83	26.03

TABLE 4 – Résultats de RAP (WER %) de Parcollet *et al.* (2024) sur les partitions de test de Common Voice 6.1 et ETAPE, avec des modèles Wav2Vec2.0 ajustés sur des données de RAP étiquetées

Encodeur	EER	minDCF^{-10} ↓	minDCF^{-100} ↓
7K-*large*	5.228	0.3833	0.5754
14K-*large*	**3.535**	**0.2965**	**0.4801**

TABLE 5 – Résultats de la tâche de vérification du locuteur de Parcollet *et al.* (2024) sur le corpus Fabiole. EER : *Equal Error Rate*, minDCF : *Minimum of Detection Cost Function*

5 Conclusion

Dans cet article, nous présentons le corpus Audiocite.net composé de plus de 6,600 heures d'enregistrements provenant de 130 locuteurs et locutrices, disponible sur OpenSLR (www.openslr.org/139/) avec la même licence que les livres audio (c'est-à-dire Creative Commons). Tous les enregistrements sont distribués dans leur format brut tels que nous les avons téléchargés depuis *audiocite.net* (avec des musiques de fond, des bruits, des participations inattendues, au format MP3, en mono ou stéréo). Aucun prétraitement n'a été appliqué aux fichiers, ni aucune transcription automatique effectuée sur ceux-ci. Cependant, nous avons ajouté des informations sur le genre en l'inférant à partir du nom et en vérifiant la voix en cas d'incertitude. Le corpus Audiocite.net a servi à l'apprentissage des modèles de 14k du projet LeBenchmark, révélant à la fois des performances élevées et certaines limites dans plusieurs tâches de traitement de la parole.

6 Remerciements

Ce travail a été soutenu en partie par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003), le projet E-SSL (ANR-22-CE23-0013) et la Banque Publique d'Investissement (BPI) dans le cadre de la convention de subvention THERADIA. Les auteurs tiennent à remercier chaleureusement William Havard pour l'idée originale, Marcely Zanon Boito et Fabien Ringeval pour leur aide lors de la première ébauche de ce travail.

7 Considérations éthiques

Les livres sélectionnés par les lecteurs et lectrices sont soit exclusivement sous licence Creative Commons (CC), soit obtenus par des accords de distribution écrits avec les auteurs ou autrices. Une fois la lecture terminée, les locuteurs et locutrices choisissent une seconde licence Creative Commons pour l'audio avant de publier leur enregistrement sur le site *audiocite.net*. Cette seconde licence comporte des restrictions égales ou supérieures à celles assignées au livre original. En publiant leurs enregistrements sur la plateforme, les lecteurs et lectrices étaient au courant que leur matériel pourrait être utilisé à diverses fins au-delà de l'intention originale, dans les limites de la licence attribuée. L'administrateur du site *audiocite.net* nous a explicitement donné l'autorisation d'utiliser et de distribuer les audios conformément à leurs conditions d'utilisation.

Concernant le contenu des audios, toutes sortes d'affirmations peuvent y être trouvées et nous ne souhaitons encourager personne à développer une position quelconque. Nous nous engageons à supprimer l'enregistrement, ses métadonnées et à mettre à jour le corpus à la demande de toute personne contributrice désirant retirer ses données du corpus sans raison explicite.

Références

ABDEL-RAHMAN M., HUNG-YI L., LASSE B., JAKOB D. H., JOAKIM E., CHRISTIAN I., KATRIN K., SHANG-WEN L., KAREN L., LARS M., TARA N. S. & SHINJI W. (2022). Self-supervised speech representation learning : A review. *IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio Processing*.

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

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common Voice : A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France.

BAEVSKI A., HSU W.-N., XU Q., BABU A., GU J. & AULI M. (2022). data2vec : A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39 th International Conference on MachineLearning*, Baltimore, Maryland, USA.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. In *proceedings of NeurIPS*, Vancouver, Canada.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., ZENG M., YU X. & WEI F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**, 1–14.

ESTÈVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

FELICE S., EVAIN S., ROSSATO S. & PORTET F. (2024). Audiocite.net : A large spoken read dataset in french. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

GEBRU T., MORGENSTERN J., VECCHIONE B., VAUGHAN J. W., WALLACH H., III H. D. & CRAWFORD K. (2021). Datasheets for datasets. *Commun. ACM*, **64**(12), 86–92. DOI : 10.1145/3458723.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, **29**, 3451–3460.

JIANG D., LI W., CAO M., ZOU W. & LI X. (2021). Speech SimCLR : Combining Contrastive and Reconstruction Objective for Self-Supervised Speech Representation Learning. In *proceedings of Interspeech*.

LIU A., YANG S.-W., CHI P.-H., HSU P.-C. & LEE H.-Y. (2020). Mockingjay : Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *proceedings of ICASSP*.

OORD A. V. D., LI Y. & VINYALS O. (2019). Representation Learning with Contrastive Predictive Coding. arXiv :1807.03748.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI : 10.1109/ICASSP.2015.7178964.

PARCOLLET T., NGUYEN H., EVAIN S., ZANON BOITO M., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTÈVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, **86**, 101622. DOI : https ://doi.org/10.1016/j.csl.2024.101622.

PRATAP V., XU Q., SRIRAM A., SYNNAEVE G. & COLLOBERT R. (2020). MLS : A large-scale multilingual dataset for speech research. In *INTERSPEECH*, Shanghai, China.

RAVANELLI M., ZHONG J., PASCUAL S., SWIETOJANSKI P., MONTEIRO FILHO J., TRMAL J. & BENGIO Y. (2020). Multi-Task Self-Supervised Learning for Robust Speech Recognition. In *proceedings of ICASSP*.

WANG C., RIVIERE M., LEE A., WU A., TALNIKAR C., HAZIZA D., WILLIAMSON M., PINO J. & DUPOUX E. (2021). VoxPopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, Online.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Comparaison de mesures pour la détection automatique de déviance dans la dysarthrie ataxique

Natacha Miniconi[1] Cédric Gendrot[1] Angélina Bourbon[1] Leonardo Lancia[2]
Cécile Fougeron[1]
(1) Laboratoire de phonétique et phonologie, UMR 7018 / Université Sorbonne Nouvelle
prenom.nom@sorbonne-nouvelle.fr
(2) Laboratoire Parole et Langage/ (CNRS AMU)
lancia.leonardo@univ-amu.fr

RÉSUMÉ

Cette étude explore l'utilisation d'un Réseau de Neurones Convolutifs (CNN) pour distinguer la parole de patients dysarthriques ataxiques de celle de locuteurs neurotypiques, en utilisant diverses entrées. L'objectif est d'extraire automatiquement des informations pertinentes sur les troubles de la parole. Le CNN est utilisé pour exploiter les caractéristiques temporelles et spectrales des signaux de parole via des spectrogrammes, des trajectoires de formants et des courbes de modulation cepstrale. Comparé à un Multi-Layer Perceptron (MLP) alimenté par des mesures acoustico-phonétiques ciblées sur la modulation cepstrale, le CNN présente de meilleurs scores de classification dans la distinction entre dysarthrie et non dysarthrie, en particulier avec la modulation cepstrale. La population CTRL obtient de meilleurs taux de classification que la population SCA avec un MLP, alors qu'on on observe l'inverse avec un CNN.

ABSTRACT

Comparison of measures for automatic detection of deviance in ataxic dysarthria.

This study explores the use of a Convolutional Neural Network (CNN) to distinguish the speech of ataxic dysarthric patients from that of neurotypical speakers, using various inputs. The aim is to automatically extract relevant information about speech disorders. The CNN is used to exploit the temporal and spectral characteristics of speech signals via spectrograms, formant trajectories and cepstral modulation curves. Compared with a Multi-Layer Perceptron (MLP) fed with targeted acoustic-phonetic measurements on cepstral modulation, the CNN shows better classification scores in the distinction between dysarthria and non-dysarthria, particularly with cepstral modulation. The CTRL population obtained better classification rates than the SCA population with an MLP, while the opposite was observed with a CNN.

MOTS-CLÉS : dysarthrie, deep learning, detection de déviance, acoustique.

KEYWORDS: dysarthria, deep learning, deviance detection, acoustic.

1 Introduction

L'analyse et la compréhension des signaux de parole permettent d'extraire de multiples informations sur le locuteur. Cela rejoint l'un des principaux buts de la phonétique clinique qui est d'affiner la caractérisation de la parole des patients en identifiant avec précision les aspects déviants. Les troubles de l'articulation, par exemple, sont prépondérants et peuvent être caractérisés par des difficultés à

atteindre les cibles articulatoires dans le temps et l'espace.

Dans cette étude, nous nous intéresserons à une parole pathologique particulière, celle des patients atteints d'ataxies spinocérébelleuses (SCA), un groupe hétérogène de maladies neurodégénératives dûes à une atteinte du cervelet et pouvant présenter des symptômes sur l'ensemble de la sphère motrice. Parmi ces symptômes, nous pouvons retrouver une dysarthrie qualifiée d'ataxique (Darley *et al.*, 1969). Elle se caractérise, entre autres, par des altérations de l'articulation des consonnes et voyelles et des allongements des durées segmentales (Shalling *et al.*, 2007; Brendel *et al.*, 2015; Schmitz-Hübsch *et al.*, 2011).

Généralement, l'évaluation clinique des troubles de la parole est faite à l'oreille et quantifiée à l'aide de scores basés sur des relevés d'erreurs de prononciation (Laganaro *et al.*, 2020). Cette caractérisation peut aussi être réalisée par des mesures acoustiques temporelles ou spectrales quantifiables, comme le débit de parole ou des mesures de formants sur les voyelles (Brendel *et al.*, 2015; Audibert & Fougeron, 2012). Ces dernières mesures nécessitent souvent une segmentation et/ou une intervention manuelle sur le signal de parole qui requiert une expertise approfondie et un investissement temporel conséquent. À ce titre, il est possible d'appliquer des mesures plus globales sur des séquences de parole non segmentées en phonèmes ou syllabes ; cette approche est déjà adoptée par plusieurs études sur la parole pathologique avec l'utilisation de modulation cepstrale (Slis *et al.*, 2021) ou de paramètres opensmile (Kodrasi *et al.*, 2021), de mesures acoustico-phonétiques ciblées prises sur les formants (Wang *et al.*, 2016) afin de discriminer la parole pathologique. D'autres études ont utilisé des segmentations à l'aide d'aligneur automatiques permettant de distinguer dysarthrique et non dysarthrique au niveau phonémique (Laaridh *et al.*, 2016). En contraste de ces mesures interprétables, des mesures non interprétables sont également utilisées pour capturer l'articulation comme X-vector (Favaro *et al.*, 2023).

L'utilisation des CNN pour les recherches phonétiques a introduit d'autres types d'informations pertinentes en appliquant des mesures non interprétables pour la caractérisation de l'articulation en y introduisant en entrée des images de spectrogrammes (Faragó *et al.*, 2022; Kim & Gendrot, 2022), mais aussi, des caractéristiques d'énergie de la banque de filtres Mel affectée en entrée au CNN (Abderrazek *et al.*, 2020). La présente étude cherche à évaluer la pertinence de ces différentes informations extraites d'un enregistrement audio avec un minimum d'intervention manuelle pour caractériser les troubles articulatoires dans la dysarthrie.

Ainsi, la question de recherche centrale se formule comme suit : quelles informations spécifiques s'avèrent pertinentes pour distinguer la parole de patients dysarthriques de celle de locuteurs non dysarthriques ?

Pour répondre à cette interrogation, nous proposons d'utiliser un Réseau de Neurones Convolutifs (CNN) ayant comme tâche une classification binaire : dysarthrique ou non-dysarthrique avec plusieurs types d'entrées tels que :

— Images de spectrogrammes
— Images capturant des trajectoires des formants
— Images capturant la modulation cepstrale des productions

Afin d'évaluer la pertinence des classifications produites par le CNN, nous confrontons ses résultats à une classification basée sur des mesures issues d'une expertise phonétique. Pour ce faire, nous utiliserons un perceptron multicouche (MLP) configuré pour utiliser des mesures ciblées sur certains événements de la modulation cepstrale (par exemple les maxima). Ce type de modèle MLP a déjà été utilisé sur d'autres types de dysarthrie en l'alimentant de mesures phonétiques (Alshammri

et al., 2023).

2 Méthodologie des entrées du CNN et du MLP

Les enregistrements utilisés pour cette étude proviennent des projets SpeechN'Co (n° ID-RCB : 2019-A02553-54) et ChaSpeePro (CRSII5_202228). Ils incluent 30 locuteurs, 16 hommes et 14 femmes (*moy.* = 53.4, 23<>72 ans), tous porteurs d'ataxies spinocérébelleuses. Ils présentent tous une dysarthrie, avec une sévérité variable évaluée à l'aide du score perceptif de la batterie d'évaluation BECD (Bourbon *et al.*, 2023). Nous avons comparé ces enregistrements des locuteurs dysarthriques à ceux de deux groupes contrôles, chacun constitués de 30 locuteurs qui ont été tirés aléatoirement à partir des deux bases de données, comprenant des locuteurs neurotypiques âgés de 24 à 90 ans. Le groupe CTRL1 comprend 14 hommes et 16 femmes (*moy.* = 60.38 ans, 24<>90 ans) et le groupe CTRL2 comprend 11 hommes et 19 femmes (*moy.* = 57.7 ans, 25<>82 ans) Le matériel linguistique enregistré est extrait du protocole MonPaGe-MoSpeeDi et consiste en trois séquences glides-voyelles ayant un sens en français. Ces séquences comportent chacune trois syllabes : 'aille-aille-aille' /ajajaj/, 'ouille-ouille-ouille' /ujujuj/ et 'oui-oui-oui-' /wiwiwi/. Chaque locuteur avait pour consigne de produire ces séquences de façon continue, sans pause entre les syllabes, à un débit et une intensité confortable. Ce matériel a été construit pour évaluer des modulations articulatoires sur la suite de trois syllabes via des modulations acoustiques continues sur le signal acoustique produit (Lévêque *et al.*, 2022). La Table 1 présente le nombre d'enregistrements utilisés dans l'étude.

Logatomes	SCA	CTRL1	CTRL2
ajajaj	117	119	120
ujujuj	112	119	119
wiwiwi	112	119	120
Total	341	357	359

TABLE 1 – Répartition du nombre de fichiers audios pour les groupes de locuteurs avec parole dysarthrique(SCA) et sans parole dysarthrique(CTRL1, 2) correspondant chacun à une séquence que le locuteur doit produire.

2.1 Types d'informations acoustiques extraites du signal

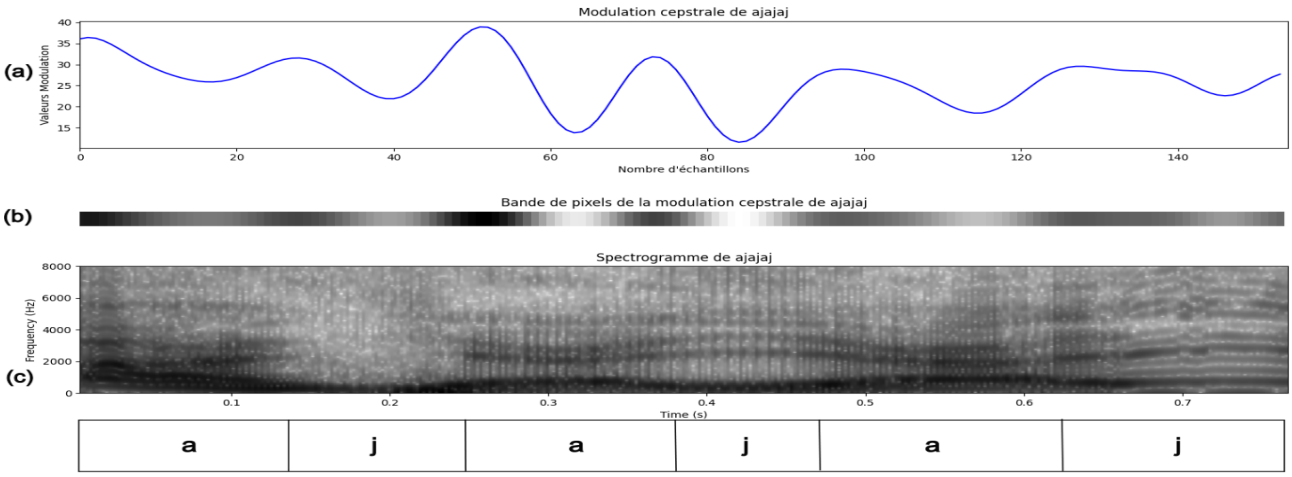


FIGURE 1 – Représentation des différents types de mesures utilisées dans le CNN pour le locuteur : SCA_F_AB03. La modulation cepstrale notée (a) devient (b) pour l'entrée du CNN. Les trajectoires formantiques subissent la même transformation. Les spectrogrammes ont été extraits sur la séquence entière (c) en tant qu'entrée pour le réseau.

Noms entrées	Types d'informations acoustiques extraites du signal
Mod_Cepstr	La modulation cepstrale permet de quantifier le degré de changement global des caractéristiques spectrales (Goldstein, 2019). Elle est ici sous la forme d'une courbe où les pics indiquent des changements d'état avec une forte différence d'énergie spectrale, tandis que les creux reflètent des périodes de stabilité avec une faible différence. Afin d'obtenir les modulations cepstrales, suivant la procédure élaborée par Leonardo Lancia (Slis *et al.*, 2021), les coefficients cepstraux en Mel (MFCC) ont été calculés grâce à la bibliothèque *Librosa* entre 300Hz et 8KHz, avec la possibilité de personnaliser des paramètres tels que la longueur de la fenêtre d'analyse (définie à 25 ms), le pas temporel (fixé à 5 ms) et le nombre de MFCC à extraire (choisi à 13). Suite à l'extraction des MFCCs, à chaque pas de l'analyse, la valeur absolue de la différence entre les valeurs successives de chaque coefficient a été calculée. Ensuite, chaque série chronologique obtenue a été soumise à un filtre passe-bas avec une fréquence de coupure fixée à 12 Hz. Afin de pouvoir utiliser les informations des modulations cepstrales, celles-ci ont par la suite subi une transformation visuelle en représentant chaque courbe sous la forme d'une bande de pixels. Dans cette représentation, l'intensité de la courbe est reflétée par la teinte des pixels, où une teinte plus foncée correspond à une intensité plus élevée de la courbe. L'échelle a été normalisée sur la totalité du corpus. Sa représentation est visible sur la Figure 1b.
Spectro	Les spectrogrammes, capturant les déformations potentielles des voyelles et des consonnes induites par la maladie (Shalling *et al.*, 2007). Les fréquences en Hz retenues pour les spectrogrammes se situent dans la plage de 0 à 8 000 Hz.
F1, F2, F3	Les spectrogrammes sont susceptibles de comporter beaucoup de bruit. De ce fait, nous allons également extraire les trajectoires des formants F1, F2, F3 sur la durée totale de la séquence produite grâce à la fonction *To Formant (burg)* de Praat qui extrait la valeur de chaque formant toutes les 5ms. Après un lissage sur python, pour chaque formant, la trajectoire a été transformée en une bande de pixels, suivant le même processus appliqué à la courbe de modulation cepstrale. Les bandes de formant ont été empilées, cela permettant d'avoir une image comportant les trois bandes de formants (F1+F2+F3), comme le spectrogramme illustré figure 1c.
Mean pics	Moyennes des pics sur la courbe de modulation cepstrale correspondant au moment dans la production avec le plus de changement cepstraux (=aux transitions entre les phonèmes). Voir Figure 2.
Meanch	Moyenne de toutes les valeurs de la courbe de modulation cepstrale sur la production. Voir Figure 2.
SD Meanch	Ecart-type de toutes les valeurs de la courbe de modulation cepstrale sur la production.
EventDUR	La moyenne des durées entre deux minimums consécutifs. Voir Figure 2.
SD EventDUR	Ecart-type des durées entre les pics qui reflète la régularité des durées entre les segments.

TABLE 2 – Descriptions des différents types d'informations acoustiques extraites du signal

Les mesures ciblées (Mean pics, Meanch, SD Meanch, EventDUR, SD EventDUR) réalisées sur la modulation cepstrale développées dans des études ultérieures (Slis *et al.*, 2021; Lévêque *et al.*, 2022) sont basées sur une quantification de la modulation cepstrale durant la séquence produite, de façon à approximer les changements dans le temps du conduit vocal lors de la production. Les mesures ont permis de mettre en évidence des différences dans les schémas articulatoires des personnes atteintes de dysarthrie. Le même type de méthode a déjà été réalisé dans des études antérieures où ont été utilisées comme indice articulatoire les transitions : consonnes, voyelles (Mathad *et al.*, 2022; Xu *et al.*, 2022). Ci-dessous la figure 2 représentant la prise des mesures ciblées.

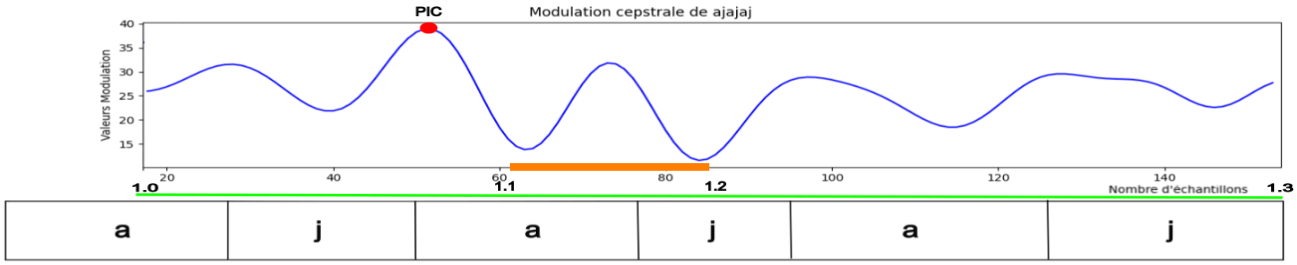


FIGURE 2 – Modulation cepstrale d'un *ajajaj* du locteur SCA_F_AB03. La valeur moyenne des 5 pics maximums représente la mesure Mean pics. EventDUR représente la durée moyenne de chaque intervalle de temps entre deux points de minimum consécutifs dans la modulation cepstrale. Une de ces durées est représentée entre le point 1.1 et 1.2 surlignée en orange. Meanch représente la valeur moyenne de toutes les valeurs sur la durée entre 1.0 et 1.3 surlignée en vert.

2.2 Élaboration du CNN et du MLP

Les CNN prennent comme entrées : Spectro, Mod_Cepstr et les Trajectoires F1, F2, F3 sous forme d'image. Le MLP quant à lui reçoit comme entrée des données structurées sous forme numérique,

comprenant des mesures spécifiques affectées sur la modulation cepstrale décrite dans la Table 2 (Mean pics, Meanch, SD Meanch, EventDUR, SD EventDUR). Ces mesures, préalablement calculées et organisées dans un tableau, ont servi d'entrée pour l'entraînement et pour le test du MLP.

Pour les deux modèles, la tâche de classification est identique : il s'agit de déterminer si les enregistrements vocaux proviennent d'un locuteur avec une parole dysarthrique ou non. Pour y parvenir, nous avons utilisé un corpus de 30 locuteurs ayant une parole dysarthrique (SCA) et de 30 locuteurs n'ayant pas de parole dysarthrique (CTRL1) constituant le premier essai appelé test1. Ensuite, pour valider notre méthode, nous avons effectué la même tâche de classification en utilisant cette fois les 30 locuteurs ayant une parole dysarthrique et 30 autres locuteurs n'ayant pas de parole dysarthrique issus groupe CTRL2. Cette classification constitue notre deuxième essai appelé test2. Ce test2 permet d'évaluer la stabilité des modèles en regardant si les résultats sont les mêmes avec un autre groupe CTRL.

Lors de la réalisation des deux modèles, la procédure "*Leave One Out*" (LOO) a été employée. Le modèle a été exécuté sur 60 itérations pour les corpus comprenant SCA et CTRL1 et CTRL2, en utilisant chaque locuteur comme locuteur de test une fois, tout en utilisant toutes les données des autres locuteurs pour l'entraînement. Cela nous permet d'éviter du biais ainsi que des bonnes performances dues au hasard ou bien à un choix particulier sur le corpus test (Berrar, 2019).

Pour chaque sortie de ces modèles, nous obtenons plusieurs scores de performance : l'*accuracy* de chaque locuteur, le *F-score* et le *rappel*. Ces deux derniers scores n'étant pas pertinents pour cette étude, ils ne seront pas mentionnés. L'*accuracy* correspond au nombre d'enregistrements correctement prédits sur l'ensemble d'enregistrement du locuteur.

L'architecture du CNN débute par une couche de convolution avec 64 filtres de taille 3×3 et un padding de type valid qui produit une sortie de taille réduite par rapport à l'entrée, suivi d'une normalisation par lots. Elle incorpore ensuite un *ZeroPadding2D* pour gérer la variation de la longueur des enregistrements audios. Chaque locuteur a fourni des enregistrements audios de durées différentes, ce qui a conduit à des variations dans la taille des images correspondantes aux différentes mesures. La couche de *ZeroPadding2D* a été employée pour pallier ce problème. Son rôle est d'ajuster la taille des images spectrogrammes plus courtes en ajoutant des "pixels de remplissage" autour d'elles. Ensuite, un *MaxPooling2D* a été ajouté avec un pool de taille 2×2, et une autre normalisation par lots. La seconde couche de convolution utilise 16 filtres de 3×3 avec padding de type valid, suivi d'une nouvelle normalisation. Après avoir aplati les données, le modèle applique un Dropout de 0.5, puis trois couches denses avec respectivement 16 et 8 unités avec activation ReLU, et enfin une unité avec activation sigmoïdale.

Quant au MLP, il a été utilisé la bibliothèque *sklearn*. La fonction d'activation des couches cachées est ReLU. La taille des couches cachée est de 100 neurones, nous avons choisi le même nombre de couches cachées que pour le CNN. D'autres paramètres, tels que le taux d'apprentissage constant, le nombre maximal d'itérations, l'optimiseur "adam" sont fixés aux valeurs par défaut. Une analyse en composantes principales (ACP) a été faite sur les mesures numériques données au MLP et a révélé que la première composante principale explique 50% de la variance, tandis que la deuxième en explique 30%. Ensemble, ces deux composantes couvrent donc 80% de la variance totale. Pour atteindre le seuil souhaité de 95% de la variance expliquée, nous nous retrouvons dans une situation où le nombre de composantes nécessaires dépasse celui des mesures originales, rendant l'approche peu avantageuse. L'apprentissage du modèle a donc été effectué sans réduction de dimension.

3 Résultats

Entrées CNN	SCA		CTRL1	
	Accuracy	SD Accuracy/min/max	Accuracy	SD Accuracy/min/max
Mod_Cepstr	**84%**	17% / 50% / 100%	**72%**	23% / 1% / 100%
Spectro	74%	30% / 1% / 100%	47%	37% / 0% / 100%
F1 + F2 + F3	72%	26% / 1 % / 100 %	42%	30% / 0% / 92 %
F1	73%	22% / 16% / 100%	38%	22% / 0% / 90%
F2	71%	23% / 17% / 100%	42%	20% / 1% / 83%
F3	72%	17% / 36% / 100%	30%	18% / 0% / 83%
Entrées MLP				
Mean pics + Meanch + SD Meanch + EventDUR + SD EventDUR	73%	30% / 1% / 100%	76%	30% / 30% / 100%
Mean pics	65%	31% / 1% / 100%	78%	20% / 33% / 100%
Meanch + SD Meanch	**74%**	27% / 0% / 100 %	**81%**	26% / 1% / 100%
EventDUR + SD EventDUR	70%	27% / 0 % / 100 %	71%	30% / 1% / 100 %

TABLE 3 – Performances du CNN et MLP lors du test 1 en fonction des différentes entrées et des groupes (SCA et CTRL1). Les performances sont estimées à partir du pourcentage de fichiers bien classés (*accuracy*), de la variabilité des *accuracy* entre les locuteurs en termes d'écart-type, minimum et maximum d'*accuracy* (SD/min/max *accuracy*).

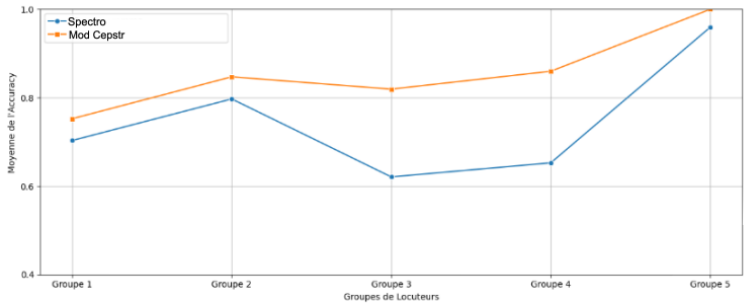
3.1 Effet du type d'entrée sur les performances de classification (CNN et MLP)

La Table 3 présente la moyenne des résultats de l'évaluation des performances du CNN et MLP avec chacune des entrées pour le groupe SCA et le groupe CTRL1 lors du test1. Les entrées provenant d'informations sur la modulation cepstrale montrent une meilleure performance face à celles provenant d'un spectrogramme. Les résultats montrent que la meilleure entrée du CNN est Mod_Cepstr avec 84% des enregistrements bien classés pour les locuteurs SCA et 72% pour les CTRL1. Lors du test2, avec la classification sur les mêmes locuteurs du groupe SCA et un groupe d'autres locuteurs contrôles (CTRL2), les résultats sont similaires. L'entrée Mod_Cepstr classifie correctement 80% des productions des locuteurs SCA et 71% des productions des locuteurs du groupe CTRL2. Les entrées du MLP extraites de la modulation cepstrale montrent également de bonnes performances en comparaison de celles issues d'un spectrogramme (Spectro, F1, F2, F3). La meilleure entrée du MLP est celle composée de Meanch et SD Meanch avec une *accuracy* moyenne de 74% pour les locuteurs SCA et 81% pour les locuteurs CTRL1. Pour la comparaison avec les locuteurs CTRL2, nous avons à nouveau peu de différences : une *accuracy* de 72 % pour les locuteurs SCA et 77 % pour les locuteurs CTRL2. Ce faible écart entre les groupes CTRL1 et CTRL2 s'observe par ailleurs sur la totalité des entrées. L'*accuracy* de l'entrée Spectro avec 74% pour les locuteurs SCA et 47% pour les locuteurs CTRL1 est légèrement inférieure à celle de l'entrée Mod_Cepstr pour les locuteurs SCA mais pour les locuteurs CTRL1, une chute des performances est observée où un enregistrement sur deux est mal classé. Quant aux entrées F1, F2, F3, elles sont inférieures ou égales à l'entrée Spectro, nous observons cette même chute de performances pour les groupes CTRL1 et 2, en particulier avec F3.

3.2 Performance de classification en fonction des locuteurs (CNN et MLP)

Un calcul du coefficient de corrélation de Spearman est réalisé afin de savoir si la sévérité des SCA est corrélée au taux d'*accuracy* pour les différentes entrées. Nous pouvons voir que les performances ne sont pas similaires en fonction de la sévérité des troubles de la parole : sans surprise, les locuteurs les plus sévères sont les mieux reconnus comme étant dysarthriques, à hauteur de 100% des enregistrements pour les locuteurs les plus sévères sur la base de l'entrée Mod_Cepstr. Pour les locuteurs les moins sévères, une baisse de performance à hauteur de 18% est observée. Une corrélation positive modérée significative est constatée entre l'*accuracy* et la sévérité pour l'entrée Mod_Cepstr ($\rho = 0.42$, $p < 0.05$), mettant en exergue l'influence de la sévérité. En revanche,

pour toutes autres entrées du CNN, la corrélation est quasi-nulle avec une p-value non significative ($\rho = -0.2$, $p > 0.05$). Le taux d'*accuracy* des performances de classification avec les entrées : Spectro, F1, F2, F3 n'est pas sensible à la sévérité de la dysarthrie. Sur la Figure 3 nous pouvons remarquer que Mod_Cepstr classe beaucoup mieux les plus sévères avec un faible creux dans les sévérités intermédiaires, à contrario de l'entrée Spectro qui n'est effectivement pas affecté par la sévérité.



Groupes	Sévérité (S_1)
Groupe 1	S_1 compris entre 1.4 et 3.4
Groupe 2	S_2 compris entre 4.6 et 6.5
Groupe 3	S_3 compris entre 7 et 8.8
Groupe 4	S_4 compris entre 9.2 et 12.80
Groupe 5	S_5 compris entre 13.4 et 17

FIGURE 3 – Graphique de l'accuracy des 5 groupes SCA dans l'ordre croissant de sévérité en fonction de l'accuracy. Les sévérités des locuteurs de chaque groupe sont indiquées dans le tableau à droite.

Nous retrouvons cette corrélation positive et significative entre les performances de classification du MLP et la sévérité de locuteurs SCA. Pour l'entrée détenant l'intégralité des mesures, le coefficient de corrélation indique une forte corrélation positive entre le taux d'*accuracy* et la sévérité ($\rho = 0.72$, $p < 0.001$). Enfin, l'entrée incluant EventDUR et SD EventDUR présente une corrélation positive et légèrement moins élevée que celle de l'entrée détenant l'intégralité des mesures ($\rho = 0.65$, $p < 0.001$). Nous obtenons les mêmes valeurs pour Mean pics. L'entrée Meanch + SD Meanch montre une corrélation positive légèrement moins élevée ($\rho = 0.58$, $p < 0.001$). L'apport du CNN par rapport au MLP est notable sur la classification des locuteurs peu sévères, ce qui est observable avec l'entrée Mod_Cepstr du CNN apportant une *accuracy* de 18% pour les locuteurs les moins sévères en comparaison à la meilleure entrée du MLP (Meanch + SD Meanch).

Nous avons observé une importante différence d'*accuracy* minimum entre les entrées : pour la mesure Mod_Cepstr le locuteur le moins bien classé est à hauteur de 50% dans le groupe SCA, alors que, pour les entrées du MLP le minimum est entre 0 et 1% indiquant une très mauvaise classification pour au moins un des locuteurs SCA. Au sujet des CTRL, les entrées du CNN rencontrent davantage de difficultés par rapport à celles du MLP pour classifier correctement les locuteurs des groupes CTRL1 et CTRL2 sur l'ensemble des données soumises. Sa meilleure performance se trouve avec l'entrée Mod_Cepstr. Les entrées du MLP quant à elles classent mieux les CTRL que cela soit pour le test1 avec les CTRL1 ou le test2 (CTRL2). A contrario des entrées du CNN, deux entrées du MLP ont un minimum d'*accuracy* nettement plus élevé à 30%, 33% (toutes les entrées en entraînement et Mean pics) par rapport aux performances du minimum d'*accuracy* pour le CNN qui tournent autour de 0% pour l'intégralité de ses entrées. L'intégralité des entrées du MLP ont au moins un locuteur CTRL qui est classé à 100% pour le CNN ; nous avons ce taux uniquement pour deux entrées : Mod_Cepstr et Spectro montrant une fois de plus les difficultés du modèle sur cette population.

Au regard de l'écart-type de l'*accuracy* pour chaque entrée présent dans la Table 3, la mesure Mod_Cepstr a un écart-type parmi les plus bas, ce qui suggère une faible variation dans la classification des locuteurs SCA et une stabilité dans les performances. Pour la totalité des mesures données au MLP, nous observons de plus grands écarts-types indiquant une variabilité des classifications entre locuteurs, aussi bien pour le groupe SCA que les groupes CTRL1 et CTRL2.

4 Conclusion et Discussion

Nous avons observé des performances différentes des modèles CNN et MLP en fonction des types d'entrée. L'entrée Mod_Cepstr pour le groupe SCA s'est avérée meilleure face aux autres types d'entrées du CNN ciblées sur les informations de spectrogramme (Spectro, F1, 2, 3). Ces résultats s'observent pour les deux groupes CTRL. Ces performances peuvent s'expliquer par les multiples informations que porte la modulation cepstrale (Slis *et al.*, 2021; Lévêque *et al.*, 2022).

Le CNN avec des mesures non interprétables s'est montré plus performant pour la population SCA que le MLP avec des mesures issues d'une expertise phonétique (interprétables). Le MLP classe beaucoup mieux les patients très sévères que les patients peu sévères avec un écart-type qui montre une plus grande variation dans sa classification. L'efficacité de mesures non interprétables effectuées par un modèle de traitement automatique a également déjà été observée dans une étude antérieure où les mesures non interprétables ont surpassé celles étant interprétables (Favaro *et al.*, 2023). Nous avons observé que les mesures numériques sur la modulation cepstrale assimilées par le MLP sont plus performantes pour la population CTRL. Les difficultés du CNN sur la population CTRL pourraient s'expliquer par une variabilité moins prononcée ou différente au sein des locuteurs dans les groupes CTRL1 et CTRL2 par rapport aux SCA. Malgré les tests de plusieurs architectures de modèles et l'augmentation du volume de données pour cette catégorie durant l'entraînement, aucune amélioration des résultats n'a été observée. La quantité relativement faible de données disponibles dans cette étude pourrait également contribuer à ce problème. Il a été noté dans l'étude précédemment citée (Favaro *et al.*, 2023) que pour les corpus détenant plus de données, des mesures non interprétables (les traits provenant du modèle TRILLsson) étaient plus performantes et avaient un plus gros écart de performances en comparaison des mesures interprétables (des traits prosodiques).

Les difficultés du MLP à classer les locuteurs SCA avec une faible dysarthrie peuvent être dues à des mesures similaires entre CTRL et SCA dues à la précocité de la maladie (Slis *et al.*, 2021), ce qui peut causer des difficultés pour le MLP à apprendre un motif particulier pour cette population et à les discriminer convenablement. À contrario, l'aisance du MLP dans la classification de la population CTRL peut résider dans le fait d'avoir des mesures plus stables dans la population CTRL avec très peu de variations, ce qui peut l'aider à apprendre un pattern. La performance observée de l'entrée Meanch et SD Meanch démontre son importance (Slis *et al.*, 2021). Ces mesures peuvent refléter des troubles articulatoires comme des difficultés à atteindre des cibles articulatoires lors de transition d'un phonème à l'autre, ce qui est observable chez la population SCA (Shalling *et al.*, 2007). En effet, le modèle arrive à apprendre un pattern pour chaque classe, ce qui permet cette bonne classification. Globalement, avec la totalité des mesures condensées, il est observé un score inférieur de 8% du MLP par rapport au CNN avec l'entrée Mod_Cepstr pour la population SCA. L'entrée Mod_Cepstr dans le CNN peut capturer des motifs spatiaux et des relations locales dans les données qui ne sont pas évidentes ou directement accessibles via les mesures brutes de la modulation cepstrale utilisées dans le MLP. Cela peut inclure des nuances subtiles et des motifs complexes qui sont importants pour la classification.

Ces résultats laissent plusieurs ouvertures pour la suite : les mesures prises par le CNN sur la modulation cepstrale ont un avantage pour la population SCA, ce qui montre une possibilité d'affiner les mesures numériques ciblées par la suite pour pouvoir capter des caractéristiques clefs permettant de différencier un locuteur peu atteint d'un témoin. Des tests prometteurs pour la même tâche ont été réalisés à l'aide de grands modèles de langues pré-entrainés (w2v2) et cette piste est en cours d'exploration.

Références

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2020). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders - step 1 : Cnn model-based phone classification. *Interspeech 2020*, **27**, 522–2526,. DOI : hal-03017394.

ALSHAMMRI R., ALHARBI G., ALHARBI E. & ALMUBARK I. (2023). Machine learning approaches to identify parkinson's disease using voice signal features. *Sec. Medicine and Public Health*, **6**. DOI : 10.3389/frai.2023.1084001.

AUDIBERT N. & FOUGERON C. (2012). Distorsions de l'espace vocalique : quelles mesures ? application à la dysarthrie. *JEP TALN*, **27**, 217–224. DOI : hal-02436294.

BERRAR D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, p. 542–545. DOI : 10.1016/B978-0-12-809633-8.20349-X.

BOURBON A., FOUGERON C. & CREVIER_BUCHMAN L. (2023). *Effects of instruction and content on repetition performances of ataxic dysarthric and healthy speakers*. Thèse de doctorat, 8th International Conference on Speech Motor Control Groeningen.

BRENDEL B., SYNOFZIK M., ACKERMANN H., LINDIG T., SCHÖLDERLE T., SCHÖLS L. & ZIEGLER W. (2015). Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with friedreich ataxia. *J Neurol*, **262**, 21–26. DOI : 10.1007/s00415-014-7511-8.

DARLEY F., ARONSON A. & BROWN J. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, p. 246–269.

FARAGÓ P., ȘTEFĂNIGĂ S.-A., CORDOȘ C.-G., MIHĂILĂ L.-I. & HINTEA S. (2022). Cnn-based identification of parkinson's disease from continuous speech in noisy environments. *JSLHR*, **5**, 1767–1783. DOI : 10.3390/bioengineering10050531.

FAVARO A., TSAI Y.-T., BUTALA A., THEBAUD T., VILLALBA J., DEHAK N. & MORO-VELÁZQUEZ L. (2023). Interpretable speech features vs. dnn embeddings : What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, **166**, 107559. DOI : 10.1016/j.compbiomed.2023.107559.

GOLDSTEIN L. (2019). The role of temporal modulation in sensorimotor interaction. *Front. Psychol*, **10**, 2068.

KIM L. & GENDROT C. (2022). Classification automatique de voyelles nasales pour une caractérisation de la qualité de voix des locuteurs par des réseaux de neurones convolutifs. *JEP*, p. 13–17.

KODRASI I., PERNON M., LAGANARO M. & BOURLARD H. (2021). Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech. *ICASSP*.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2016). Détection automatique d'anomalies sur deux styles de parole dysarthrique : parole lue vs spontanée. *JEP*.

LAGANARO M., FOUGERON C., PERNON M., LEVÊQUE N., BOREL S., CHIUVÉ M. F. S. C., URSULA LOPEZ R. T., MÉNARD L., BURKHARD P. R., ASSAL F. & DELVAUX V. (2020). Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in french : the monpage-screening protocol. *Clinical Linguistics & Phonetics*, **35**, 1060–1075. DOI : 10.1080/02699206.2020.1865460.

LÉVÊQUE N., SLIS A., LANCIA L., BRUNETEAU G. & FOUGERON C. (2022). Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases. *JSLHR*, **5**, 1767–1783. DOI : 10.1044/2022_JSLHR-21-00434.

MATHAD V. C., LISS J. M., CHAPMAN K., SCHERER N. & BERISHA V. (2022). Consonant-vowel transition models based on deep learning for objective evaluation of articulation. *IEEE*, **31**, 86–95. DOI : 10.1109/TASLP.2022.3209937.

SCHMITZ-HÜBSCH T., ECKERT O., SCHLEGEL U., KLOCKGETHER T. & SKODDA S. (2011). Instability of syllable repetition in patients with spinocerebellar ataxia and parkinson's disease. *Mov disord*, **27**, 316–319. DOI : 10.1002/mds.24030.

SHALLING E., HAMMARBERG B. & HARTELIUS L. (2007). Perceptual and acoustic analysis of speech in individuals with spinocerebellar ataxia (sca). *Logopedics Phoniatrics Vocology*, **32**, 31–46. DOI : 10.1080/14015430600789203.

SLIS A., FOUGERON C., LÉVÊQUE N., PERNON M., ASSAL F. & LANCIA L. (2021). Analysing spectral changes over time to identify articulatory impairments in dysarthria. DOI : 10.1121/10.0003332.

WANG J., KOTHALKAR P. V., CAO B. & HEITZMAN D. (2016). Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. *Interspeech2016*. DOI : 10.21437/Interspeech.2016-1542.

XU L., LISS J. & BERISHA V. (2022). Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA*, **3**, 015201. DOI : 10.1121/10.0016833.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé

Jingyi Sun[1] Yaru Wu[1, 2, 3] Nicolas Audibert[1] Martine Adda-Decker[1, 3]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 Rue des Irlandais, 75005 Paris, France
(2) CRISCO/UR4255 (Université de Caen Normandie), Esp. de la Paix, 14000 Caen, France
(3) LISN (Univ. Paris-Saclay), Rue du Belvédère, 91405 Orsay, France
{jingyi.sun, nicolas.audibert, martine.adda-decker}@sorbonne-nouvelle.fr, yaru.wu@unicaen.fr

RÉSUMÉ

La technologie ASR excelle dans la transcription précise des discours lus préparés, mais elle rencontre encore des défis lorsqu'il s'agit de conversations spontanées. Cela est en partie dû au fait que ces dernières relèvent d'un registre de langage non préparé et informel, avec disfluences et réductions de parole. Afin de mieux comprendre les différences de production en fonction des styles de parole, nous présentons la création d'un corpus de parole conversationnelle, dont des extraits sont ensuite lus par leurs auteurs. Le corpus comprend 36 heures de parole en chinois mandarin avec leur transcription, réparties entre conversations spontanées et lecture. Nous avons utilisé WHISPER pour la transcription automatique de la parole et le *Montreal Forced Aligner* pour l'alignement forcé, résultant dans un corpus de parole transcrit avec annotations multi-niveaux incluant phonèmes, caractères/syllabes et mots. De telles productions de parole parallèles (en modes spontané et lu) seront particulièrement intéressantes pour l'étude des réduction temporelle.

ABSTRACT

Creating a Speaking Styles Parallel Corpus in Mandarin through Auto-transcription and Forced Alignment

ASR technology excels in accurately transcribing prepared read speech, but it still encounters challenges when dealing with spontaneous conversations. This is partly because the latter is an unprepared, casual language register with lots of disfluencies and speech reductions. In order to improve our knowledge about speech variations, we designed an oral corpus of 36 hours of Mandarin Chinese. Spontaneous conversations are automatically transcribedand selected excerpts of these are then read by their authors. We employed WHISPER for automatic speech transcription and the *Montreal Forced Aligner* for forced alignment, which represents an effective procedure from speech to text, then to multi-tier annotation within phones, characters/syllables, and words, particularly suitable for large-scale speech corpus construction. This enabled us to collect different speaking styles with partly overlapped content produced by the same speaker. Such parallel corpus are particularly helpful to investigate temporal speech reduction phenomena in spontaneous speech.

MOTS-CLÉS : corpus parallèle, style de parole, auto-transcription, alignement forcé.

KEYWORDS: parallel corpus, speaking style, auto-transcribe, forced aligment.

Actes des 35èmes Journées d'Études sur la Parole, pages 291–300.

1 Introduction

Malgré les progrès considérables réalisés en ASR au cours des dernières décennies, la parole spontanée et, en particulier, informelle reste difficile. Différents facteurs, tels qu'une possible diminution de la qualité du canal, le bruit de fond et les chevauchements de parole, pourraient être mentionnés comme des explications possibles de la performance plus faible de l'ASR (Benzeghiba *et al.*, 2007; O'Shaughnessy, 2008). Au-delà des facteurs non linguistiques mentionnés ci-dessus, il semble que les changements dans le contenu linguistique lui-même pourraient également expliquer partiellement certains aspects de la diminution de la précision de l'ASR : les disfluences, l'hypoarticulation et la réduction temporelle de la parole, qui peuvent entraîner des variantes de prononciation inattendues ou moins documentées (Adda-Decker & Lamel, 2018). Ces caractéristiques de la parole sont particulièrement remarquables dans des styles de parole moins contrôlés.

Deux styles de parole typiques sont prédominants dans les ensembles de données d'entraînement couvrant plusieurs langues : la lecture attentive et soignée et la conversation informelle non préparée. Ces styles de parole représentatifs couvrent deux extrémités d'un continuum (Gabler *et al.*, 2023). La lecture attentive préserve les formes de prononciation relativement intactes des phonèmes et la coarticulation nécessaire (Farnetani & Recasens, 1997). Dans les dialogues spontanés, le débit de parole et l'accentuation sont très flexibles et parfois incorrects sur le plan grammatical. Les unités linguistiques non accentuées, prévisibles et de haute fréquence tendent à être prononcées rapidement et même avec une réduction extrême. De plus, cette parole contient de nombreuses pauses remplies, des bégaiements, des répétitions, des autocorrections, des hésitations et des marqueurs de discours, ainsi que des rires et des toux, tous difficiles à éviter tout en maintenant la naturalité de la parole. Ces caractéristiques sont également essentielles pour mesurer la spontanéité et l'informalité de la parole (Dufour *et al.*, 2009).

Le corpus parallèle mandarin que nous sommes en train de créer est une étape importante vers l'extraction des paramètres phonétiques entre le registre informel et formel. Il est cependant important de noter que le terme « parallèle » utilisé ici diffère à la fois de la définition de Baker (Baker, 1995), qui fait référence à des corpus contenant des textes et des traductions dans deux langues ou plus, et du concept de Johansson (Johansson *et al.*, 1998), qui fait référence à des corpus contenant des textes dans deux langues avec des relations universelles et comparables. En utilisant le style de parole comme seule variable de comparaison, nous avons créé un corpus parallèle mandarin avec la même langue, le même locuteur, aucun dialecte régional et un contenu linguistique égal, mais seulement dans des styles de parole différents.

2 Travaux Connexes

Le développement de bases de données de parole de conversation spontanée a connu une croissance rapide au cours des dernières décennies, avec des bases de données disponibles dans différentes langues, notamment l'anglais, le français, l'espagnol, l'allemand, l'italien et le mandarin taïwanais (Du Bois *et al.*, 2000; Torreira & Ernestus, 2010; Kohler, 1996; Mereu & Vietti, 2021; Tseng, 2019). Cependant, dans les corpus de parole disponibles publiquement et axés sur le chinois mandarin , le style prédominant reste la lecture scriptée. Des bases de données à grande échelle de mandarin provenant de différentes régions et groupes d'âge comprennent, par exemple, *Chinese Mandarin (South/North) database (ELRA)*, *Chinese Digital Speech Data by Mobile Phone (ELRA)*, *AISHELL*

Speech databases (Bu *et al.*, 2017) et *1997 Mandarin Broadcast News Speech* (Graff, 2002). Il y a également eu des progrès significatifs dans le développement de jeux de données de parole spontanée en mandarin. Des exemples notables incluent les données de conversation en mandarin du *Mandarin Conversational Speech Data du Primewords Chinese Corpus Set 1* (Primewords Information Technology Co., 2018) et le *Magic Data Chinese Mandarin Conversational Speech* (Yang *et al.*, 2022).

Les corpus interlinguistiques couvrant simultanément deux styles de parole différents sont néanmoins rares, et la plupart d'entre eux n'exigent pas un contenu linguistique identique. La création de corpus parallèles contenant des informations stylistiquement distinctes mais sémantiquement comparables, avec un alignement automatique sur plusieurs niveaux de texte pour permettre un lien direct avec le signal de parole, n'a été explorée que dans une quantité limitée de recherches (Barras *et al.*, 2004). La majorité des institutions de recherche ou des entreprises optent soit pour la diffusion de corpus ne comprenant qu'un style de parole, soit utilisent différentes méthodes pour recueillir les deux styles. La méthode la plus courante consiste à recueillir des discours spontanés de participants en réponse à des entretiens, puis à leur demander d'effectuer une tâche de parole spécifique, telle que donner des indications basées sur une carte (Thompson *et al.*, 1993; Ibrahim *et al.*, 2020) ou collaborer pour décorer un arbre de Noël (Ito & Speer, 2006). Enfin, les participants sont tenus de lire un texte standardisé. Cette stratégie de collecte de données avec des tâches indépendantes est largement utilisée, mais elle présente plusieurs inconvénients, notamment la possibilité que les réponses de différents individus à la même tâche soient très similaires ou excessivement simplistes.

L'amélioration de la précision et de la robustesse de la technologie de transcription automatique permet la construction de corpus parallèles relatifs à différents styles de parole. Nous utilisons le système de reconnaissance vocale multilingue *Whisper* (Radford *et al.*, 2023) pour obtenir rapidement la transcription de la conversation décontractée des locuteurs, qui est ensuite fournie aux locuteurs pour lecture après adaptation manuelle. Cela peut favoriser des variations de prononciation plus riches pour la modélisation des caractéristiques acoustiques et évaluer les performances de l'ASR tout en conservant un contenu linguistique cohérent.

La procédure peut également fournir des données linguistiques naturelles étendues pour soutenir l'étude des phénomènes phonétiques en mandarin, tels que la coarticulation, le sandhi tonal et la synérèse, en produisant un corpus mandarin couvrant à la fois la parole spontanée et la lecture. Ces informations permettent une évaluation minutieuse pour déterminer si des variations phonétiques spécifiques sont aléatoires ou résultent de processus phonologiques (Shih, 2005), éclairant les similitudes et les différences entre la coarticulation, la réduction de la parole et autres phénomènes propres à la parole connectée (Farnetani & Recasens, 1997). Une approche de recherche basée sur le corpus implique l'obtention de limites temporelles pour les voyelles et les consonnes individuelles par alignement forcé et annotation automatique, en se concentrant sur les correspondances erronées entre les instances dans le modèle acoustique et les phonèmes alignés, un phénomène particulièrement prévalent dans la parole spontanée. Ensuite, les distributions, les durées, les fréquences à l'intérieur et entre les catégories de sons sont comptabilisées séparément, et leurs distances acoustiques mesurées (Audibert *et al.*, 2015). De telles études servent de référence pour comprendre les causes et les tendances de la réduction phonétique. De plus, comme le ton porte une charge phonémique significative en chinois, les variations de ton possèdent une valeur théorique cruciale (Surendran *et al.*, 2006).

3 Protocole de Construction du Corpus

3.1 Locuteurs

Le corpus actuel comprend 40 locuteurs, répartis de manière équilibrée selon le genre (F :H=1 :1). Ils proviennent de 19 provinces de Chine, notamment Zhejiang, Henan, Shandong, Hunan, Anhui, Jiangsu, Yunnan, etc. Les locuteurs, âgés de 20 à 32 ans, sont parfaitement à l'aise en mandarin standard sans accent régional, et tous sont des étudiants universitaires en bonne santé ne présentant ni troubles du langage ou mentaux, ni pathologies des organes articulatoires.

Étant donné l'exigence de capturer une parole conversationnelle spontanée, les participants appariés doivent être familiers les uns avec les autres. Cela réduit les sentiments négatifs potentiels tels que l'anxiété et le malaise pendant les sessions d'enregistrement. Avant de commencer l'enregistrement, les participants ont reçu une description complète de la procédure de collecte de données. Nous leur avons présenté environ 20 sujets quotidiens portant sur l'hébergement, les études, la nourriture, les voyages, les loisirs, etc., tout en expliquant les précautions à prendre lors du processus d'enregistrement, les droits qu'ils ont de suspendre/retirer à tout moment et la sécurité des données de transcription de *Whisper*. Ensuite, ils ont été invités à signer le formulaire de consentement éclairé après avoir confirmé qu'ils n'avaient aucune préoccupation ou question. Chaque locuteur a reçu 15€ en espèces à la fin de la session d'enregistrement.

3.2 Paramètres d'Enregistrement

Nous avons utilisé un enregistreur de terrain Roland R-26 et deux microphones casques AKG C520. Le taux d'échantillonnage est de 48 kHz, tandis que la quantification est réglée sur 16 bits. De plus, nous avons mis en œuvre le système de transcription automatique de la parole open-source *Python*-basé, *Whisper*, sur un ordinateur exécutant un système Windows 11 équipé d'une carte graphique discrète NVIDIA.

Basé sur *Python* et *PyTorch*, *Whisper* est un modèle multitâche réalisé grâce à une supervision faible à grande échelle pour la reconnaissance de la parole, la traduction et l'identification de la langue. La précision et la vitesse de transcription varient en fonction de la langue (l'anglais, l'espagnol, le néerlandais et le coréen donnent les meilleurs résultats) et de la taille du modèle (avec cinq options : *tiny*, *base*, *small*, *medium*, *large*). Pour le chinois mandarin, le modèle de taille moyenne au minimum est recommandé pour la précision, les modèles plus grands atteignant une précision plus élevée mais des vitesses de reconnaissance plus lentes.

3.3 Processus d'enregistrement et d'alignement forcé

Le processus d'enregistrement et d'alignement forcé comprend cinq phases principales.

1. Conversation Spontanée

   Tout d'abord, nous demandons aux participants de s'engager dans des discussions ouvertes aussi longtemps que possible. Un groupe de locuteurs peut souvent discuter en continu pendant 40 à 70 minutes. Avant l'expérience, nous fournissons aux participants une variété de sujets

de référence portant sur de nombreux aspects de la vie quotidienne. L'audio des conversations spontanées de deux participants est enregistré en stéréo. Nous considérons que les données de conversation spontanée vraiment naturelles commencent environ cinq minutes après le début de la conversation. Une fois que les participants commencent à parler, notre enregistrement commence en même temps.

2. Transcription et Modification de Texte

   Après que le locuteur a fini de parler, nous commençons la transcription avec *Whisper*. Selon nos tests, transcrire une conversation d'une heure en utilisant le modèle *large* prend environ 18 minutes. Étant donné que *Whisper* ne différencie pas entre les locuteurs et qu'il y a quelques erreurs de reconnaissance de la parole dans le texte transcrit, causées soit par des homophones soit par une segmentation incorrecte des mots, la vérifiation et la correction manuelle des transcriptions sont nécessaires. Ensuite, ces transcriptions corrigées font l'objet d'un processus de révision et de correction par les locuteurs pour obtenir la version finale écrite pour la relecture par les locuteurs.

   Il convient cependant de noter que la structure grammaticale et sémantique de l'improvisation orale est informelle, avec de nombreuses hésitations telles que des répétitions ou des auto-corrections. Par conséquent, nous visons à maintenir ces hésitations et ces non-conformités grammaticales dans les traductions anglaises correspondantes de chaque phrase et à les effacer lors de l'adaptation au texte lu au style écrit. La parole de conversation spontanée comprend également de nombreuses phrases courtes, telles que des réponses brèves ou des déclarations interrompues. Dans ce cas, nous ne couvrons pas de manière exhaustive toutes les informations de la conversation spontanée. Au lieu de cela, nous sélectionnons des segments plus longs et relativement complets pour la réécriture. De plus, l'alternance rapide des tours de conversation peut entraîner des transcriptions modifiées incohérentes ou difficiles à comprendre lorsqu'elles sont extraites séparément. Pour éviter cela, nous pouvons, si nécessaire, incorporer des parties d'informations linguistiques du locuteur 1 dans le texte lu du locuteur 2, et vice versa.

   Comme le montre la Figure 1, nous avons apporté les ajustements suivants pour organiser le texte au style écrit :

   — Segmentation du discours non ponctué en phrases fluides à lire pour les locuteurs et qui ne semblent pas linguistiquement artificielles. Par exemple, nous transformons la phrase a. en deux phrases A., en utilisant la ponctuation pour séparer deux propositions indépendantes.
   — Réorganisation des phrases fragmentées et incorporation des informations linguistiques de l'autre personne pour compléter l'énoncé. Par exemple, la phrase i. répond à la question de l'autre, nous combinons donc la phrase h. et la recréons en tant que phrase grammaticalement et sémantiquement bien formée H.+I.. En général, les énoncés appropriés pour la supplémentation et la fusion devraient inclure au moins deux des éléments suivants : sujet, prédicat et objet. Cela permet de sélectionner les parties manquantes à partir des informations linguistiques fournies par l'interlocuteur et de compléter l'énoncé. Cependant, si les deux parties ont fourni des informations importantes pour les phrases modifiées, alors les deux locuteurs seront invités à lire la phrase.
   — Élimination des interjections et des particules modales de la transcription. Par exemple, nous retirons « 啊(ah) » dans la phrase e., une interjection, et « 嗯(um) » dans la phrase i., une particule modale.
   — Réorganisation des phrases inversées incorrectes. Par exemple, dans la phrase F., nous avons déplacé l'adverbe « 先(d'abord) » devant le VP « 预约一下(prendre rendez-vous) ».

— Élimination des phrases répétées ou des auto-corrections. Par exemple, le sujet de la phrase c. a fait l'objet d'une auto-correction, de « 我(je) » à « 我男朋友(mon petit ami) », nous ne conservons donc que cette dernière lors de la réorganisation de la phrase C.

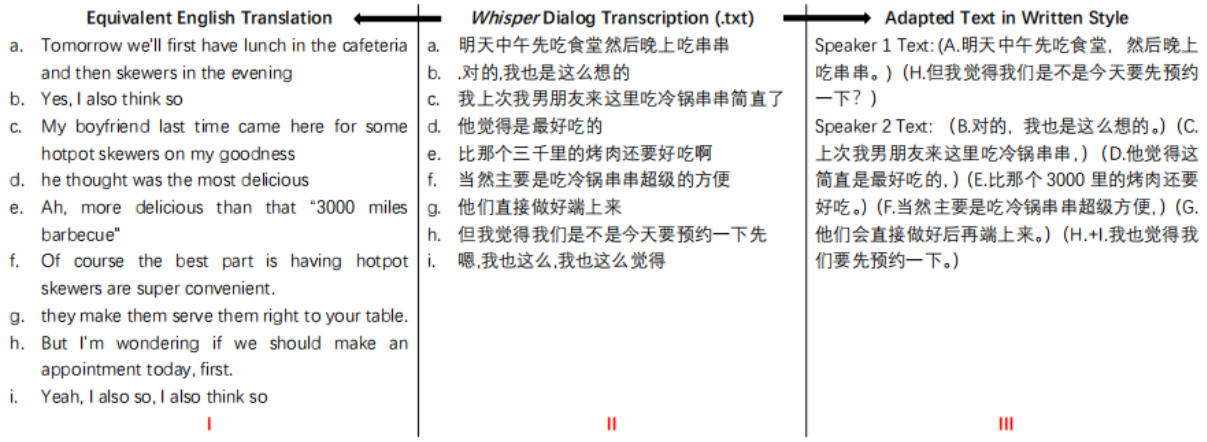


FIGURE 1 – Transcription du dialogue *Whisper* (II) au centre, la traduction anglaise correspondante (I) à gauche, et le texte adapté au style écrit (III) à droite

En nous basant sur notre expérience actuelle en matière d'édition de texte, le problème le plus difficile est de rectifier les inexactitudes dans la transcription, souvent dues aux homophones ou aux hésitations de la parole. Dans de tels cas, nous nous appuyons sur l'interprétation par le locuteur du contenu de l'interaction afin de finaliser le manuscrit du discours, car ils se souviennent généralement de ce qu'ils viennent de dire. La révision des textes des lectures des deux intervenants prend généralement environ 30 minutes au total.

3. Lecture de Texte

   Après avoir obtenu la transcription textuelle modifiée, nous demandons aux participants de la lire deux fois de manière émotionnellement neutre tout en enregistrant leur discours. Cela est destiné à obtenir autant de fluidité que possible dans la lecture de la parole typique, car la première lecture peut donner lieu à des hésitations en raison de la méconnaissance du matériel. Le discours lu est enregistré en mono, cette phase prenant généralement 10 à 15 minutes par locuteur.

4. Prétraitement de l'entrée pour l'alignement forcé

   Nous avons utilisé *Montreal Forced Aligner 3.0.6* (McAuliffe *et al.*, 2017) pour l'alignement automatique au niveau des phonèmes des mots, ce qui nécessite la préparation préalable de textes de transcription correctement formatés et d'audio sous-échantillonné. Le traitement des textes de transcription chinois nécessite une attention particulière, car ils sont écrits en chaînes continues, ce qui implique qu'il n'y a pas de repères de segmentation en dehors de la ponctuation. Ainsi, le package *spacy-pkuseg* est utilisé pour aider à la tokenisation automatique du texte chinois. L'alignement forcé est effecté en utilisant un lexique de prononciation chinois mandarin pré-entraîné et des modèles acoustiques dans MFA pour l'alignement forcé. Chaque heure de données audio nécessite environ 3 à 4 minutes de traitement pour l'alignement forcé.

5. Vérification manuelle et Réalignement

   La performance de Whisper utilisant le modèle large-v3 en mandarin a été évaluée par ses auteurs sur deux ensembles de données. Dans l'ensemble de données *Common Voice 15*, le

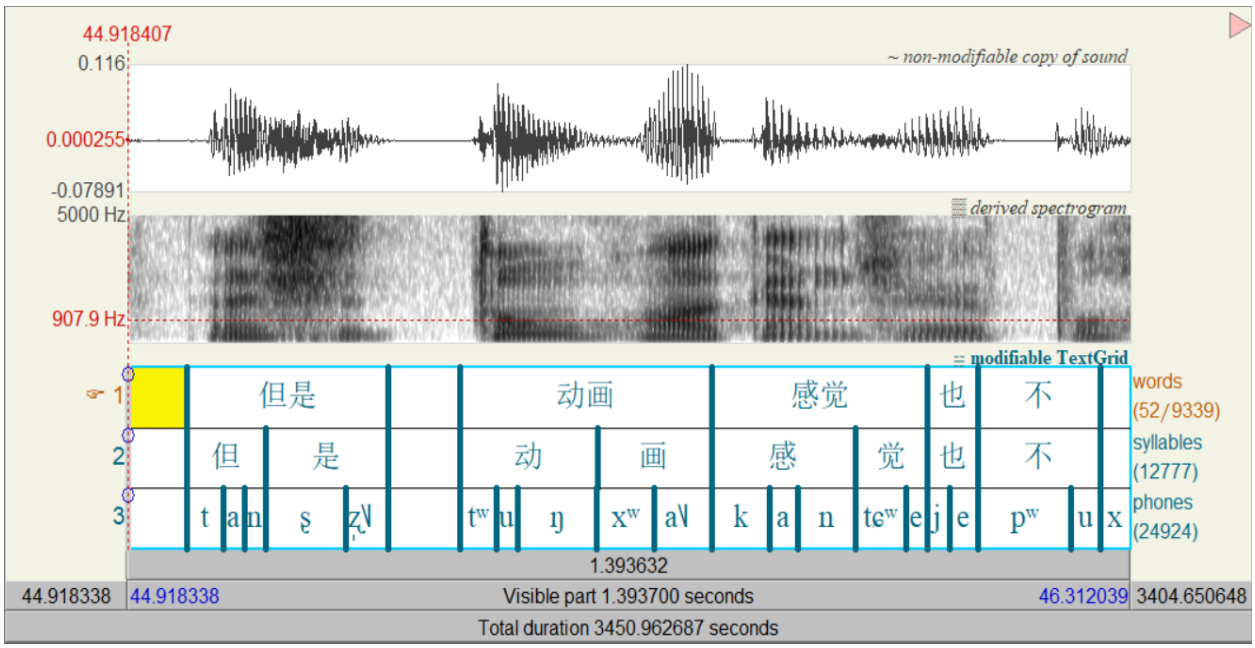


FIGURE 2 – Annotation à trois niveaux des Phones, des Caractères/Syllabes et des Mots

taux d'erreur de caractères (CER) était de 12,8%, tandis que dans l'ensemble de données *Fleurs*, le CER était de 7,7%. Cela indique que sa performance de reconnaissance varie en fonction de l'ensemble de données utilisé. Nous avons ensuite évalué *Whisper* sur la base de ce corpus parallèle de styles de parole et avons constaté que le CER pour la parole lue est de 4,18%, tandis que pour la parole de conversation spontanée, il est de 8,37%. Ces erreurs se reflètent également dans les résultats d'alignement forcé de la transcription provenant de *Whisper* et de l'audio. Par conséquent, après le premier alignement forcé, nous devons corriger les erreurs de transcription dans *Praat* (Boersma & Van Heuven, 2001) et ajouter du texte avec leurs limites pour correspondre à la parole pertinente dans les sections non reconnues. Contrairement à la correction précédente visant à faciliter la relecture par les locuteurs, cette correction de la transcription est effectuée en tenant compte autant que possible des hésitations et des autres disfluences. Cette démarche permettra d'obtenir une saisie de texte de transcription plus précise pour le réalignement.

Après vérification manuelle, un réalignement est nécessaire pour obtenir un nouveau fichier *textgrid* avec les frontières temporelles correctes pour chaque mot. De plus, étant donné que les mots ne sont pas l'unité la plus petite dans le système d'écriture chinois, nous pouvons également utiliser *Python* pour segmenter automatiquement le texte de transcription en caractères/syllabes chinois pour l'alignement. Les résultats alignés peuvent être superposés aux annotations obtenues au niveau des mots pour former des annotations à trois niveaux : phones, caractères/syllabes chinois et mots, comme illustré dans la Figure 2.

Cette approche permet la construction rapide de corpus de langage parlé spontané à grande échelle avec des données annotées. Le corpus comprend une durée totale de 36 :01 :53, comprenant 28 :59 :00 de parole de conversation spontanée et 7 :02 :53 de parole lue. Les statistiques préliminaires indiquent que la parole spontanée contient 586 554 caractères/syllabes chinois et 1 184 186 phonèmes, tandis que la parole lue contient 116 345 caractères/syllabes chinois et 245 605 phonèmes. Il est à noter que le nombre de phones dépend également du

modèle acoustique utilisé, dans ce cas, en utilisant le modèle pré-entraîné "mandarin_mfa", qui comprend 142 phones. Dans nos données de parole lue, les 28% des phones les plus représentés (40) représentent 72,11% de tous les phones. Les consonnes les plus fréquentes comprennent deux nasales, /n/ (7,75%) et /ŋ/ (4,60%), ainsi que des occlusives, des fricatives et des affriquées /t/ (4,35%), /ʂ/ (3,37%), /tɕ/ (2,24%), /z̧/ (2,60%) et /x/ (1,61%). Les voyelles les plus fréquentes comprennent /o/ (7,16%), /a/ (6,46%), /i/ (4,88%) et /ə/ (3,46%). De plus, deux approximants, /w/ (4,32%) et /j/ (3,20%), sont également très fréquents. Un affinement supplémentaire du corpus permettra d'obtenir plus de différences statistiques entre la parole spontanée et lue en chinois mandarin, inspirant ainsi une exploration plus poussée des motifs de variation allophonique.

4 Conclusion et Travaux Futurs

Cette étude propose une méthode systématique pour créer un corpus vocal parallèle concernant différents styles de parole avec le même locuteur et les mêmes informations linguistiques, avec l'aide de *Whisper* et de *Montreal Forced Aligner*. Le modèle *large* de *Whisper* permet une récupération rapide et précise du texte de transcription hors ligne, qui est ensuite adapté pour être utilisé comme matériel au style écrit pour la tâche de lecture. Dans les cas où les mots et les phrases se chevauchent fortement, ce corpus peut être utilisé non seulement pour comparer les performances de reconnaissance vocale de différents styles de parole et localiser rapidement les différences, mais il peut également fournir des variantes de prononciation riches pour l'étude des mécanismes de réduction de la parole en chinois mandarin, de la coarticulation et du sandhi tonal. En particulier, l'étude des affriquées et de l'aspiration en chinois mandarin peuvent permettre de mieux documenter les phénomènes de réduction, tandis que l'interaction entre les tons et l'intonation, la focalisation et d'autres paramètres prosodiques dans la parole spontanée extensive peut également être étudiée, éclairant des caractéristiques multidimensionnelles dans la parole informelle et formelle (Chen & Yuan, 2007). Nos prochaines étapes impliquent l'annotation automatisée, la vérification manuelle du corpus et l'analyse des fréquences d'occurrence sur les syllabes, les phonèmes, les tons et d'autres aspects pertinents pour la caractérisation de la réduction en mandarin.

Remerciements

Ce travail a bénéficié du soutien financier du Laboratoire d'Excellence Empirical Foundations of Linguistics (LabEx EFL, ANR-10-LABX-0083), contribuant ainsi à l'IdEx Université de Paris (ANR-18-IDEX-0001), ainsi que du projet ANR-21-CE38-0019 DIPVAR. Jingyi Sun a été soutenue par une bourse du China Scholarship Council (Grant No. 202208410095).

Références

ADDA-DECKER M. & LAMEL L. (2018). Discovering speech reductions across speaking styles and languages. *Rethinking reduction : Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*, **25**, 101. DOI : 10.1515/9783110524178-004.

AUDIBERT N., FOUGERON C., GENDROT C. & ADDA-DECKER M. (2015). Duration-vs. style-dependent vowel variation : A multiparametric investigation. In *18th International Congress of Phonetic Sciences (ICPhS'15)*.

BAKER M. (1995). Corpora in translation studies : An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, **7**, 223–243. DOI : 10.1075/target.7.2.03bak.

BARRAS C., ADDA G., ADDA-DECKER M., HABERT B., DE MAREÜIL P. B. & PAROUBEK P. (2004). Automatic Audio and Manual Transcripts Alignment, Time-code Transfer and Selection of Exact Transcripts. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, p. 877–880 : European Language Resources Association (ELRA).

BENZEGHIBA M., DE MORI R., DEROO O., DUPONT S., ERBES T., JOUVET D., FISSORE L., LAFACE P., MERTINS A., RIS C. *et al.* (2007). Automatic speech recognition and speech variability : A review. *Speech communication*, **49**, 763–786. DOI : 10.1016/j.specom.2007.02.006.

BOERSMA P. & VAN HEUVEN V. (2001). Speak and unspeak with praat. *Glot International*, **5**(9/10), 341–347.

BU H., DU J., NA X., WU B. & ZHENG H. (2017). Aishell-1 : An open-source mandarin speech corpus and a speech recognition baseline. http://www.aishelltech.com/kysjcp.

CHEN Y. & YUAN J. (2007). A corpus study of the 3rd tone sandhi in standard chinese. In *Interspeech*, p. 2749–2752 : Citeseer.

DU BOIS J. W., CHAFE W. L., MEYER C., THOMPSON S. A. & MARTEY N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia : Linguistic Data Consortium*. DOI : 10.35111/s2q7-gq73.

DUFOUR R., JOUSSE V., ESTÈVE Y., BÉCHET F. & LINARÈS G. (2009). Spontaneous speech characterization and detection in large audio database. *SPECOM, St. Petersburg*, **7**, 41–46.

FARNETANI E. & RECASENS D. (1997). Coarticulation and connected speech processes. *The handbook of phonetic sciences*, **371**, 404.

GABLER P., GEIGER B. C., SCHUPPLER B. & KERN R. (2023). Reconsidering read and spontaneous speech : Causal perspectives on the generation of training data for automatic speech recognition. *Information*, **14**, 137. DOI : 10.3390/info14020137.

GRAFF D. (2002). An overview of broadcast news corpora. *Speech Communication*, **37**, 15–26. DOI : 10.1016/S0167-6393(01)00057-7.

IBRAHIM O., ASADI H., KASSEM E. & DELLWO V. (2020). Arabic speech rhythm corpus : Read and spontaneous speaking styles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5337–5342, Marseille, France : European Language Resources Association.

ITO K. & SPEER S. R. (2006). Using interactive tasks to elicit natural dialogue. *Methods in empirical prosody research*, p. 229–257.

JOHANSSON S., EBELING S. O. & OKSEFJELL S., Éds. (1998). *Corpora and cross-linguistic research : Theory, method and case studies*. Rodopi.

KOHLER K. J. (1996). Labelled data bank of spoken standard german : the kiel corpus of read/spontaneous speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, p. 1938–1941 : IEEE.

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M. & SONDEREGGER M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, p. 498–502.

MEREU D. & VIETTI A. (2021). Dialogic italian : the creation of a corpus of italian spontaneous speech. *Speech Communication*, **130**, 1–14. DOI : 10.1016/j.specom.2021.03.002.

O'SHAUGHNESSY D. (2008). Automatic speech recognition : History, methods and challenges. *Pattern Recognition*, **41**, 2965–2979. DOI : 10.1016/j.patcog.2008.05.008.

PRIMEWORDS INFORMATION TECHNOLOGY CO. L. (2018). Primewords chinese corpus set 1. https://www.primewords.cn.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518.

SHIH C. (2005). Understanding phonology by phonetic implementation. In *Ninth European Conference on Speech Communication and Technology*.

SURENDRAN D., NIYOGI P. *et al.* (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. *Amsterdam studies in the theory and history of linguistic science series 4*.

THOMPSON H. S., ANDERSON A. H., BARD E. G., DOHERTY-SNEDDON G., NEWLANDS A. & SOTILLO C. (1993). The hcrc map task corpus : Natural dialogue for speech recognition. In *Human Language Technology : Proceedings of a Workshop Held at Plainsboro, New Jersey*.

TORREIRA F. & ERNESTUS M. (2010). The nijmegen corpus of casual spanish. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, p. 2981–2985.

TSENG S.-C. (2019). Ilas chinese spoken language resources. *Proceedings of LPSS 2019*, p. 13–20.

YANG Z., CHEN Y., LUO L., YANG R., YE L., CHENG G., XU J., JIN Y., ZHANG Q., ZHANG P. *et al.* (2022). Open source magicdata-ramc : A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv :2203.16844*. DOI : 10.48550/arXiv.2203.16844.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Déplacement vertical du larynx dans la production des plosives en thaï

Paula Alejandra Cano Córdoba[1], Thi-Thuy-Hien Tran[1], Nathalie Vallée[1], Christophe Savariaux[1], Silvain Gerber[1], Nicha Yamlamai[2], Yu Chen[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
(2) Faculté de Lettres, Univ. Silpakorn, Bangkok, Thaïlande
prénom.nom@gipsa-lab.fr, yamlamai_n@su.ac.th,
yu.chen1@etu.univ-grenoble-alpes.fr

RESUME

Les plosives, généralement accompagnées d'un *burst* (relâchement audible) après la phase d'occlusion, sont néanmoins produites sans *burst* dans certaines langues d'Asie comme le thaï. Cette absence de bruit est attribuée au non-relâchement brusque des articulateurs et est observée exclusivement lorsque les plosives sont en finale de syllabe, jamais en initiale. Nous formulons l'hypothèse qu'un mouvement d'abaissement du larynx pourrait provoquer une diminution de la pression intraorale pendant la tenue de l'occlusion induisant le non-relâchement articulatoire. Nous avons examiné le mouvement vertical du larynx chez deux locutrices natives lors de la production des plosives /p, t, k/ dans une tâche de lecture d'une liste de pseudo-mots de structure CVC. Les résultats montrent une grande variabilité dans le mouvement d'abaissement du larynx en fonction des segments consonantiques, vocaliques et du contexte tonal, suggérant que plusieurs facteurs pourraient être impliqués dans l'explication de la diminution de la pression intraorale.

ABSTRACT

Larynx vertical movement in the production of Thai plosives
Plosives are generally produced with a burst (i.e. an audible release) which follows the occlusion phase. However, this is not the case in some Asian languages like Thai. The lack of burst comes from non-abrupt release of articulators and is observed exclusively when plosives are syllable-final, never initial. We suggest that a lowering movement of the larynx could cause a decrease in intra-oral pressure during the closure phase, inducing articulatory non-release. We examined the vertical movement of the larynx in two native speakers during the production of the plosives /p, t, k/ in a reading task of a CVC pseudo-word list. The results reveal a great variability in laryngeal lowering movement as a function of consonant, vowel and tonal contexts, suggesting several factors that may be involved in explaining decrease in intra-oral pressure.

MOTS-CLÉS : Plosives non relâchées, EGG, suivi du larynx, pression intraorale
KEYWORDS: Unreleased stops, EGG, larynx tracking, intraoral pressure

1 Introduction

Les consonnes plosives, présentes dans toutes les langues du monde (Maddieson, 1984 ; Vallée et al., 2002), sont généralement produites avec un blocage de l'air provenant des poumons au niveau

supraglottique, entraînant une augmentation de la pression derrière l'obstruction, ce qui provoque un écartement soudain des articulateurs (Ladefoged & Maddieson, 1996 ; Stevens, 2000). Sur le plan acoustique, ces consonnes se caractérisent par un silence, correspondant à l'occlusion, suivi d'un bruit de plosion court et audible, également appelé *burst*, résultant de la séparation brusque des articulateurs (Ladefoged & Maddieson, 1996 ; Stevens, 2000). Le son de relâchement varie en fonction du point d'articulation de la consonne, reflétant la forme de la cavité de résonance située après le blocage sur le parcours du flux d'air (Stevens, 1997). Néanmoins, dans certaines langues, les plosives sont dépourvues de relâchement audible, spécifiquement en position finale (coda) de syllabe. C'est le cas, par exemple, en vietnamien, en cantonais, en karitiana ou en thaï (Iwata et al., 1990 ; Tingsabadh & Abramson, 1993 ; Storto & Demolin, 2002 ; Tsukada, 2004 ; Tran, 2011 ; Yamlamai & Tran, 2018). La position de coda est propice à des phénomènes d'affaiblissement acoustique et articulatoire, ou de lénition (Kingston, 2008). Ces processus conduisent à la restriction des inventaires consonantiques dans les langues, où les différentes configurations trouvées en position initiale (attaque) sont neutralisées en position de coda (Kingston, 2008). Par exemple, en coréen, trois configurations laryngales sont observées en attaque (fortis /p, t, k/, lenis /p*, t*, k*/ et aspirée /p^{h}, t^{h}, k^{h}/) tandis qu'une seule réalisation allophonique non voisée et non relâchée [p̚, t̚, k̚] est attestée en coda (Cho et al., 2002). Ce processus de neutralisation des plosives en attaque aboutissant à des allophones non relâchés en coda est également observé dans d'autres langues telles que le cantonais, le thaï et le vietnamien (Iwata et al., 1990 ; Tran, 2011). En cantonais, par exemple, les plosives initiales non voisées, qu'elles soient aspirées ou non (ex. /p, p^{h}/) sont toutes réalisées comme non relâchées en coda de syllabe ([p̚]) (Iwata et al., 1990).

Malgré le fait que cette tendance soit bien décrite sur le plan phonologique, l'explication du processus de non relâchement n'est pas encore très claire, parfois sujette à controverse. D'une part, dans certaines langues comme le thaï ou le hakka, il existe un processus de renforcement glottique avec adduction des plis vocaux qui arrête le flux d'air empêchant l'accumulation de pression dans les cavités supraglottiques et conduisant à la suppression du burst (Iwata et al., 1990 ; Edmondson et al., 2010, 2011). En l'absence de glottalisation comme dans le cas des plosives finales du vietnamien (Michaud, 2004), ce type de production non relâchée pourrait nécessiter un contrôle particulier du relâchement de l'occlusion, soulevant ainsi des interrogations sur la complexité de la réalisation allophonique en position lénifiante. Ladefoged & Maddieson (1996), par exemple, suggèrent que ce phénomène résulterait d'une fuite d'air dans les fosses nasales pendant l'occlusion, entraînant ainsi une diminution de la pression intraorale (PIO). Ils notent que « (…) *nasal release occurs in some of the languages which are usually described as having unreleased final stops. A good example is Vietnamese. In this language, word-final stops are usually released, but the release is by lowering the velum while the oral closure is maintained, so that a short voiceless nasal is produced* » (Ladefoged & Maddieson, 1996 : 129). La même tendance a été observée chez des locuteurs du karitiana, langue amérindienne parlée au Brésil, où une augmentation systématique du débit d'air nasal pendant la phase de maintien de l'occlusion des plosives a été remarquée par Storto & Demolin (2002). Cette étude a également révélé une durée plus longue de la phase d'occlusion pour les consonnes non relâchées. Cependant, les fuites nasales n'ont pas été systématiquement observées lors de la production des plosives non relâchées dans des études plus récentes portant sur les données aérodynamiques et glottographiques du coréen, du vietnamien et du thaï (Tran et al., 2020 ; Cano Córdoba et al., 2022). En revanche, ces travaux ont mis en évidence un abaissement plus important du larynx pour les plosives non relâchées en finale de syllabe par rapport à celles relâchées en position initiale, ce qui pourrait expliquer une diminution de la pression intraorale et ainsi remplir les conditions aérodynamiques favorables au non relâchement.

Bien que les recherches sur les actions laryngées aient été abondantes, recourant à diverses méthodes instrumentales à la fois invasives et non invasives ainsi qu'à des techniques de

modélisation computationnelle (voir Esling et al., 2019 pour une revue), peu d'études ont mesuré le mouvement vertical du larynx. Ceci est principalement dû aux défis technologiques inhérents à cette mesure et à l'absence, jusqu'à présent, d'une méthodologie de mesure standard établie (Kleiner et al., 2023). Il semble y avoir un consensus sur des découvertes telles que la corrélation positive entre hauteur du larynx et f_0, c'est-à-dire l'abaissement du larynx en tant que corrélat de l'abaissement de la f_0 (Ohala, 1978 ; Sagart et al., 1986 ; Hallé, 1994 ; Moisik et al., 2014) explicable par la contraction des muscles infra-hyoïdiens qui provoque une descente globale du larynx et une diminution du pitch. En revanche, la contraction des fibres des muscles supra-hyoïdiens provoque une élévation du larynx et une augmentation de f_0 (Trigo, 1991 ; Honda et al., 1999). En outre, la hauteur du larynx est activement impliquée dans la production des tons comme en mandarin, dans la mesure où elle pourrait compenser l'absence d'implication de la musculature principale responsable de la régulation de la hauteur (Moisik et al., 2014).

Par ailleurs, des régularités concernant le lien entre la position verticale du larynx et la production des voyelles ont été établies dans des travaux antérieurs. Les trois voyelles cardinales /i, a, u/ ont respectivement une position moyenne haute, intermédiaire et basse du larynx (Ewan & Krones, 1974). Il est également reconnu que les voyelles postérieures telles que /u, o/ ont une position laryngée plus basse que les voyelles antérieures telles que /i, e/ selon les travaux de Hoole & Kroos (1998). Leur étude révèle également une corrélation entre la position du larynx et la labialité des voyelles antérieures en allemand : le larynx est plus bas pour les voyelles arrondies /y, ø/ que pour les non arrondies /i, e/. Ces résultats confirment les travaux antérieurs de Riordan (1977) et de Petersen (1983) respectivement sur le français et le néerlandais, indiquant que le larynx est en position plus haute pour /i/ que pour /u/. Néanmoins, Hoole (2006) affirme que même si la voyelle postérieure /u/ a clairement tendance à avoir une position plus basse du larynx, la position laryngée relative de /i/ et /a/ n'est pas aussi nette.

2 Objectifs et hypothèses

La présente étude consiste à explorer des pistes permettant de mieux comprendre les mécanismes sous-jacents au non relâchement des plosives en position finale en thaï, parmi lesquelles le mouvement du larynx lors de la production de ces consonnes.

Le lexique du thaï est majoritairement mono- et dissyllabique avec respectivement 41,37 % et 40,35 % des lemmes (Rousset, 2004) et présente une structure syllabique prédominante CVC (64,41 % des syllabes). De manière plus générale, les syllabes fermées sont favorisées (69 % des syllabes relevées dans le lexique. Rousset, 2004). En thaï, huit plosives sont autorisées en initiale /p, p^h, b, t, t^h, d, k, k^h/ (Abramson & Tingsabadh, 1999) présentant trois modes phonatoires différents (Lisker & Abramson, 1964). La distinction phonologique observée en initiale de syllabe se voit neutralisée en position de coda où seulement trois allophones plosifs sont permis (Iwata et al., 1990), réalisés comme non voisés et non relâchées [p̚, t̚, k̚] (Tingsabadh & Abramson, 1993).

Une descente plus importante du larynx aurait pour effet de diminuer ce qui, selon Stevens (1997), est nécessaire pour produire un burst audible : la montée de la pression intraorale derrière le lieu d'articulation de la consonne. En considération des résultats de Shipp et collègues (1987) qui ont mis en évidence une descente systématique du larynx lors de la production de segments non voisés par rapport à des segments voisés, il est attendu que l'abaissement du larynx soit plus marqué pour une plosive sourde en finale si elle est suivie d'un segment non voisé que pour une plosive sourde en initiale devant voyelle. Il a été observé que la position de la consonne dans la syllabe a une influence sur le degré d'abaissement du larynx en vietnamien, coréen et thaï. En effet, en l'absence de fuite nasale ou d'adduction des plis vocaux, le larynx descend davantage en C_2 où se trouvent

les plosives non relâchées (Tran et al., 2020 ; Cano Córdoba et al., 2022). De plus, une corrélation a été établie entre le lieu d'articulation de la plosive et le déplacement vertical du larynx : plus la consonne est postérieure, plus le larynx descend (Cano Córdoba et al., 2022).

L'hypothèse générale proposée dans cette étude suggère que l'absence de relâchement audible des plosives finales pourrait résulter d'un mouvement d'abaissement du larynx pendant la phase de maintien de l'occlusion (Cano Córdoba et al., 2022), une caractéristique également observée en vietnamien et en coréen (Tran et al., 2020). En outre, étant donné que la position verticale du larynx est proposée directement corrélée à la hauteur de la f_0 et à l'articulation des voyelles (voir les études antérieures citées plus haut), nous cherchons à savoir si le mécanisme laryngé dans les productions non relâchées peut être impacté par le registre tonal porté par la syllabe, ainsi que par le contexte vocalique.

3 Méthodologie

3.1 Matériel

Les données glottographiques analysées ont été recueillies avec un électroglottographe multicanal (EGG réf. EG2-PCX2 de Glottal Enterprises Inc.)(Rothenberg, 1992). Deux électrodes fixées par un dispositif autoagrippant à un collier, ont été placées de part et d'autre du larynx des participants pour détecter à la fois les oscillations des plis vocaux (mouvements d'ouverture et de fermeture de la glotte) et le déplacement vertical du larynx lors de la production de parole. Le signal acoustique de la parole a été enregistré à l'aide d'un micro AKG C1000S et d'un enregistreur Marantz PMD 670, à une fréquence d'échantillonnage de 44.1 kHz. Les données acoustiques ont également été enregistrées en synchronisation avec les données EGG en utilisant le système d'acquisition Biopac MP150 via le logiciel *Acqnowlegdge* à une fréquence d'échantillonnage de 25 kHz. Aucun filtrage n'a été réalisé sur les données avant analyse. Le corpus a été enregistré dans la chambre anéchoïque de l'Université Silpakorn, à Nakhon Pathom, en Thaïlande, durant l'été 2023.

3.2 Corpus et locuteurs

Pour cette étude, nous avons sélectionné une liste de pseudo-mots de type C_1VC_2. Nous avons inclus dans notre étude toutes les combinaisons possibles où C_1 et C_2 correspondent aux plosives non voisées et non aspirées /p, t, k/, ainsi que tous les contextes vocaliques possibles en thaï /a ɛ e i ɔ o u ɤ ɯ/. Pour observer toute différence éventuelle selon le registre tonal, deux tons, bas descendant T2 /a^{21}/ et haut montant T4 /a^{45}/ (Ladefoged & Maddieson, 1996), ont été sélectionnés pour la syllabe cible. Les pseudo-mots ainsi obtenus ont été insérés dans la phrase porteuse พูดว่า __ ดี ๆ [p^{h}û:t^{51} wâ:^{51} __ di:^{32} di:^{32}] "Dis __ attentivement". Le corpus est constitué de deux répétitions de la liste de phrases, où les pseudo-mots cibles de chaque répétition ont été mis en ordre aléatoire différent, mais identique, pour tous les participants. La consigne a été donnée de lire à voix haute, à un débit normal et fluide les phrases qui s'affichaient une par une sur l'écran (dont la taille était de 15'').

Nous présentons ci-dessous les premières analyses effectuées auprès de deux locutrices natives (au total 20 locuteurs ont été enregistrés pour ce projet). Originaires de la région de Bangkok, elles parlent la même variété dialectale du centre de la Thaïlande. L'une d'entre elles (âgée de 49 ans)

n'avait pas été au contact de langues étrangères (TH_FG) tandis que l'autre locutrice (TH_FE), âgée de 18 ans, était étudiante en première année du département de français de l'Université Silpakorn de Nakhon Pathom, Thaïlande, au moment de l'enregistrement. Les résultats sont issus de 81 pseudo-mots * 2 tons * 2 répétitions * 2 locutrices, soit un total de 648 stimuli cibles.

3.3 Mesures et analyses

Le signal acoustique des productions de chaque locutrice a été segmenté et annoté manuellement avec Praat (v.6.2.23, Boersma & Weenink, 2022). Le traitement des données physiologiques a été effectué avec TRAP (TRaitement Automatique de la Parole), développé sous Matlab en interne au GIPSA-lab (Savariaux, 2017) et consistait à l'étiquetage des événements glottographiques sur le signal EEG. (FIGURE 1).

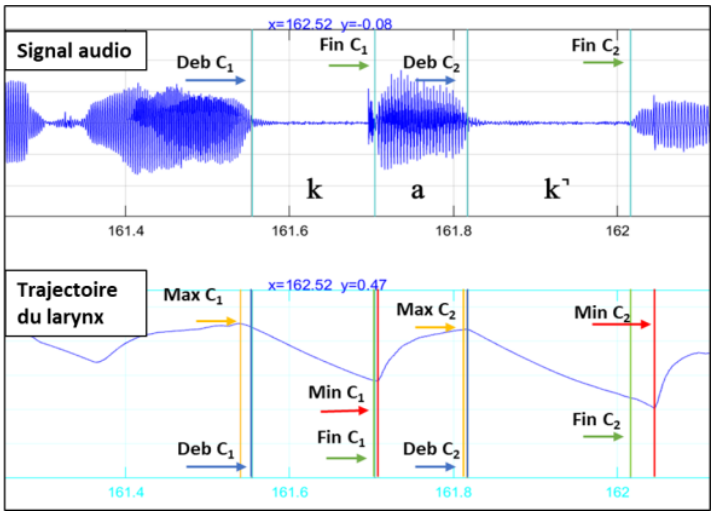


FIGURE 1: Visualisation de la fenêtre de travail du programme TRAP avec les 4 points repérés (Maximum PVL, Minimum PVL, Début de la consonne, Fin de la consonne) pour C_1 et C_2 sur la trajectoire du larynx lors de la réalisation de la syllabe /kak/ réalisé [kak˺] par TH_FE4.

Les tracés du déplacement du larynx pouvant être liés à des mouvements de la tête et/ou des électrodes de l'EGG ont été exclus des analyses. Au total, 269 réalisations ont été écartées, comprenant 140 réalisations de C_1 (21 %) et 129 réalisations de C_2 (20 %). Les données glottographiques analysées correspondent à la position verticale du larynx (dorénavant PVL). Quatre événements ont été repérés le long de la trajectoire PVL : (1) le début acoustique de la consonne cible (Deb) ; (2) la fin acoustique de la consonne (Fin) ; (3) la position maximale de la PVL (Max) ; (4) la position minimale de la PVL (Min). Généralement, la valeur maximale de PVL se situe autour du début acoustique de la consonne et la valeur minimale autour de la fin acoustique de celle-ci. Pour décrire le mouvement de déplacement vertical du larynx pendant la réalisation des consonnes cibles, quatre calculs de différence de PVL (ΔPVL) ont été effectués à partir de ces points de mesure : (1) ΔPVL_{FD} = Fin - Deb, (2) ΔPVL_{mM} = Min - Max, (3) ΔPVL_{MD} = Max - Deb, (4) ΔPVL_{mF} = Min - Fin. Étant donné que le matériel EGG utilisé ne dispose d'aucun paramètre de calibrage, les quatre différences d'amplitude du mouvement du larynx pour un locuteur donné ont ensuite été calibrées en les divisant par la valeur maximale de PVL (PVLMax) du locuteur. Par exemple, le pourcentage de l'abaissement du larynx ΔPVL_{FD} du locuteur j est calculé avec la formule suivante : $100*\Delta PVL_{FDj}/PVL_{Maxj}$. Cependant, parmi les quatre calculs de ΔPVL initialement proposés, celui qui semble le plus pertinent pour rendre compte de l'amplitude et de la trajectoire du larynx pendant la production des plosives est ΔPVL_{FD}. Par conséquent, nous présentons ci-dessous les analyses de cette variable.

3.4 Analyses statistiques

Pour chacune des deux participantes, nous avons étudié l'impact de plusieurs variables explicatives et de leurs interactions sur la variable réponse ΔPVL_{FD}. Les variables explicatives sont CONSONNE (p, t, k), TON (T2 et T4), POSITION (initiale C_1 et finale C_2) et VOYELLE (a, e, i, o, u, ɛ, ɔ, ɤ, ɯ). Les valeurs de la variable réponse sont bornées dans l'intervalle [-100, 0]. Compte tenu qu'un mot est prononcé plusieurs fois, nous avons introduit la variable *STIMULI* comme effet aléatoire dans le modèle. Dès lors, comme nous sommes en présence d'une variable réponse dont les valeurs sont bornées, et de mesures répétées, nous avons choisi d'utiliser une régression beta avec effets aléatoires réalisée à l'aide de la fonction *glmmTMB* du package *glmmTMB* du logiciel R. Pour pouvoir utiliser la régression beta, nous avons appliqué une bijection sur les valeurs de la variable réponse pour passer de l'intervalle [-100, 0] à l'intervalle [0, 1]). Afin d'évaluer si les facteurs fixes ainsi que leur interaction ont un impact significatif sur la variable réponse, nous avons utilisé des tests de modèles emboités en partant du modèle initial et en réalisant une sélection descendante avec $p \leq 0.05$. Les tests ont été réalisés à l'aide de la fonction *anova* et du package *DHARMa* de R. En complément de l'étape de sélection du modèle, nous avons réalisé des comparaisons multiples (Hothorn et al., 2008) en utilisant la fonction *glht* du package *multcomp* de R d'où ont été tirées les valeurs de p présentées ci-dessous. La matrice de contrastes a été calculée avec la fonction *emmeans* du package *emmeans*.

4 Résultats

Afin d'observer les effets en inter-sujet, l'amplitude du déplacement vertical du larynx est estimée par rapport au point de référence correspondant spécifiquement à la valeur la plus élevée de la PVL relevée chez l'ensemble des locutrices. Parmi les deux sujets, TH_FG présente le plus grand abaissement de la PVL. Le point de référence est donc observé chez cette locutrice avec une valeur maximale absolue de 0,89.

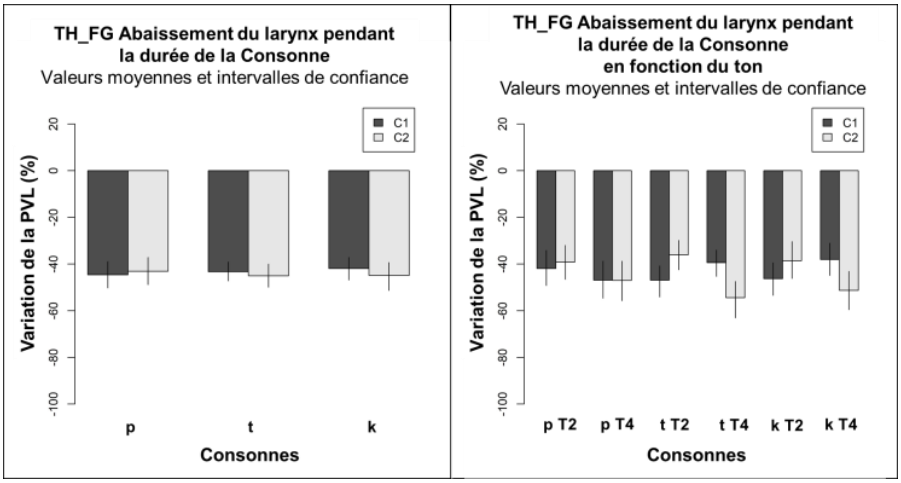


FIGURE 2: Valeurs moyennes et intervalles de confiance pour la descente du larynx, pendant la réalisation de /p, t, k/ chez la locutrice TH_FG, en fonction de la position de la consonne dans la syllabe (à gauche) et en fonction du contexte tonal (bas T2 ou haut T4) et de la position de la consonne dans la syllabe (à droite).

Lors de la segmentation des données acoustiques, nous avons remarqué dans les répétitions de la locutrice TH_FE que le segment sonore /d/ qui suit la consonne finale C_2 de la syllabe cible se dévoise régulièrement. Ainsi, un silence est présent sur le signal acoustique entre C_2 non relâchée et /d/ dévoisée, ce qui rend difficile la détermination de la durée de C_2 pour cette locutrice. Par conséquent, les données glottographiques de cette locutrice ne nous permettent pas d'étudier l'hypothèse principale de ce travail concernant une comparaison de la PVL entre C_1 et C_2. Afin de pouvoir continuer la segmentation et observer les C_2 non relâchées entre elles, il a été décidé de placer la frontière entre C_2 et C_3 en nous basant sur l'intensité. La fin de C_2 est ainsi déterminée par le point où la courbe d'intensité commence à remonter.

Concernant la locutrice TH_FG, la tendance de la descente du larynx présente des variations entre les consonnes en position initiale et finale. Le larynx descend un peu plus en C_2 pour /t/ et /k/ mais en C_1 pour /p/. Cependant, ces différences entre C_1 et C_2 ne sont pas significatives.

Une analyse plus détaillée en fonction du contexte tonal, révèle des différences dans le mouvement du larynx en position initiale et finale. Par exemple, pour le ton bas (T2) en position initiale, le larynx descend davantage pour la coronale (47 %), puis la vélaire (46 %), et moins pour la bilabiale (41 %), tandis qu'en position finale, cette distribution est inversée : le larynx descend davantage pour /p/ et /k/ (39 %) et moins pour /t/ (36 %). En fonction de la position de la plosive dans la syllabe, c'est en C_1 où une descente plus importante est relevée pour les trois lieux d'articulation (/p/ C_1 : 42 % *vs* C_2 : 39 %, /t/ C_1 : 50 % *vs* C_2 : 36 % et /k/ C_1 : 46 % *vs* C_2 : 38 %). En contexte tonal haut (T4), la descente du larynx est marquée pour la consonne bilabiale en C_1 : le larynx descend plus pour la consonne /p/ (47 %), suivie de /t/ (39 %), et finalement de /k/ (38 %). En C_2, le larynx descend plus notablement pour la consonne coronale (54 %) et moins pour la bilabiale (46 %). À l'intérieur des consonnes, une descente plus importante en C_2 qu'en C_1 est observée pour /t/ (54 % *vs* 39 %) et /k/ (51 % *vs* 38 %), tandis que la différence pour /p/ est quasi inexistante (46,98 % en C_2 *vs* 46,91 en C_1). En dépit de ces constatations, le facteur CONSONNE et toutes les interactions associées ont été statistiquement éliminées par la sélection descendante. En conséquence, les analyses des comparaisons entre C_1 et C_2 pour cette locutrice ne concernent que les interactions des variables TON et VOYELLE. Suite au traitement statistique résultant, les différences significatives trouvées sont limitées aux voyelles /i/ et /e/ sous le ton bas T2 ($z = -3.282$, $p = 0.016$ et $z = -3.01$, $p = 0.039$ respectivement) où le larynx descend davantage en position initiale, et à la voyelle /ɔ/ sous registre tonal haut T4 ($z = 2.955$, $p = 0.047$) avec un abaissement plus important en position finale qu'initiale.

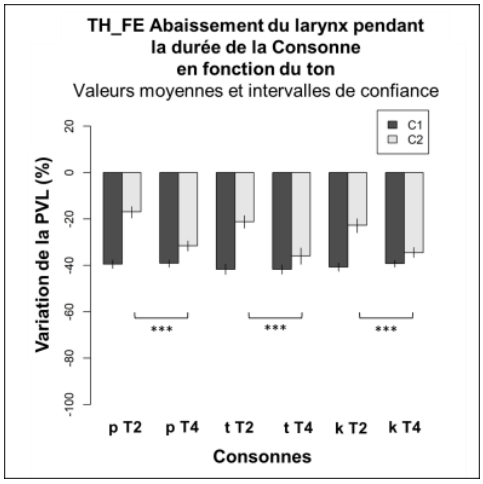


FIGURE 3: Valeurs moyennes et intervalles de confiance de la variation de la descente du larynx pendant la réalisation de /p, t, k/ chez la locutrice TH_FE, en fonction du contexte tonal (bas T2 ou haut T4) et de la position de la consonne dans la syllabe.

Rappelons que pour TH_FE, en raison du voisement du segment qui suit la syllabe cible, nous n'avons pas inclut ces données dans les analyses qui comparent la PVL entre C_1 et C_2. Néanmoins, ces données ont également révélé des tendances en fonction du contexte tonal. En position C_2, la différence de mouvement du larynx est plus ou moins prononcée selon le ton porté par la syllabe : une descente plus importante pour le ton haut que pour le ton bas est observée. Les analyses statistiques ont révélé des différences significatives pour les trois consonnes /p, t, k/ en fonction du ton ($z = 16.672$, $p = <0.001$ pour les trois consonnes), ainsi que pour la combinaison de facteurs CONSONNE*VOYELLE, à l'exception des voyelles fermées postérieures /u, ɯ/ ($z = 2.986$, $p = 0.073$ et $z = 2.966$, $p = 0.077$ respectivement). La différence entre T2 et T4 pour les consonnes en position finale s'est révélée également significative chez la locutrice TH_FG pour toutes les voyelles ($z = 3.817$; $p = 0.002$).

5 Conclusion

Cette étude qui est une partie d'un projet en cours vise à déterminer les aspects physiologiques des plosives non relâchées du thaï. Face aux résultats de travaux antérieurs qui attestent une absence de fuite nasale ou de fermeture glottique (Iwata et al., 1990 ; Michaud, 2004 ; Tran et al., 2020 ; Cano Córdoba et al., 2022), une analyse du mouvement vertical du larynx est proposée en considérant plusieurs variables, notamment le lieu d'articulation, la position syllabique de la consonne, le contexte vocalique, ainsi que le registre tonal. En accord avec les travaux antérieurs de Shipp et al. (1987), une descente du larynx a été observée lors de la production des plosives sourdes. Par ailleurs, entre la position initiale et la position finale de syllabe, une variation selon le contexte tonal a été trouvée.

Malgré la variabilité et l'absence de significativité statistique pour les données de TH_FG, une descente plus marquée du larynx a été constatée en C_2 qu'en C_1 pour les consonnes coronale et vélaire. Une analyse plus détaillée du comportement du larynx en fonction du ton a révélé des tendances spécifiques selon le registre tonal porté dans la syllabe. Pour la plupart des contextes vocaliques sous le ton haut (T4), une descente plus importante en C_2 qu'en C_1 a été observée, tandis que pour le ton bas (T2), le cas inverse a été relevé : c'est en position initiale que le larynx s'abaisse davantage. Cet effet du registre tonal sur le mouvement du larynx pourrait s'expliquer par le rôle de sa position dans la diminution de la hauteur de la fréquence fondamentale (Ohala, 1978 ; Trigo, 1991 ; Honda et al., 1999). Ainsi, lorsque le noyau syllabique se trouve sous le registre haut, le larynx est dans une position plus élevée que pour un registre tonal bas. En l'absence de fuite nasale ou de fermeture glottique, si un geste de descente de larynx est opéré pour donner lieu au non relâchement, celui-ci devrait être plus ample pour une consonne qui succède à une voyelle sous le ton haut que sous le ton bas. En conséquence des observations réalisées au cours de cette étude, il semble important de poursuivre l'exploration des stratégies individuelles qui pourraient expliquer le non relâchement des plosives finales en thaï. Le traitement et l'analyse des productions de deux autres locuteurs issues de la même campagne d'acquisition de données sont en cours.

Remerciements

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-15-IDEX-02 » (PAALAS : Perceptual-Acoustic cues of Asian-LAnguage final Stop consonants).
Les auteurs remercient l'IDEX pour sa contribution au financement de la mobilité de PACC.
Nous remercions vivement tous les locuteurs qui ont participé aux enregistrements.

Références

ABRAMSON, A. S., & TINGSABADH, K. (1999). Thai Final Stops: Cross-Language Perception. Phonetica, 56(3-4), 111-122. https://doi.org/10.1159/000028446

BOERSMA, P., & WEENINK, D. (2022). Praat: Doing phonetics by computer [Computer program]. Version 6.2.23. http://www.praat.org/

CANO CÓRDOBA, P. A., TRAN, T. T. H., VALLÉE, N., SAVARIAUX, C., GERBER, S., & YAMLAMAI, N. (2022). Caractérisation des consonnes plosives non relâchées du thaï : Une étude électro glottographique. XXXIIIIèmes Journées d'Etude sur la Parole, 361-370.

CHO, T., JUN, S.-A., & LADEFOGED, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. Journal of Phonetics, 30(2), 193-228. https://doi.org/10.1006/jpho.2001.0153

EDMONDSON, J. A., CHANG, C. B., HUANG, H. J., HSIEH, F., & PENG, Y. (2010). Reinforcing voiceless stop coda in Taiwanese, Vietnamese and other East and Southeast Asian languages: Laryngoscopic case studies. Labphon, 12. https://labphon.org/sites/default/files/previous_conferences/LabPhon12.pdf

EDMONDSON, J. A., CHANG, Y., HSIEH, F., & HUANG, H. J. (2011). Reinforcing voiceless finals in Taiwanese and Hakka: Laryngoscopic case studies. Proceedings of the 17th International Congress of Phonetic Sciences, 627-630.

ESLING, J. H., MOISIK, S. R., BENNER, A., & CREVIER-BUCHMAN, L. (2019). Voice quality: The laryngeal articulator model. Cambridge University Press.

EWAN, W. G., & KRONES, R. (1974). Measuring larynx movement using the thyroumbrometer. Journal of Phonetics, 2(4), 327-335. https://doi.org/10.1016/S0095-4470(19)31302-6

HALLÉ, P. A. (1994). Evidence for Tone-Specific Activity of the Sternohyoid Muscle in Modern Standard Chinese. Language and Speech, 37(2), 103-123. https://doi.org/10.1177/002383099403700201

HONDA, K., HIRAI, H., MASAKI, S., & SHIMADA, Y. (1999). Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control. Language and Speech, 42(4), 401-411. https://doi.org/10.1177/00238309990420040301

HOOLE, P. (2006). Experimental studies of laryngeal articulation [Thèse d'habilitation pour l'obtention de la Venia Legendi dans le domaine de la phonétique]. Ludwig-Maximilians-Universität.

HOOLE, P., & KROOS, C. (1998). Control of larynx height in vowel production. 5th International Conference on Spoken Language Processing (ICSLP 1998), paper 1097-0. https://doi.org/10.21437/ICSLP.1998-360

HOTHORN, T., BRETZ, F., & WESTFALL, P. (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal, 50(3), 346-363. https://doi.org/10.1002/bimj.200810425

IWATA, R., HIROSE, H., NIIMI, S., SAWASHIMA, M., & HORIGUCHI, S. (1990). Syllable final stops in East Asian languages: Southern Chinese, Thai, and Korean. Proceedings of the 1990 International Conference on Spoken Language Processing, 621-624. http://www.isca-speech.org/archive/icslp_1990

KINGSTON, J. (2008). Lenition. 1-31.

KLEINER, C., HÄSNER, P., & BIRKHOLZ, P. (2023). Intrinsic velocity differences between larynx raising and larynx lowering. PLOS ONE, 18(2), e0281877. https://doi.org/10.1371/journal.pone.0281877

LADEFOGED, P., & MADDIESON, I. (1996). The sounds of the world's languages. Blackwell.

LISKER, L., & ABRAMSON, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. WORD, 20(3), 384-422. https://doi.org/10.1080/00437956.1964.11659830

MADDIESON, I. (1984). Patterns of sounds. Cambridge University Press.

MICHAUD, A. (2004). Final Consonants and Glottalization: New Perspectives from Hanoi Vietnamese. Phonetica, 61(2-3), 119-146. https://doi.org/10.1159/000082560

MOISIK, S. R., LIN, H., & ESLING, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). Journal of the International Phonetic Association, 44(1), 21-58.

OHALA, J. J. (1978). Production of Tone. In V. A. Fromkin (Éd.), Tone (p. 5-39). Academic Press. https://doi.org/10.1016/B978-0-12-267350-4.50006-6

PETERSEN, N. R. (1983). The effect of consonant type on fundamental frequency and larynx height in Danish. Annual Report of the Institute of Phonetics University of Copenhagen, 17, 55-86. https://doi.org/10.7146/aripuc.v17i

RIORDAN, C. J. (1977). Control of vocal-tract length in speech. The Journal of the Acoustical Society of America, 62(4), 998-1002. https://doi.org/10.1121/1.381595

ROTHENBERG, M. (1992). A multichannel electroglottograph. Journal of Voice, 6(1), 36-43. https://doi.org/10.1016/S0892-1997(05)80007-4

ROUSSET, I. (2004). Structures syllabiques et lexicales des langues du monde. Données, typologies, tendances universelles et contraintes substantielles [Phdthesis, Université Stendhal - Grenoble III]. https://tel.archives-ouvertes.fr/tel-00250154

SAGART, L., HALLÉ, P. A., BOYSSON-BARDIES, B. DE, & ARABIA-GUIDET, C. (1986). Tone production in modern standard chinese: An electromyographic investigation. Cahiers de Linguistique - Asie Orientale, 15(2), 205-221. https://doi.org/10.3406/clao.1986.1204

SAVARIAUX, C. (2017). Trap (version 7.2) [Logiciel de traitement des signaux de parole] (Gipsa-lab Brevet).

SHIPP, T., GUINN, L., SUNDBERG, J., & TITZE, I. R. (1987). Vertical laryngeal position—Research findings and their relationship to singing. Journal of Voice, 1(3), 220-222. https://doi.org/10.1016/S0892-1997(87)80003-6

STEVENS, K. N. (1997). Articulatory-Acoustic-Auditory Relationships. In W. J. Hardcastle & J. Laver (Éds.), The Handbook of Phonetic Sciences (1re éd., p. 462-506). Blackwell. https://doi.org/10.1002/9781444317251.ch12

STEVENS, K. N. (2000). Acoustic phonetics (1. paperback ed). MIT Press.

STORTO, L., & DEMOLIN, D. (2002). The Phonetics and Phonology of Unreleased Stops in Karitiana. Annual Meeting of the Berkeley Linguistics Society, 28(1), 487-497. https://doi.org/10.3765/bls.v28i1.3860

TINGSABADH, M. R. K., & ABRAMSON, A. S. (1993). Thai. Journal of the International Phonetic Association, 23(1), 24-28. https://doi.org/10.1017/S0025100300004746

TRAN, T. T. H. (2011). Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde : De l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE (Numéro 2011GRENL028) [Theses, Université de Grenoble]. https://tel.archives-ouvertes.fr/tel-01568326

TRAN, T. T. H., VALLÉE, N., SAVARIAUX, C., KIM, I., & KIM, S. (2020). Caractérisation des plosives finales dans des langues d'Asie : Une étude multilingue du non relâchement. Actes des XXVIIIèmes Journées d'Etude sur la Parole, Nancy, France, 597-605.

TRIGO, L. (1991). On Pharynx-Larynx Interactions. Phonology, 8(1), 113-136.

TSUKADA, K. (2004). A Cross-linguistic Acoustic Comparison of Unreleased Word-final Stops: Korean and Thai. Proceedings of INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, 1293-1296. https://doi.org/10.21437/Interspeech.2004-76

VALLÉE, N., BOE, L. J., SCHWARTZ, J. L., BADIN, P., & ABRY, C. (2002). The Weight of Phonetic Substance in the Structure of Sound Inventories. ZAS Papers in Linguistics, 28, 145-168. https://doi.org/10.21248/zaspil.28.2002.163

YAMLAMAI, N., & TRAN, T. (2018). Effet de la position de la syllabe sur la réalisation acoustique des consonnes finales du thaï. Actes des XXXIIe Journées d'Études sur la Parole, 151-159. https://doi.org/10.21437/JEP.2018-18

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Détection automatique des schwas en français - Application à la détection des troubles du sommeil

Colleen Beaumard[1,2], Vincent P. Martin[3], Yaru Wu[4], Jean-Luc Rouas[1], Pierre Philip[2]
(1) LaBRI/UMR5800, 351 cours de la Libération, 33400 Talence, France
(2) SANPSY/6033, Place Amélie Raba-Léon, 33076 Bordeaux, France
(3) DDP Research Unit, Department of Precision Health, LIH, 1 A-B Rue Thomas Edison, 1445 Strassen, Luxembourg
(4) CRISCO/UR4255, Esplanade de la Paix, 14032 Caen, France
{colleen.beaumard, jean-luc.rouas}@labri.fr, vincentp.martin@lih.lu, yaru.wu@unicaen.fr, pierre.philip@u-bordeaux.fr

RÉSUMÉ

La Somnolence Diurne Excessive affecte négativement les individus et est un problème de santé publique. L'analyse de la parole pourrait aider les cliniciens à la surveiller. Nous nous sommes concentrés sur la détection du schwa /ə/ et avons trouvé un lien entre le nombre d'occurrences annoté manuellement et le niveau de somnolence des patients hypersomnolents d'un sous-ensemble du corpus TILE. Dans un second temps, afin de pouvoir généraliser ces résultats à l'intégralité du corpus, nous avons conçu un système de détection des schwas, robuste à la somnolence. Dans un troisième temps, nous avons étendu notre analyse à deux autres phonèmes supplémentaires /ø/ et /œ/. Nous avons ainsi observé une relation significative entre /ø/ et la combinaison des trois phonèmes et la somnolence subjective à court terme.

ABSTRACT

Automatic Schwa Spotting in French - Application to Pathological Sleepiness Detection

Excessive Daytime Sleepiness negatively affects both individuals and public health. Speech analysis could help clinicians monitor it. We focused on the detection of schwa /ə/ since it is a vowel whose realization is optional in French. We found a link between its manually annotated number and the sleepiness level of hypersomnolent patients of a subset of the MSLTc. We need a schwa spotting system robust to sleepiness to attest to this link in the whole corpus. We compared two different systems outputting either phonemes (baseline) or words (proposed) and found the second system was more robust to sleepiness, using manual annotation as our reference. Two other phonemes were considered due to their closeness to /ə/. Using automatic detection, we found a significant relationship between /ø/ and their combination, and short-term subjective sleepiness.

MOTS-CLÉS : Détection de schwa, Somnolence, Reconnaissance Automatique de la Parole.

KEYWORDS: Schwa Spotting, Sleepiness, Automatic Speech Recognition.

1 Introduction

La Somnolence Diurne Excessive (SDE) est associée à une grande variété de maladies (neurologiques, cardiovasculaires...) et affecte négativement la vie quotidienne et professionnelle des personnes qui en

souffrent (Barnes & Watson, 2019). Elle augmente également le risque de mortalité (pour les accidents de véhicule (Bioulac *et al.*, 2017), etc.) et de handicap (Jike *et al.*, 2018), qui sont des problèmes de santé publique. Le suivi des symptômes de la SDE est difficile étant donné que les tests sont effectués à l'hôpital pendant une journée, ce qui est à la fois chronophage pour le patient et coûteux pour l'hôpital. Ainsi, les cliniciens ont besoin d'un outil pour collecter ces symptômes régulièrement et dans des conditions écologiques. L'analyse automatique de la parole peut être employée pour surveiller la SDE. La collecte de données vocales peut également être simplifiée en étant effectuée via des smartphones en enregistrant la parole lue ou spontanée.

Plusieurs corpus ont été créés pour détecter automatiquement la somnolence en analysant la parole, chacun utilisant différents types d'annotation. La mesure de somnolence utilisée pour annoter les corpora Sleepy Language (SLC) (Schuller *et al.*, 2011) et SLEEP (Schuller *et al.*, 2019) n'est pas validée cliniquement, donc non-interprétable par les médecins du sommeil. L'annotation du dataset Voiceome (Tran *et al.*, 2022) ne permet pas de distinguer la somnolence de la fatigue (Maclean *et al.*, 1992). Le corpus TILE (Martin *et al.*, 2021b) (Test Itératif de Latence d'Endormissement) est le seul à notre connaissance à mesurer à la fois la somnolence subjective et physiologique avec des mesures validées par les cliniciens.

Bien que de nombreuses tentatives aient été faites pour étudier la détection automatique de la somnolence par l'analyse de la voix, les corpora utilisés sont presque exclusivement composés de parole lue. Ainsi, malgré les performances encourageantes obtenues par les derniers systèmes, certaines des caractéristiques utilisées sont spécifiques à ce type de parole (Martin *et al.*, 2021a, 2022).

Puisque notre objectif est de pouvoir classifier automatiquement la somnolence de manière écologique, nous avons besoin d'évaluer des caractéristiques liées à la somnolence spécifiques à l'analyse de la parole spontanée. Le contrôle neuro-moteur et la planification cognitive peuvent avoir un impact sur la parole élicitée par les sujets somnolents (Harrison & Horne, 1997), nous avons donc décidé de nous concentrer en premier lieu sur l'étude de phonèmes spécifiques qui peuvent être influencés par un tel phénomène, et notamment le schwa.

En français, le schwa /ə/ est une voyelle centrale optionnelle - ce qui signifie que sa prononciation ou son absence ne change pas le sens du mot (Bürki *et al.*, 2011; Durand, 2014). Par exemple, /dəmɛ̃/ « demain » peut être prononcé [dəmɛ̃] ou [dmɛ̃] sans changer son sens. Une pause remplie peut également être produite comme une voyelle de type schwa ([ə ː], « euh »). Sa présence ou son absence est liée à de nombreux facteurs tels que l'accent, la coarticulation, ou le type de parole. L'analyse phonétique peut être appliquée à la fois à la parole spontanée et à la lecture, ce qui pallie au manque de corpus de parole spontanée spécifique à la somnolence. Un corpus nommé Medispeech est actuellement en cours d'enregistrement au Service Universitaire de Médecine du Sommeil (SUMS) du CHU de Bordeaux et contient à la fois de la parole lue et de la parole spontanée. En attendant d'avoir un nombre suffisant d'enregistrement, nous avons utilisé un sous-ensemble du corpus TILE contenant 20 patients avec les plus importantes variations de somnolence à court terme (Martin *et al.*, 2023a) et avons montré que le nombre de schwas annotés manuellement est lié au niveau de somnolence (Beaumard *et al.*, 2023).

Notre objectif est de concevoir un système de détection automatique de schwas – c'est-à-dire détecter la présence (ou non) de schwa dans un signal de parole – sachant qu'elle ne doit pas être impactée par le niveau de somnolence des patients (robuste à la somnolence), et de le valider sur une base de données étiquetée à la main. Ce système peut ensuite être utilisé pour détecter automatiquement le schwa sur un corpus plus large (le corpus TILE) afin de confirmer nos précédents résultats obtenus.

Cet article est organisé comme suit : la Section 2 décrit le corpus utilisé, le sous-ensemble annoté manuellement à des fins d'évaluation ainsi que la méthodologie d'annotation. Dans la Section 3, nous présentons les systèmes de détection des schwas de réference et proposés. Leurs performances sont présentées et comparées dans la Section 4. Enfin, la Section 5 discute de l'utilité du système de détection de schwas conçu pour l'évaluation automatique de la somnolence.

2 Description des données

2.1 Le corpus TILE

Le corpus TILE (Martin *et al.*, 2021b) contient 660 enregistrements de 132 patients hypersomnolents du SUMS du CHU de Bordeaux. Chaque patient a effectué un Test Itératif de Latence d'Endormissement (test éponyme du corpus) consistant en 5 opportunités de sieste de 20 minutes toutes les 2 heures, de 9h à 17h. Leur latence d'endormissement, c'est-à-dire le temps entre le début du test et le moment où le patient s'endort ou la fin du test, est mesurée à chaque occurrence du test. C'est une mesure de référence de la somnolence physiologique à long terme (Arand *et al.*, 2005; Martin *et al.*, 2023b). Avant chaque opportunité de sieste, les patients sont enregistrés en train de lire à voix haute un texte différent du *Petit Prince* (A. de Saint-Exupéry) d'environ 250 mots. Les patients ont également renseigné deux échelles cliniques : l'échelle de somnolence de Karolinska (KSS) et l'échelle de somnolence d'Epworth (ESS). Le premier mesure le niveau de somnolence subjective instantanée sur une échelle de Lickert à 9 points. Les patients remplissent ce questionnaire avant chaque lecture (somnolence subjective à court terme). Le second questionnaire mesure le niveau de somnolence subjective à long terme avec des questions sur des situations quotidiennes. Les patients remplissent ce questionnaire une fois.

Nous utilisons un sous-ensemble du corpus TILE (Martin *et al.*, 2023a) contenant 100 enregistrements de 20 patients hypersomniaques ayant les plus fortes variations de somnolence à court terme (KSS) dans le but de mesurer l'impact du niveau de somnolence sur les performances de détection automatique de schwas.

Ce corpus et les descriptions de son sous-ensemble sont présentés dans la Table 1.

Corpus	**TILE**	**Sous-ensemble**
Durée	14h 10m 12s	2h 11m 47s
#Patients (hommes/femmes)	132 (81/51)	20 (10/10)
#Enregistrements	660	100
KSS moyenne (sd)	4 (2)	5 (2)
Latence d'endormissement moyenne (sd)	12 (6)	14 (6)

TABLE 1 – Description du corpus TILE et du sous-ensemble utilisé

2.2 Annotation manuelle

Nous avons tout d'abord transcrit les textes originaux en utilisant le Lexique 3.83 (New *et al.*, 2004), qui contient la prononciation française standard d'environ 140 000 mots. Le /ə/ étant phonétiquement

proche du /ø/ (« deux ») et du /œ/ (« neuf ») (Fougeron *et al.*, 2007), nous pensons que notre système de détection de schwas pourrait confondre le /ə/ avec l'un d'eux et avons donc étendu les phonèmes considérés au /ə/, /ø/, et /œ/. La Table 2 contient le nombre ainsi que la proportion de /ə/, /ø/, /œ/, et « e » (la combinaison des trois phonèmes) pour chaque texte et pour l'ensemble des textes. Nous avons ensuite annoté manuellement la présence ou l'absence de ces phonèmes sur les enregistrements audio du sous-ensemble.

Phonème	**Prononciation Réalisations**	**Texte 1**	**Texte 2**	**Texte 3**	**Texte 4**	**Texte 5**	**Tous**
Tous	Texte lu	735	726	634	689	734	3 518
/ə/	Texte lu	30 (4,1%)	53 (7,3%)	42 (6,6%)	42 (6,1%)	35 (4,8%)	202 (5,7%)
	Annotées manuellement	28 (3,8%)	48 (6,6%)	38 (5,5%)	37 (5,4%)	28 (3,8%)	179 (5,1%)
/ø/	Texte lu	4 (0,5%)	7 (1,0%)	7 (1,1%)	7 (1,0%)	6 (0,8%)	31 (0,9%)
	Annotées manuellement	4 (0,5%)	6 (0,8%)	6 (0,9%)	6 (0,9%)	8 (1,1%)	30 (0,8%)
/œ/	Texte lu	4 (0,5%)	3 (0,4%)	3 (0,5%)	2 (0,3%)	0 (0,0%)	12 (0,3%)
	Annotées manuellement	3 (0,4%)	3 (0,4%)	3 (0,5%)	2 (0,3%)	0 (0,0%)	11 (0,3%)
« e »	Texte lu	38 (5,2%)	63 (8,7%)	52 (8,2%)	51 (7,4%)	41 (5,6%)	245 (7,0%)
	Annotées manuellement	36 (4,9%)	58 (8,0%)	48 (7,6%)	46 (6,7%)	36 (4,9%)	224 (6,4%)

TABLE 2 – Nombre et proportion de tous les phonèmes (33), /ə/, /ø/, /œ/ et « e » pour chaque texte et pour l'ensemble des textes. Nombre moyen et ratio moyen pour l'annotation manuelle.

3 Systèmes de détection de schwa

3.1 Système de détection de schwas de référence

Notre système de détection de schwas de référence est un système de Reconnaissance Automatique de la Parole (RAP) basé uniquement sur un modèle TDNN-HMM entraîné avec la fonction LF-MMI (Boyer & Rouas, 2019). Le réseau neuronal est un réseau à délai temporel échantillonné avec 7 couches TDNN, chacune ayant 1024 unités. La valeur de pas temporel est réglée sur 1 pour les trois premières couches, 0 pour la quatrième, et 3 pour les suivantes. Le modèle acoustique est basé sur un vecteur MFCC de haute résolution à 40 dimensions concaténé avec un i-vecteur de 100 dimensions (Gupta *et al.*, 2014). Il a été entraîné en utilisant la boîte à outils Kaldi (Povey *et al.*, 2011) sur un sous-ensemble d'ESTER 1 et 2 (Galliano *et al.*, 2009). Nous utilisons directement la sortie phonétique du système pour les analyses. Le nombre de /ə/, /ø/ et /œ/ est ensuite extrait de cette transcription. La Figure 1 schématise le système de référence. Ce système atteint un taux d'erreur de phonèmes de 19,5% sur le corpus Rhapsodie composé de parole préparée, semi-spontanée et spontanée (Martin *et al.*, 2024).

3.2 Système de détection de schwas proposé

L'idée de notre système proposé est d'améliorer les performances de la détection de schwa en supprimant les différences individuelles en utilisant la prononciation standard des mots. Pour ce faire, nous utilisons le système de RAP avec un lexique contenant des variantes de prononciation du dictionnaire phonémique fourni par le Laboratoire d'Informatique de l'Université du Mans (LIUM). De plus, un modèle de langage en mots 3-gram prenant en compte le contexte est implémenté.

Ce dernier a été entraîné sur les corpus ESTER en utilisant la méthode de comptage n-gram de SRILM (Stolcke, 2002) avec une réduction KN et a été limité aux 50 000 mots les plus fréquents dans les textes d'entraînement et le dictionnaire. Ce système atteint un taux d'erreur de mots de 13,7% (Boyer & Rouas, 2019).

Nous utilisons ainsi la sortie en mots de ce système complet de RAP que nous transcrivons ensuite en unités phonétiques en utilisant le Lexique 3.83 (voir la Figure 2). Si un mot n'est pas inclus dans celui-ci, nous l'avons transcrit manuellement. Le nombre d'occurrences détecté de /ə/, /ø/, et /œ/ est ensuite extrait de cette transcription. Cette méthode permet d'associer l'audio à la prononciation standard, ce qui réduirait la différence dans les transcriptions entre le lexique du système complet RAP et le lexique 3.83. Par exemple, le lexique à l'intérieur du système transcrivait le mot « premier » comme /pʁəmje/ alors que la prononciation standard dans le Lexique 3.83 est /pʁømje/.

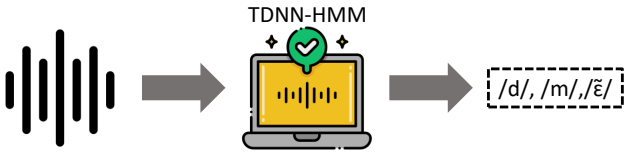


FIGURE 1 – Système de détection de schwas de référence

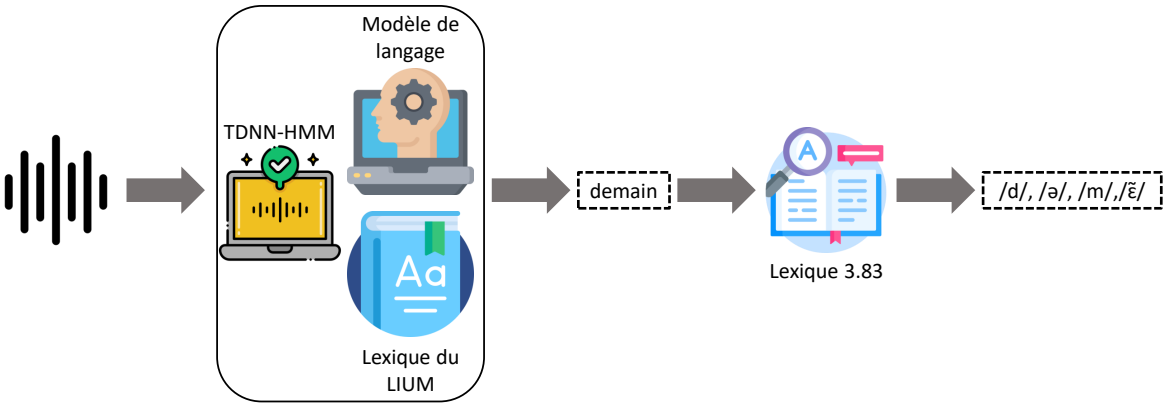


FIGURE 2 – Système de détection de schwas proposé

4 Performances de reconnaissance des phonèmes étudiés

Afin d'évaluer les performances des deux systèmes, nous avons calculé l'Erreur Absolue Moyenne normalisée (%MAE) et la Racine de l'Erreur Quadratique Moyenne (%RMSE) entre le nombre de schwas annotés manuellement et le nombre détecté par chaque système. Plus ces valeurs sont basses, meilleur est le système. La Table 3 référence le %MAE et le %RMSE pour le système initial tandis que la Table 4 référence les mêmes métriques pour le système proposé.

Dans la Table 3, le pourcentage d'erreur de détection du schwa est élevé, indiquant que ce système est peu performant par rapport à l'annotation manuelle, en détectant plus de schwas que leur nombre annoté. De plus, ces performances dépendent du texte (le système initial est plus performant sur le texte

#2 que sur le texte #5). Les performances pour /ø/ et /œ/ sont également faibles, particulièrement pour /ø/, et dépendent elles aussi du texte. La catégorie « e » montre également de faibles performances, avec les mêmes disparités entre les textes.

Phonème	Métrique	Texte 1	Texte 2	Texte 3	Texte 4	Texte 5	Tous
/ə/	%MAE	29,3	11,2	28,4	33,5	50,2	28,3
	%RMSE	35,7	13,5	30,5	36,2	52,0	32,4
/ø/	%MAE	50,0	14,5	44,8	42,6	52,5	40,0
	%RMSE	50,0	14,5	46,3	44,1	42,5	44,6
/œ/	%MAE	2,6	33,3	3,2	20,0	-	12,5
	%RMSE	5,1	36,7	9,7	30,0	-	25,0
e	%MAE	17,5	9,6	16,5	21,5	28,0	17,8
	%RMSE	23,6	11,4	18,6	23,6	30,7	20,9

TABLE 3 – %MAE et %RMSE pour /ə/, /ø/, /œ/ et « e » pour chaque texte et pour l'ensemble avec le système de détection de schwas de référence.

Phonème	Métrique	Texte 1	Texte 2	Texte 3	Texte 4	Texte 5	Tous
/ə/	%MAE	8,9	6,4	11,5	10,2	25,3	11,4
	%RMSE	12,9	9,5	13,4	13,7	30,2	15,8
/ø/	%MAE	0,4	17,1	16,7	7,4	11,3	10,8
	%RMSE	0,1	17,1	21,2	10,3	15,0	16,9
/œ/	%MAE	2,6	6,7	6,5	25,0	-	8,3
	%RMSE	7,7	16,7	12,9	35,0	-	20,8
e	%MAE	7,5	5,9	7,4	7,8	18,8	8,9
	%RMSE	10,6	8,8	9,3	11,7	22,4	12,4

TABLE 4 – %MAE et %RMSE pour /ə/, /ø/, /œ/ et « e » pour chaque texte et pour l'ensemble avec le système de détection de schwas proposé.

Les mêmes métriques d'évaluation calculées sur la sortie du système de RAP proposé sont rapportées dans la Table 4. Les %MAE et %RMSE des trois phonèmes et de leur combinaison ont grandement diminué : la pire %RMSE pour tous les textes avec le système initial était de 44,6% (/ø/) tandis qu'avec le système proposé, il est de 20,8% (/œ/). De plus, même s'il existe encore des disparités entre les textes, elles sont moindres que celles du système initial. Le système proposé est donc plus proche de l'annotation manuelle, donc plus performant, pour la détection de /ə/, /ø/, /œ/, et de leur combinaison. Nous devons maintenant évaluer sa robustesse à la somnolence avant de le considérer comme un système fiable.

4.1 Robustesse du système proposé à la somnolence

Pour mesurer l'effet des textes, de la KSS, et de la latence d'endormissement sur les performances de détection de /ə/, /ø/, et /œ/ par le système proposé, nous avons calculé quatre ANOVA multivariées à mesures répétées pour expliquer les variations intra- et inter-locuteurs des performances de détection des phonèmes avec celles de la somnolence et l'influence des textes. Les résultats de ces ANOVA sont reportés dans la Table 5.

Les différents facteurs n'ont pas d'effet significatif sur les variations inter-locuteurs des phonèmes étudiés. Seuls les textes influencent significativement les variations intra-locuteur des performances de détection de /ə/ et /ø/. Puisque la détection de chaque phonème n'est pas affectée par la somnolence,

le système de détection de schwas proposé est robuste à la somnolence et fiable pour la détection de ces phonèmes.

Nous allons maintenant mesurer si le nombre de /ə/, /ø/, et /œ/ détecté automatiquement est corrélé au niveau de somnolence des patients sur l'ensemble du corpus TILE en utilisant le système proposé.

Facteur	Perf. sur /ə/	Perf. sur /ø/	Perf. sur /œ/	Perf. sur « e »
Texte	***	***	-	***
KSS	-	-	-	-
Latence d'endormissement	-	-	-	-

TABLE 5 – Résultats des ANOVA multivariées à mesures répétées avec le système de détection de schwas proposé sur le sous-ensemble. Nous n'avons trouvé aucune influence significative du Texte, de la somnolence subjective ou objective sur les variations inter-locuteurs ; seule l'influence sur les variations intra-locuteurs est rapportée. *** : $p < .001$

5 Lien entre somnolence et détection automatique de phonèmes

Chaque /ə/, /ø/, et /œ/ extrait des textes transcrits avec le Lexique 3.83 a été comptabilisé afin d'obtenir le nombre attendu pour chacun d'entre eux. Nous avons appliqué le système proposé au corpus TILE et calculé la moyenne et l'écart-type du nombre de phonèmes détectés sur les enregistrements. La moyenne et l'écart-type de la valeur absolue de la différence entre le nombre de phonèmes attendus et détectés ont également été calculés. La Table 6 référence ces mesures.

Texte	Mesures	#/ə/	#/ø/	#/œ/	#« e »
Texte 1	Attendu	30	4	4	38
	Détecté	31 (1)	4 (1)	4 (0)	38 (2)
Texte 2	Attendu	53	7	3	63
	Détecté	50 (2)	8 (1)	3 (1)	61 (3)
Texte 3	Attendu	42	7	3	52
	Détecté	42 (3)	6 (1)	3 (1)	50 (3)
Texte 4	Attendu	42	7	2	51
	Détecté	41 (3)	7 (1)	2 (1)	50 (3)
Texte 5	Attendu	35	6	0	41
	Détecté	34 (2)	7 (1)	0 (0)	42 (2)
Tous	Attendu	202	31	12	245
	Détecté	197 (6)	32 (2)	12 (1)	242 (7)

TABLE 6 – Nombre de /ə/, /ø/, /œ/ et « e » attendus. Moyenne et écart-type (valeurs arrondies) des phonèmes détectés par le système de détection de schwas proposé.

Le nombre détecté de /œ/ est le plus proche du nombre attendu tandis que la détection de /ə/ est la plus éloignée. Cela peut s'expliquer par son nombre attendu élevé par rapport à /ø/ et /œ/ (202, 31, et 12 respectivement pour tous les textes). En moyenne, le nombre de phonèmes détectés par ce système est égal ou proche du nombre attendu, ce qui signifie que le système proposé est fiable.

Pour évaluer l'impact de la somnolence sur le nombre de phonèmes détectés par le système proposé, nous avons réalisé quatre ANOVA multivariées à mesures répétées avec les mêmes facteurs qu'au-

paravant (textes, KSS, et latence d'endormissement). Les résultats sont référencés dans la Table 7. Aucun effet significatif des variations inter-locuteurs n'a été trouvé contrairement aux variations intra-locuteurs. Un effet significatif de la somnolence subjective à court terme (KSS) sur les nombres détectés automatiquement de /ø/ et « e » a été mesuré tandis que la latence d'endormissement n'affecte la détection automatique d'aucun phonème. Le perception subjective de la somnolence à court terme impacterait donc la détection automatique du nombre de /ø/ et « e » contrairement à la somnolence physiologique mesurée par EEG.

Facteur	Nombre de /ə/	Nombre de /ø/	Nombre de /œ/	Nombre de « e »
KSS	.	**	-	*
Latence d'endormissement	-	-	-	.

TABLE 7 – Résultats de l'ANOVA multivariée à mesures répétées (variations intra-locuteur uniquement) avec le système de détection de schwas proposé sur le corpus TILE. . : $p < .1$; * : $p < .05$; ** : $p < .01$

6 Conclusion

Dans le but d'aider les cliniciens à suivre l'évolution des symptômes de la Somnolence Diurne Excessive (SDE) en analysant la parole spontanée, nous avons cherché à compléter les approches précédentes centrées sur la durée et l'emplacement des pauses en étudiant le comportement phonétique. En particulier, nous nous sommes intéressés au schwa, phonème optionnel en français, dont sa réalisation peut refléter un effort vocal de la part du locuteur. Nous émettons l'hypothèse que la somnolence altère cet effort. Du fait de l'absence de corpus de parole spontanée spécifique à la somnolence, nous avons utilisé le corpus TILE contenant des enregistrements de parole lue car l'analyse phonétique est transposable à la parole spontanée.

Pour ce faire, nous avons validé un système de détection de schwas fiable et robuste à la somnolence sur l'annotation manuelle d'un sous-ensemble du corpus TILE et l'avons étendu à d'autres phonèmes liés au schwa. Nous l'avons ensuite appliqué à l'ensemble du corpus TILE et avons trouvé une relation significative entre les phonèmes considérés détectés automatiquement par notre système et la somnolence subjective à court terme.

Notre prochaine étape consiste à concevoir un système de classification automatique en prenant en compte les résultats actuels afin d'améliorer la détection de la somnolence. De plus, nous visons à étudier le lien entre la durée et les propriétés acoustiques des trois phonèmes et la somnolence. La méthode utilisée pour le système de détection de schwas pouvant être appliquée à d'autres phonèmes, nous projetons également d'élargir les phonèmes considérés ainsi que les caractéristiques étudiées (le débit de parole, etc.).

Remerciements

CB a reçu le soutien financier de la MITI du CNRS (projet PRIME 80 DSM-HEALTH). VPM a reçu le soutien financier du programme de recherche et d'innovation européen Horizon Europe à travers le projet Marie Skłodowska-Curie MATER (No. 101106577).

Références

ARAND D., BONNET M., HURWITZ T., MITLER M., ROSA R. & SANGAL R. B. (2005). The Clinical Use of the MSLT and MWT. *Sleep*, **28**(1), 123–144. DOI : 10.1093/sleep/28.1.123.

BARNES C. M. & WATSON N. F. (2019). Why healthy sleep is good for business. *Sleep Med. Rev.*, **47**, 112–118. DOI : 10.1016/j.smrv.2019.07.005.

BEAUMARD C., MARTIN V., WU Y., ROUAS J.-L. & PHILIP P. (2023). Automatic detection of schwa in French hypersomniac patients. *Plate-Forme Intelligence Artificielle (PFIA)*.

BIOULAC S., MICOULAUD-FRANCHI J.-A., ARNAUD M., SAGASPE P., MOORE N., SALVO F. & PHILIP P. (2017). Risk of Motor Vehicle Accidents Related to Sleepiness at the Wheel : A Systematic Review and Meta-Analysis. *Sleep*, **40**(10). DOI : 10.1093/sleep/zsx134.

BOYER F. & ROUAS J.-L. (2019). End-to-End Speech Recognition : A review for the French Language. *arXiv*. DOI : 10.48550/ARXIV.1910.08502.

BÜRKI A., ERNESTUS M., GENDROT C., FOUGERON C. & FRAUENFELDER U. H. (2011). What affects the presence versus absence of schwa and its duration : A corpus analysis of French connected speech. *J. Acoust. Soc. Am.*, **130**(6), 3980–3991. DOI : 10.1121/1.3658386.

DURAND J. (2014). À la recherche du schwa : données, méthodes et théories. *SHS Web of Conferences*, **8**, 23–43. DOI : 10.1051/shsconf/20140801396.

FOUGERON C., GENDROT C. & BÜRKI A. (2007). On the acoustic characteristics of French schwa. In *ICPhS 2007*, Saarbrücken.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech 2009*, p. 2583–2586. DOI : 10.21437/Interspeech.2009-680.

GUPTA V., KENNY P., OUELLET P. & STAFYLAKIS T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP*, p. 6334–6338. DOI : 10.1109/ICASSP.2014.6854823.

HARRISON Y. & HORNE J. A. (1997). Sleep Deprivation Affects Speech. *Sleep*, **20**(10), 871–877. DOI : 10.1093/sleep/20.10.871.

JIKE M., ITANI O., WATANABE N., BUYSSE D. J. & KANEITA Y. (2018). Long sleep duration and health outcomes : A systematic review, meta-analysis and meta-regression. *Sleep Med. Rev.*, **39**, 25–36. DOI : 10.1016/j.smrv.2017.06.011.

MACLEAN A. W., FEKKEN G. C., SASKIN P. & KNOWLES J. B. (1992). Psychometric evaluation of the Stanford Sleepiness Scale. *J. Sleep Res.*, **1**(1), 35–39. DOI : 10.1111/j.1365-2869.1992.tb00006.x.

MARTIN V., ARNAUD B., ROUAS J.-L. & PHILIP P. (2022). Does sleepiness influence reading pauses in hypersomniac patients ? In *Speech Prosody 2022*, p. 62–66. DOI : 10.21437/SpeechProsody.2022-13.

MARTIN V., FERRON A., ROUAS J.-L., SHOCHI T., DUPUY L. & PHILIP P. (2023a). Physiological vs. Subjective sleepiness : what can human hearing estimate better ? Insights from the French Endymion study. In *ICPhS 2023*.

MARTIN V., LOPEZ R., DAUVILLIERS Y., ROUAS J.-L., PHILIP P. & MICOULAUD-FRANCHI J.-A. (2023b). Sleepiness in adults : An umbrella review of a complex construct. *Sleep Med. Rev.*, **67**, 101718. DOI : 10.1016/j.smrv.2022.101718.

MARTIN V., ROUAS J.-L., BOYER F. & PHILIP P. (2021a). Automatic Speech Recognition Systems Errors for Objective Sleepiness Detection Through Voice. In *Interspeech 2021*, p. 2476–2480. DOI : 10.21437/Interspeech.2021-291.

MARTIN V., ROUAS J.-L., MICOULAUD-FRANCHI J.-A., PHILIP P. & KRAJEWSKI J. (2021b). How to Design a Relevant Corpus for Sleepiness Detection Through Voice ? *Front. digit. health.*, **3**, 686068. DOI : 10.3389/fdgth.2021.686068.

MARTIN V. P., BEAUMARD C., ROUAS J.-L. & WU Y. (2024). Is automatic phoneme recognition suitable for speech analysis ? Temporal and performance evaluation of an Automatic Speech Recognition model in spontaneous French. In *Speech Prosody 2024*.

NEW B., PALLIER C., BRYSBAERT M. & FERRAND L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 516–524. DOI : 10.3758/BF03195598.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLÍČEK P., QIAN Y., SCHWARZ P., SILOVSKÝ J., STEMMER G. & VESELÝ K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii : IEEE Signal Processing Society.

SCHULLER B., STEIDL S., BATLINER A., SCHIEL F. & KRAJEWSKI J. (2011). The INTERSPEECH 2011 speaker state challenge. In *Interspeech 2011*, p. 3201–3204. DOI : 10.21437/Interspeech.2011-801.

SCHULLER B. W., BATLINER A., BERGLER C., POKORNY F. B., KRAJEWSKI J., CYCHOSZ M., VOLLMANN R., ROELEN S.-D., SCHNIEDER S., BERGELSON E., CRISTIA A., SEIDL A., WARLAUMONT A. S., YANKOWITZ L., NÖTH E., AMIRIPARIAN S., HANTKE S. & SCHMITT M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech 2019*, p. 2378–2382. DOI : 10.21437/Interspeech.2019-1122.

STOLCKE A. (2002). SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, p. 901–904. DOI : 10.21437/ICSLP.2002-303.

TRAN B., ZHU Y., LIANG X., SCHWOEBEL J. W. & WARRENBURG L. A. (2022). Speech Tasks Relevant to Sleepiness Determined With Deep Transfer Learning. In *ICASSP*, p. 6937–6941. DOI : 10.1109/ICASSP43922.2022.9747000.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Effet du vieillissement sur l'anticipation d'arrondissement intra-syllabique en français.

Louise Wohmann-Bruzzo[1] Cécile Fougeron[1] Nicolas Audibert[1]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4, rue des Irlandais, 75005 Paris, France
prénom.nom@sorbonne-nouvelle.fr

RÉSUMÉ

Cette étude examine l'évolution de l'anticipation d'arrondissement intra-syllabique en fonction de l'âge, en se basant sur les travaux précédents de D'Alessandro et Fougeron (2021), qui montre une diminution de la coarticulation inter-syllabique chez les personnes âgées. Nous avons analysé acoustiquement les syllabes /sy/ et /si/ de 40 locuteurs français répartis en deux groupes : 20 jeunes adultes (23-34 ans) et 20 personnes âgées (72-86 ans). Nos résultats montrent une diminution significative de l'anticipation d'arrondissement chez les âgés, indépendante d'une diminution de débit articulatoire. Moins de coarticulation au sein d'une même syllabe ne peut pas s'expliquer par un encodage syllabe par syllabe, comme pouvait l'être la diminution de coarticulation entre syllabe. Nous avançons donc que cette diminution globale de la co-articulation reflète un changement de paramétrisation de la parole chez les personnes âgées limitant le chevauchement entre gestes articulatoires et garantissant l'atteinte des cibles articulatoires successives.

ABSTRACT

Ageing effects on intra-syllabic anticipatory rounding in French.

This study examines the evolution of intra-syllabic rounding anticipation as a function of age, building upon previous work by D'Alessandro and Fougeron (2021), which demonstrated a decrease in inter-syllabic coarticulation among older individuals. Acoustic analysis was conducted on the syllables /sy/ and /si/ from 40 French speakers divided into two groups: 20 young adults (aged 23-34) and 20 elderly individuals (aged 72-86). Our findings reveal a significant decrease in rounding anticipation among the elderly, independent of a decrease in articulatory rate. The reduced coarticulation within a syllable cannot be explained by a syllable-by-syllable encoding, as was the case with the decrease in inter-syllabic coarticulation. We therefore propose that this overall reduction in coarticulation reflects a shift in speech parameterization among the elderly, limiting overlap between articulatory gestures and ensuring the achievement of successive articulatory targets.

MOTS-CLES : coarticulation, vieillissement, anticipation d'arrondissement, acoustique, fricative

KEYWORDS : coarticulation, speech aging, rounding anticipation, acoustic analysis, fricative

1 Introduction

1.1 Vieillissement de la parole

La phonétique, particulièrement dans son application clinique, explore de plus en plus les liens entre langage, parole et maladies neurodégénératives, lesquelles affectent principalement les personnes âgées. En comparaison, moins d'études se penchent sur le lien entre parole et vieillissement sain, privant ainsi la littérature de bases de données saines à comparer aux données pathologiques. Or, il est difficile de décrire les effets du vieillissement pathologique sur la parole sans savoir comment la parole est supposée vieillir sans pathologie. Notre étude préliminaire s'inscrit dans un projet plus large, qui vise à mieux comprendre les effets du vieillissement sur la parole. Nous savons que les muscles s'affaiblissent et que les cartilages se solidifient avec l'âge, et des changements de voix dus à l'âge ont été expliqués par ces phénomènes physiologiques (Xue & Hao, 2003; Linville & Rens, 2001). Nos capacités cognitives (Rodríguez-Aranda & Jakobsen, 2011) et nos capacités motrices déclinent également : nos mouvements sont moins précis, et plus lents. Les personnes âgées prennent alors plus de temps pour planifier leurs gestes, notamment leurs gestes articulatoires. Le ralentissement du débit de parole est un des résultats les plus robustes des études faites sur l'âge : les personnes âgées parlent plus lentement que les personnes jeunes. Cela a été constaté sur plusieurs tâches de production de parole (Amerman & Parnell, 1992; Ramig, 1983; voir Tucker, 2021, pour une revue), avec plusieurs groupes d'âge, et sur des études transversales (Bóna, 2014) comme longitudinales (Quené, 2013). Mais la nature de ce ralentissement de débit articulatoire peut être multifactorielle. Le ralentissement moteur général lié à l'âge est un des facteurs connus (Yan, Thomas & Stelmach, 1998), même si ses causes sont débattues, et on sait qu'il pourrait affecter les gestes articulatoires. Un ralentissement des réponses cognitives pourrait s'ajouter et pousser les personnes âgées à adopter de nouvelles stratégies, dont le ralentissement des gestes articulatoires ferait partie, afin de répondre à un changement dans leurs ressources. D'autres stratégies, discutées dans la littérature, seraient l'utilisation de pauses ou de fillers plus importante chez les personnes âgées, qui, selon Bortfeld et. al. (2001) et Martin-Reis & Andrade (2011), répondraient à une difficulté accrue à planifier les unités de parole. Tous ces changements pourraient affecter la manière dont les personnes âgées parlent, mais on connaît encore mal leur rôle et leur conséquence. Pour essayer de démêler les effets de certains de ces facteurs, particulièrement le ralentissement du débit et les éventuels changements dans la planification de la parole, nous avons choisi d'utiliser la coarticulation anticipatoire comme indice des effets du vieillissement sur la parole.

1.2 Arrondissement anticipatoire

La coarticulation est le phénomène par lequel les gestes articulatoires peuvent s'exécuter ensemble. Ainsi, les sons vont s'influencer dans le flux de parole. Elle est nécessaire à la fluidité du signal produit. Nous avons choisi d'étudier l'arrondissement anticipatoire, en français, d'une voyelle arrondie sur la fricative qui la précède. Dans une étude récente, D'Alessandro et Fougeron (2021) ont montré que la coarticulation anticipatoire de voyelle à voyelle, entre deux syllabes en français, diminue avec l'âge, et ceci de manière non linéaire entre 20 et 90 ans. Afin de tester si la diminution de la coarticulation était la conséquence directe d'un ralentissement du débit – car un débit plus lent impliquerait moins de chevauchements de gestes et inversement, le débit serait plus lent car les gestes se chevauchent moins –, les autrices ont étudié les liens entre coarticulation et débit des locuteurs. Leurs résultats montrent une réduction abrupte du degré de coarticulation chez les locuteurs dépassant 54 ans. Entre 20 ans et 70 ans, la réduction de la coarticulation et le

ralentissement du débit sont bien corrélés. Mais, à partir de 70 ans, le ralentissement du débit n'explique plus la diminution de la coarticulation : les locuteurs âgés coarticulent peu, quel que soit leur débit. Les autrices proposent l'hypothèse suivante pour expliquer leurs résultats. Cette diminution pourrait être due à la modification de la taille de l'unité sur laquelle la parole est planifiée. En effet, les voyelles impliquées dans ce processus de coarticulation faisant partie d'un seul et même mot bisyllabique, la présence d'anticipation d'une syllabe sur la précédente suppose une planification d'une unité de parole dans laquelle les gestes peuvent se chevaucher qui est à l'échelle du mot, a minima. Si la coarticulation entre les syllabes, dans ce mot, disparaît, cela peut venir du fait que les gestes sont planifiés non plus à l'échelle du mot où les syllabes sont coordonnées, mais sur des unités plus petites, syllabe par syllabe. Or, pour appuyer cette hypothèse, il est nécessaire de savoir si la coarticulation à l'intérieur d'une syllabe varie aussi avec l'âge. Ceci est l'objectif du projet dans lequel s'inscrit le présent travail. Nous cherchons à vérifier si une coarticulation anticipatoire à l'échelle syllabique serait affectée par l'âge. Pour cela, nous examinons l'anticipation de l'arrondissement de la voyelle /y/ sur la fricative sourde /s/ dans des syllabes /sy/ en parole continue en français. Dans ce travail préliminaire, les effets du vieillissement sont testés par une comparaison de groupes, entre adultes âgés et jeunes.

1.3 Quantification de l'arrondissement dans les réalisations de /s/

Le choix de la syllabe /sy/ pour considérer l'anticipation d'arrondissement dans notre étude est motivé par les propriétés mêmes des sons impliqués : la voyelle /y/ et la fricative sourde /s/ sont compatibles articulatoirement. Peu de gestes articulatoires diffèrent entre leurs productions, et cela favorise l'apparition de coarticulation (Yeni-Komshian & Soli, 1981). Le geste de la labialisation, qui se traduit en français par une protrusion des lèvres et une réduction de leur aperture, peut être anticipé dès la production du /s/, et il a un effet directement observable acoustiquement : l'allongement de la cavité antérieure a un impact sur le bruit du /s/, et donc sur les mesures acoustiques qui le quantifient. La mesure habituelle dans la littérature pour les fricatives, et particulièrement les sibilantes, est le premier moment spectral : le centre de gravité (CoG). Mais le CoG ne permet pas toujours une interprétation claire des événements articulatoires. Il a été vérifié que le CoG baisse en contexte arrondi (Koenig et. al., 2013; Jongman et. al., 2000) ; mais cet abaissement peut traduire à la fois l'augmentation de la taille de la cavité antérieure à la constriction (effet sur le filtre), ou le placement de l'apex lingual plus en arrière (effet sur le chenal fricatif). Une mesure permettant une interprétation plus directe et plus transparente est introduite par Koenig et. al. (2013), qui cherchent à capturer la résonance de la cavité antérieure à la constriction : la mesure $Freq_M$ est une mesure du pic spectral du /s/, qui se trouve généralement dans les fréquences moyennes (entre 3 et 7 kHz pour les hommes et entre 3 et 8 kHz pour les femmes). Cette mesure capture la résonance de la cavité antérieure à la constriction. Les études de Koenig et al. (2013) et de Shadle et al. (2023) montrent que le $Freq_M$ des sibilantes labialisées diminue, permettant de distinguer celles avec et sans anticipation d'arrondissement. Son abaissement, lié à l'allongement de la cavité antérieure à la constriction, se fait dynamiquement : on peut l'observer tout au long de la fricative, avec un abaissement maximal à la fin. C'est pourquoi nous avons choisi d'utiliser $Freq_M$ dans notre étude. Cette mesure étant dépendante du sexe des locuteurs, il est important de traiter séparément les données des hommes et des femmes, comme il est recommandé pour la plupart des études sur les sibilantes (Shadle et. al., 2023 ; Stuart-Smith, 2007, Stuart-Smith et. al., 2007). La prise en compte de l'évolution temporelle des caractéristiques acoustiques du /s/ est conseillé pour ce type d'étude sur les fricatives : l'analyse sur une seule mesure prise en moyenne sur la totalité de la fricative, ou à un seul point temporel, amènerait un degré d'erreur trop important, et ne tiendrait pas compte des variations aérodynamiques naturelles de la production du /s/ (Bendat & Piersol,

2000). De plus, la coarticulation est un phénomène dynamique, et il semble difficile de l'analyser de manière statique sans en lisser les principales caractéristiques.

2 Méthode

2.1 Corpus

L'étude analyse un corpus de 40 locuteurs, répartis en deux groupes : les personnes âgées, comprenant 11 femmes et 9 hommes, et les jeunes, comprenant 10 hommes et 10 femmes. Les personnes âgées ont entre 72 et 86 ans (moyenne = 78.2), et les jeunes ont entre 23 et 34 ans (moyenne = 26.7). Chaque locuteur a lu trois fois 7 phrases formant une histoire. Dans ces phrases, nous avons sélectionné 8 mots monosyllabiques, ayant pour attaque un /s/ et comme noyau la voyelle /y/ ou la voyelle /i/. Nous avons 4 /sy/ et 4 /si/ pour une répétition des 7 phrases. L'analyse des /si/ nous permet d'établir une base de référence pour les /s/ de chaque locuteur, afin d'évaluer clairement l'effet de la voyelle /y/ sur le /s/ qui la précède. Nous ferons référence à un /s/ suivi d'un /i/ en le notant /s_i/, et à un /s/ suivi d'un /y/ en le notant /s_y/. Nous obtenons donc 12 /s_i/ et 12 /s_y/ par locuteur, soit 960 /s/ analysés au total.

2.2 Analyses acoustiques et statistiques

Les [s] ont été segmentés manuellement avec Praat. Suivant la méthode de Koenig et. al. (2013), un script Praat a été utilisé pour relever la durée de chaque [s], et sur la bande de fréquence 500 Hz-15 kHz, 12 spectres ont été calculés à 12 points temporels équidistants, avec des fenêtres de Hanning de 25 ms. Sur chaque spectre, nous avons extrait automatiquement la valeur de $Freq_M$, après avoir procédé à un lissage cepstral avec une largeur de bande de 1000 Hz, suivant les paramètres retenus par Al-Tamimi & Khattab (2015). Le premier et le dernier point temporel ont été exclus, nous amenant à 10 points de mesures analysés. La mesure de $Freq_M$ étant la mesure de la fréquence du pic spectral de la résonance principale, dépendante de la taille de la cavité antérieure à la constriction, nous attendons que les $Freq_M$ des [s_i] soient plus hauts que les $Freq_M$ des [s_y]. Ainsi, notre mesure de la quantité d'anticipation d'arrondissement sera la différence de $Freq_M$ entre ces deux contextes, soit l'effet de [y] sur le [s]. Pour comprendre la relation entre débit et degré de coarticulation, nous avons mesuré un débit articulatoire par locuteur sur une des 7 phrases produites. Celle-ci a été segmentée manuellement en pauses et unités inter-pausales (UIP), et le débit articulatoire (ph/sec) correspond aux 29 phonèmes attendus divisés par la somme de la durée des UIP. Une modélisation GAMM a été envisagée sur les mesures dynamiques, mais la nécessité d'inclure une structure aléatoire contenant a minima l'occurrence et le locuteur ne permettait pas aux modèles de converger. Nous avons donc opté pour des modèles linéaires mixtes pour prédire les valeurs de $Freq_M$ en fonction de la voyelle suivante ([y] ou [i]), du groupe d'âge et de leur interaction en tant que facteurs fixes, le locuteur et l'occurrence étant des intercepts aléatoires. Afin de ne pas représenter une fricative par un seul point, pour chaque [s], nous avons 10 valeurs de $Freq_M$, prises aux 10 points temporels. La durée a également été incluse en tant que prédicteur continu dans le modèle, afin de tenir compte de la variation attendue de la durée segmentale en fonction de l'âge. L'écart moyen entre les valeurs de $Freq_M$ des [s_i] et de [s_y] par locuteur a été calculée avec un modèle linéaire. Nous avons cherché à confirmer l'existence d'une distinction entre les sexes et avons ainsi analysé les données des hommes et des femmes de manière distincte. Pour étudier la variabilité individuelle au sein de chaque groupe, des modèles par locuteur ont également été ajustés, avec la voyelle suivante comme facteur fixe. En revanche, ces modèles ne permettent pas de rendre compte directement du

lien temporel entre points successifs, et donc de la dynamique des trajectoires. En attendant de faire des modélisations dynamiques sur un nombre plus important de locuteurs, nous proposons ici, en complément, une analyse descriptive des données dynamiques, à partir de régressions *loess* des trajectoires de $Freq_M$ par locuteur et par syllabe.

3 Résultats

Puisque le sexe des locuteurs a un effet significatif sur les valeurs de $Freq_M$ ($\chi^2(1) = 14.9, p = 0.02$), avec des valeurs de $Freq_M$ des femmes plus aigües que celles des hommes, nous continuons nos analyses sur les groupes hommes et femmes, séparément. Comme attendu, en contexte [s_y], le bruit de la fricative est abaissé par rapport au contexte [s_i], chez les hommes ($\chi^2(1) = 22.24$, $p < .0001$), comme chez les femmes ($\chi^2(1) = 9.68$, $p < .001$). Plus intéressant, il y a une interaction entre l'âge et le contexte vocalique ($\chi^2(1) = 46.45$, $p < .0001$ pour les femmes et $\chi^2(1) = 14.44$, $p = .0001$ pour les hommes). En effet, comme on le voit sur la figure 1, les personnes âgées anticipent l'arrondissement, mais dans une moindre mesure par rapport aux jeunes : la baisse de $Freq_M$ entre les productions de [s_i] et de [s_y] est plus petite chez les personnes âgées. Chez les locutrices âgées, cette différence est même non significative ($z = 1.870$, $p = .06$). Cette diminution de l'anticipation d'arrondissement chez le groupe âgé est confirmée par l'effet significatif de l'âge sur l'écart moyen de $Freq_M$ entre [s_i] et [s_y] ($F(3, 36) = 2.21$, $p < 0.05$, Figure 1).

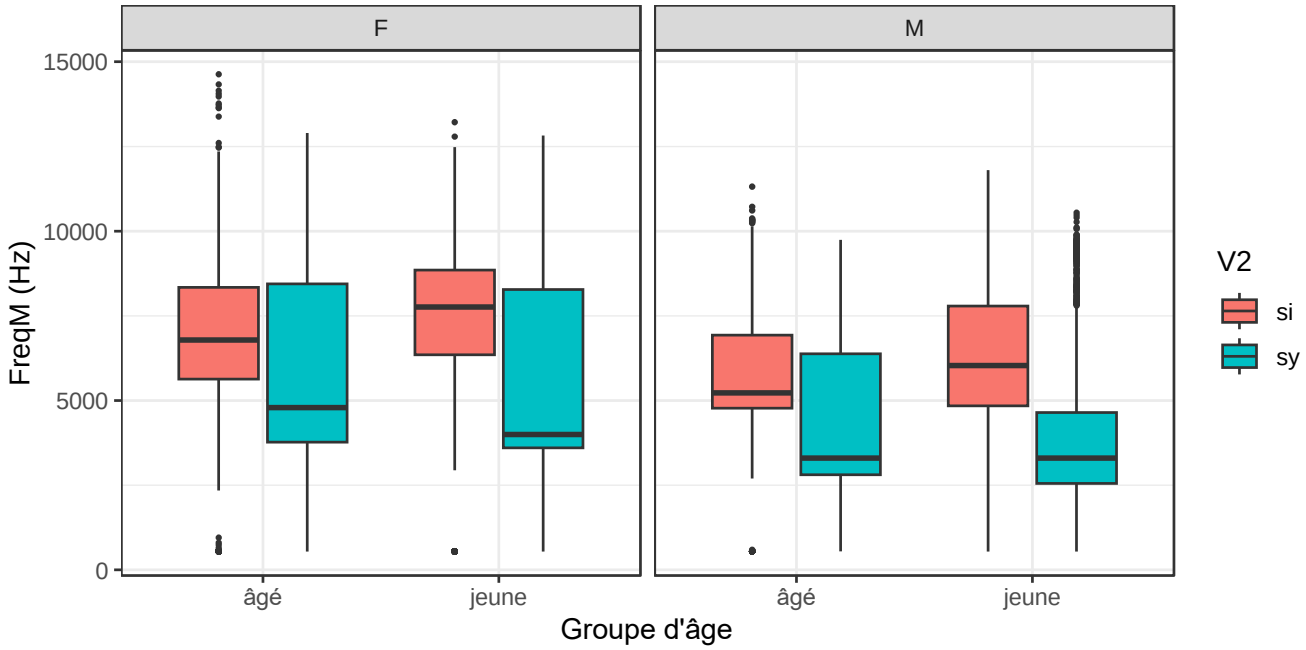


FIGURE 1 : Relation entre $Freq_M$ (axe y, en Hz), les femmes (à gauche) et les hommes (à droite), pour les groupes âgé (à gauche dans les encadrés) et jeune (à droite dans les encadrés), et en fonction de la syllabe ([s_i] en rouge, [s_y] en bleu).

Pour ce qui est de la durée, les personnes âgées produisent des /s/ significativement plus longs que le groupe jeune, chez les hommes comme chez les femmes. Cette différence n'est dépendante de la voyelle suivante que dans le groupe des hommes, pour qui les [s_i] sont significativement plus longs chez les hommes âgés, par rapport aux hommes jeunes ($z = 2.59$, $p < .01$). On trouve une corrélation positive modérée entre la quantité d'anticipation d'arrondissement et la durée des /s/ chez le groupe jeune ($r = .53$), mais pas chez le groupe âgé ($r = .21$). Les jeunes qui produisent des /s/ courts produisent plus de coarticulation, et inversement, mais cette relation n'est pas retrouvée chez les

locuteurs âgés. Les trajectoires moyennes des $Freq_M$ de chaque locuteur sont représentées sur la Figure 2.

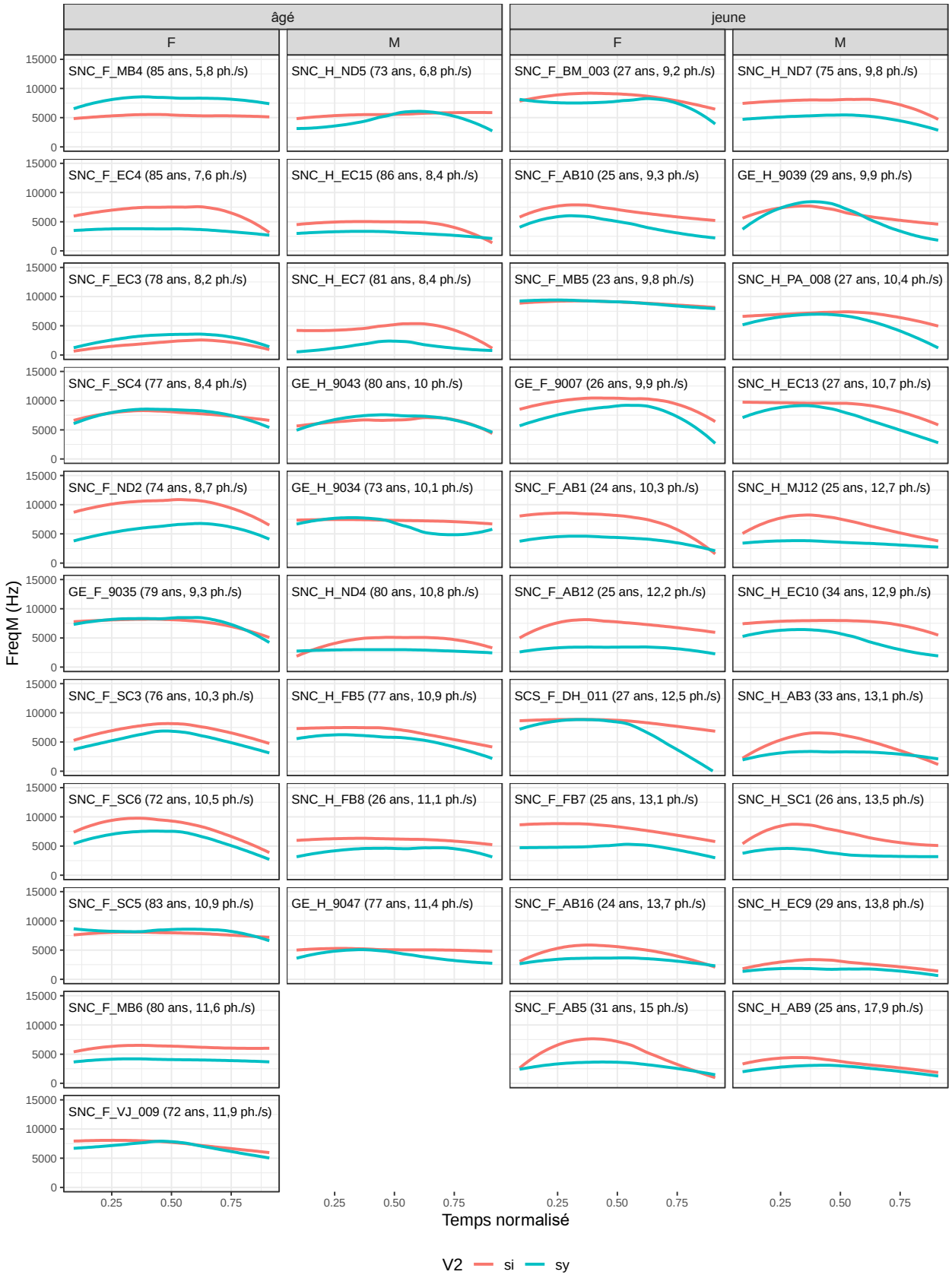


FIGURE 2 : Trajectoires moyennes, issues de modélisations *loess*, de la fréquence du pic spectral $Freq_M$ pour les locuteurs âgés (colonne de gauche) et jeunes (droite), avec les femmes à gauche et

les hommes à droite dans chaque colonne. L'âge et le débit en phonèmes/seconde est indiqué pour chaque locuteur.

Dans la figure 2, l'âge du locuteur et le débit articulatoire en phonèmes/seconde ont été ajoutés à la suite du code de chaque locuteur. Par ailleurs afin de faciliter l'interprétation du lien potentiel entre débit et anticipation d'arrondissement, dans chaque sous-groupe les locuteurs sont ordonnés en fonction de leur débit, de haut en bas par débit croissant.

Les modèles linéaires mixtes par locuteur ont montré que, sur les 21 femmes de notre corpus, 9 locutrices ne présentent pas de différences significatives entre le $Freq_M$ de leurs [s_y] et de leurs [s_i]. 6 de ces 9 locutrices sont âgées, ce qui confirme que la tendance à diminuer la quantité de coarticulation est une tendance majoritairement retrouvée chez les personnes âgées. Les 3 locutrices jeunes qui ne présentent pas de différence significative entre les $Freq_M$ de leurs [s_i] et de leurs [s_y] sont SNC_F_BM_003 (27 ans), SNC_F_AB16 (24 ans) et SNC_F_MB5 (23 ans). Chez les hommes, cette variabilité intra-groupe est moindre : seul GE_H_9043, qui est âgé, ne produit pas de différence significative entre ses [s_y] et ses [s_i].

L'observation qualitative de ces trajectoires nous permet d'apprécier la variabilité individuelle dans chacun des groupes : deux locuteurs présentant une différence significative entre leurs valeurs de $Freq_M$ pour [s_i] et [s_y] peuvent produire cette différence de multiples manières. Certains locuteurs vont produire l'anticipation d'arrondissement tôt dans la fricative, comme SNC_H_EC7 chez les hommes âgés, tandis que d'autres vont la produire plus tard, comme SCS_F_DH_011 chez les jeunes locutrices. Certains vont différencier leur [s_y] et leurs [s_i] par la hauteur de $Freq_M$, comme SNC_H_MJ12 chez les hommes jeunes, ou SNC_F_EC4 chez les femmes âgées.

4 Discussion & Conclusion

Nos résultats montrent que 6 femmes âgées de notre corpus ne produisent pas d'anticipation d'arrondissement dans leurs productions de [s_y]. 3 locutrices jeunes ne produisent pas non plus cette anticipation. De manière générale, on constate une grande variabilité inter-individuelle dans les effets de l'anticipation d'arrondissement sur les trajectoires de $Freq_M$ dans le groupe des femmes, que l'on retrouve dans une bien moindre mesure chez le groupe des hommes. Cette variabilité peut être due au /s/ lui-même, dont la production est dépendante de l'individu (Kavanagh, 2012), ainsi qu'à la variabilité individuelle que l'on retrouve dans la production de coarticulation (Guitard-Ivent et. al., 2023).

La diminution d'anticipation d'arrondissement chez le groupe âgé, par rapport au groupe jeune, est similaire à ce qui a été trouvé par D'Alessandro & Fougeron (2021) : la coarticulation diminue avec l'âge, aussi bien au sein d'une syllabe qu'entre deux syllabes. Cela n'est pas lié au débit articulatoire : la faible corrélation entre la durée des /s/ et la quantité d'anticipation d'arrondissement le montre. Aussi, l'observation des valeurs de débit articulatoire nous permet de constater que, dans les groupes âgés hommes et femmes, certains locuteurs ont des valeurs de débit articulatoire proches, mais des quantités d'anticipation d'arrondissement très différentes (SNC_ND2_F et SNC_SC4_F dans le groupe des femmes âgées par exemple). Les personnes âgées ne produisent donc pas moins d'anticipation d'arrondissement simplement parce qu'ils parlent moins rapidement que les jeunes. La diminution de la coarticulation intra-syllabique que nous trouvons permet de remettre en question une des hypothèses émises par D'Alessandro & Fougeron (2021) : si l'explication de la réduction de coarticulation anticipatoire entre syllabes relevait d'une organisation de la parole sur des unités plus petites, chez les locuteurs âgés par rapport aux plus jeunes, alors on s'attendrait à ne pas trouver

de diminution d'une coarticulation anticipatoire à l'échelle intra-syllabique. Or, nous montrons que cette diminution intra-syllabique est présente. Cela sous-entendrait que l'unité de planification est encore plus petite chez les personnes âgées : ces locuteurs et locutrices planifieraient leur parole segment par segment. Or, une telle interprétation est peu plausible pour des adultes âgés sans pathologie. Une autre explication serait qu'avec l'âge, le contrôle moteur de la parole évolue (D'Alessandro et. al., 2020). Si les locuteurs jeunes peuvent adopter un "mode" de parole où les gestes articulatoires se chevauchent fortement, et où les cibles articulatoires peuvent ne pas être atteinte (hypoarticulées), les locuteurs âgés adoptent un "mode" différent pour leur mouvements articulatoires (entre autres mouvements). La paramétrisation motrice de leurs gestes articulatoires se fait de sorte que les gestes articulatoires se chevauchent moins, et que les cibles successives soient atteintes (hyper- ou non-hypoarticulées). Des recherches supplémentaires sont nécessaires pour savoir quel paramétrage rend compte de ce mode de parole (augmentation de la rigidité (*stiffness*), décalage des phases, allongement des périodes d'activation, d'initiation/cessation (Hermes et. al., 2018), etc.). Il s'agira aussi de savoir si ce mode répond à une stratégie de compensation face à un déclin fonctionnel pour des gestes globalement plus précautionneux, s'il permet d'augmenter l'intelligibilité de la parole, ou s'il est seulement la conséquence d'un déclin au niveau du contrôle de cette activité motrice complexe qu'est la parole. Il est aussi important de prendre en compte les évolutions physiologiques qui viennent avec l'âge, spécifiquement la rigidification des tissus des lèvres et la détérioration des fibres musculaires, qui peuvent affecter le geste de protrusion des lèvres et ainsi jouer un rôle dans la réduction de coarticulation labiale que nous observons.

À la suite de cette étude, il sera nécessaire d'étendre nos analyses à une plus grande cohorte de locuteurs, de manière à pouvoir prendre en compte l'aspect dynamique dans nos modèles statistiques. Le but sera d'analyser les enregistrements d'une centaine de locuteurs, ayant entre 23 et 86 ans, afin de permettre une analyse de l'évolution des effets du vieillissement sur l'anticipation d'arrondissement, plutôt qu'une comparaison entre groupes d'âge, comme ici. En combinant notre méthodologie et un échantillon plus important, nous pensons pouvoir mieux cerner les effets de l'âge sur l'anticipation d'arrondissement, et ainsi proposer une explication à ces effets, qui nous éclairerait mieux sur le fonctionnement de la parole et son vieillissement.

Remerciements

Ce travail a été soutenu par le Laboratoire d'Excellence Empirical Foundations of Linguistics (LabEx EFL, ANR-10-LABX-0083). Il contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

AL-TAMIMI J. & KHATTAB G. (2015). Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. The Journal of the Acoustical Society of America, 138(1), 344-360. DOI : 10.1121/1.4922514.

AMERMAN, J. D. & PARNELL M. M. (1992). Speech timing strategies in elderly adults. *Journal of Phonetics*, 20(1), 65–76. DOI : 10.1016/s0095-4470(19)30254-2.

BENDAT, J. S., & PIERSOL, A. G. (2000). *Random data: Analysis and measurement procedures (3rd ed.)*. New York, NY: Wiley.

BÓNA, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America*, 136, EL116–EL121. DOI : 10.1121/1.4885482

BORTFELD, H., LEON, S. D., BLOOM, J. E., SCHOBER, M. F., and BRENNAN, S. E. (2001). Disfluency rates in conversation: effects of age, relationship, topic, role, and gender. *Lang. Speech* 44(2), 123–147. DOI : 10.1177/00238309010440020101

D'ALESSANDRO, D., & FOUGERON, C. (2021). Changes in Anticipatory VtoV Coarticulation in French during Adulthood. *Languages*, 6(181). DOI : **10.3390/languages6040181**

D'ALESSANDRO, D., BOURBON, A., & FOUGERON C. (2020). Effect of age on rate and coarticulation across different speechtasks. Paper present at the *12th International Seminar on Speech Production*, Virtual. December 14–18; New Haven: Haskins Press. ISBN : 978-1-7360794-2-3.

FUCHS, S. & TODA, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. FUCHS, M. ZYGIS & M. TODA (Eds.), *Turbulent sounds: An interdisciplinary guide.* Mouton de Gruyter. 281-302. DOI : 10.1515/9783110226584.281

GUITARD-IVENT, F., WOHMANN-BRUZZO, L., AUDIBERT, N., FOUGERON, C. (2023). Speaker-specific anticipatory labial coarticulation in French. In: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 654–658). Guarant International.

HERMES, A., MERTENS, J., & MÜCKE, D. (2018). Age-related Effects on Sensorimotor Control of Speech Production. *Proc. Interspeech 2018*, 1526-1530. DOI: 10.21437/Interspeech.2018-1233

JONGMAN, A., WAYLAND, R., & WONG, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am*. 108, 1252–1263. DOI : 10.1121/1.1288413

KAVANAGH, C. M. (2012). New consonantal acoustic parameters for forensic speaker comparison. Doctoral dissertation, University of York.

KOENIG, L. L., SHADLE, C. H., PRESTON, J. L., et MOOSHAMMER, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *J. Speech Lang. Hear. Res.* 56(4), 1175–1189. DOI : 10.1044/1092-4388(2012/12-0038)

LINVILLE, S. E. (1987). Acoustic-perceptual studies of aging voice in women. *Journal of Voice*, 1, 44–48. DOI : 10.1016/S0892-1997(87)80023-1

MARTINS-REIS, V. D. O. & DE ANDRADE, C. R. F. D. (2011). Study of pauses in elderly. *Rev. Soc. Bras. Fonoaud.* 16(3), 344–349. DOI : 10.1044/1092-4388(2012/12-0038)

QUENÉ, H. (2013). Longitudinal trends in speech tempo: The case of queen Beatrix. *The Journal of the Acoustical Society of Americ*a, 133, EL452–EL457. DOI : 10.1121/1.4802892

RAMIG, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders*, 163. 217–26. DOI : 10.1016/0021-9924(83)90035-7

RODRIGUEZ-ARANDA, C. & JAKOBSEN, M. (2011). Differential contribution of cognitive and psychomotor functions to the age-related slowing of speech production. *J. Int. Neuropsych. Soc*. 17, 1–15. DOI : 10.1017/S1355617711000828

SHADLE, C. H., CHEN, W.-R., KOENIG, L. L., & PRESTON, J. L. (2023). Refining and extending measures for fricative spectra, with special attention to the high-frequency range. *Journal of the Acoustical Society of America,* 154(3), 1932-1944. DOI: 10.1121/10.0021075

STUART-SMITH, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian, in *Laboratory Phonology 9*, edited by J. Cole and J. Hualde (Mouton de Gruyter, Berlin), 65–86. ISBN: 9783110186833

STUART-SMITH, J., TIMMINS, C., & TWEEDIE, F. (2007). Talkin' 'Jocney' ? Variation and change in Glaswegian accent. *Journal of Sociolinguistics*, 11(2), 221-260. DOI : 10.1111/j.1467-9841.2007.00319.

TUCKER, B. V., FORD, C., HEDGES, S. (2021). Speech aging: Perception and Production. *WIREs Cognitive Science,* e1557. DOI : 10.1002/wcs.1557

YAN, J. H., THOMAS, J. R., & STELMACH, G. E. (1998). Aging and rapid arm movement control. *Experimental Aging Research*, 24, 155–168. DOI : 10.1080/036107398244292

YENI-KOMSHIAN, G. H., & SOLI, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, 70, 966–975. DOI : 10.1121/1.387031

XUE, S. A., & HAO, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. *J. Speech Lang. Hear. Res.* 46, 689–701. DOI : 10.1044/1092-4388(2003/054)

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Effets du shadowing et de l'imitation en tant que méthodes d'entraînement à la prononciation du /ɥi/ en français

Wenxun Fu[1] Martine Adda-Decker[1, 2] Barbara Kühnert[1]

(1) Laboratoire de Phonétique et Phonologie : UMR 7018 CNRS / Sorbonne Nouvelle, 75005 Paris, France
(2) Laboratoire Interdisciplinaire des Sciences du Numérique : CNRS, Université Paris-Saclay, Orsay, France
wenxun.fu@sorbonne-nouvelle.fr,
martine.adda-decker@sorbonne-nouvelle.fr, barbara.Kuhnert@sorbonne-nouvelle.fr

RESUME

Trente étudiantes mandarines apprenant le français ont participé à un entraînement autonome de quatre semaines, utilisant l'imitation tardive et le shadowing (répétition immédiate). Cette étude se concentre sur le résultat de la réalisation du /i/ dans /ɥi/, souvent réalisé proche du /y/. Les posttests montrent des améliorations dans la perception et la production de /ɥi/. Pour les apprenants de niveaux intermédiaires ayant pratiqué le shadowing, la distinction entre le troisième formant (F3) et le deuxième formant (F2) du /i/ dans /ɥi/ est significativement plus élevé après l'entraînement, indiquant une meilleure distinction avec /y/. Le shadowing semble efficace dans l'amélioration de la perception chez les débutants et apprenants intermédiaires, mais uniquement dans la production pour les niveaux intermédiaires. Nous suggérons que le shadowing, en tant que méthode hautement cognitive et active, puisse servir d'alternative à la méthode d'imitation, sous réserve que la compétence linguistique des apprenants leur permette d'accomplir la tâche avec succès.

ABSTRACT

This paper presents the results of an outside-classroom French pronunciation training program using recorded native speakers' speech. Thirty female Mandarin learners of French participated in a self-directed four-week training program , using either imitation or shadowing. This study focuses on Mandarin learners' tendency to realize /i/ similarly to /y/ in the glide /ɥi/. Posttests show considerable improvements in both perceiving and producing /ɥi/ in some conditions. Notably, shadowing proved effective in enhancing perception for both beginner and intermediate levels. However, in production only intermediate or higher-level learners refined their differentiation between /i/ and /y/ (the difference between the third formant F3 and the second formant F2 of /i/ in /ɥi/ increased significantly post training). We suggest that shadowing, as a highly cognitive and active technique, could be a viable alternative to the more commonly used imitation method, provided the language proficiency of learners permits them to execute the task successfully.

MOTS-CLES : Shadowing, imitation, L2 pronunciation, français, apprenants mandarins
KEYWORDS : Shadowing, imitation, L2 pronunciation, French, Mandarin learners.

Actes des 35èmes Journées d'Études sur la Parole, pages 332–341.

1 Introduction

1.1 L’entraînement de la prononciation L2

Alors que l'importance de la pratique de la prononciation a été reconnue par un nombre croissant d'enseignants et d'étudiants en langues secondes (L2), en réalité, la précision phonétique est souvent reléguée au second plan en faveur de la fluidité et de la compréhension en classe en raison du temps limité à la langue parlée. En effet, l'entraînement à la prononciation impliquant des répétitions et un guide approprié est généralement chronophage, et les résultats des efforts déployés dans la pratique ne sont pas immédiatement perceptibles. Compte tenu des emplois du temps intensifs de la plupart des classes de L2, il serait idéal de pouvoir intégrer la pratique de la prononciation simultanément à d'autres cours , tels que les cours de communication orale, ou bien de permettre aux étudiants de la pratiquer de manière autonome en dehaor de la classe, en se quidant eux-mêmes. Cette étude se conscentre sur cette dernière option.

En dépit des nombreuses méthodes d'entraînement à la prononciation récemment développée, notamment celles impliquant l'assistance des ordinateurs, l'imitation de la parole des locuteurs natifs, souvent enregistrée, reste l'une des méthodes les plus populaires en raison de sa simplicité d'application et de son efficacité relative, comme l’attestent de nombreuses études. En effet, l'imitation est un comportement naturel dans la parole et existe même entre les locuteurs natifs, se produisant souvent inconsciemment (Nguyen et Delvaux. 2015). Les changements adaptatifs en ligne dans la production, en particulier les effets de la convergence phonétique, sont considérés comme facilitant l'échange conversationnel en favorisant la création d'un terrain commun entre locuteur et interlocuteur (Sato et al. 2013). Dans l'expérience de Sato et al. 2013, des locuteurs natifs français, à qui l'on demandait de produire des voyelles françaises sans instructions spécifiques d'imitation, imitaient automatiquement les caractéristiques acoustiques d'une voix préenregistrée. Cependant, l'effet de la convergence phonétique s’est avéré plus prononcé dans la condition d'imitation volontaire de la même expérience.

Par ailleurs, des recherches antérieures montrent que le degré de convergence phonétique lors de la tâche d'imitation dépend de l'intervalle entre l'entrée extérieure et la production imitée. Par exemple, Goldinger (1998) a démontré que dans une tâche de répétition, les participants natifs de l'anglais qui pratiquaient la répétition immédiate avaient un degré d'imitation plus élevé que ceux qui pratiquaient la répétition tardive, où un délai de 3 à 4 secondes était imposé avant de parler. La répétition immédiate, également appelée shadowing, a été définie par Tamai, l’un des premiers chercheurs l’utiliser dans les contextes japonais d’apprentissage d'anglais langue seconde, comme "un acte ou une tâche d'écoute dans lequel l'apprenant suit le discours entendu et le répète aussi exactement que possible tout en écoutant attentivement les informations entrantes" (cité dans Sumiyoshi, 2019). Lors du shadowing, les étudiants écoutent du matériel en L2 et tentent de le répéter instantanément, ce qui nécessite un traitement et une production rapides de la langue. Par conséquent, cette pratique peut aider les apprenants à développer leurs compétences de traitement ascendant (bottom-up) en renforçant leur capacité à percevoir et à produire les sons individuels et les caractéristiques phonétiques de la langue cible. Bien que le shadowing ait été initialement développé comme une méthode d’entraînement à l'interprétation simultanée, il est devenu une stratégie courante pour améliorer les capacités d'écoute des étudiants de l’anglais langue étrangère (Hamada, 2019 ; Hu, 2014 ; Lambert, 1992). En même temps, certains chercheurs notent que le shadowing auditif est une tâche cognitivement très exigeante que certains apprenants peuvent avoir du mal à mettre à éxectuer. Mori et al. (2011) soulignent que l'effet positif du shadowing n'atteindra son maximum que si les

apprenants en langues sont capables de le faire au sens réel. Afin de réduire la charge cognitive, des scripts sont parfois fournis dans les formations utilisant le shadowing. Par exemple, Hamada et Suzuki (2021) ont constaté que, au lieu d'être utilisé seul, le shadowing assisté par des scripts apporte plus d'avantages à l'adaptation perceptive des apprenants.

Les recherches sur l'efficacité du shadowing sur la production en L2 sont moins nombreuses. Cependant, des théories telles que celle de Hintzman (1986) suggèrent le potentiel du shadowing pour une imitation précise par rapport à l'imitation tardive. Selon Hintzman, une réponse en écho, qui est la seule base pour répondre dans la tâche de répétition, est composé d'informations mélangées- celle du stimulus et des épisodes déjà stockés dans la mémoire à long terme. Si la réponse est immédiate, sa similarité au stimulus devrait être considérable, qui entraîne une imitation forte. En revanche, si la réponse est générée lentement, l'écho devrait circuler entre la mémoire de travail et la mémoire à long terme, son contenu devenant moins similaire que le stimulus original. Compte tenu de l'utilité de shadowing actuelle et de la découverte de Goldinger (1998) selon laquelle les locuteurs ont tendance à imiter davantage pendant la pratique du shadowing que pendant l'imitation, il semble raisonnable de considérer le shadowing comme une méthode prometteuse pour l'apprentissage de la prononciation, en plus de l'imitation tardive. Dans l'étude de Hida (2020), les élèves japonais du premier cycle du secondaire ont amélioré leur prononciation des voyelles l'anglais non accentuées après une série d'activités en shadowing. Le présent article vise à explorer les avantages potentiels du shadowing en tant que méthode d'auto-apprentissage de la prononciation par rapport à l'imitation tardive, plus traditionnellement utilisée.

1.2 Quelques différences entre les systèmes vocaliques français et mandarins

La présence et la fréquence des voyelles orales arrondies constituent des caractéristiques distinctives du système phonologique français par rapport à celui du mandarin. En français, six des dix voyelles orales (à l'exclusion de la voyelle centrale /ə/), /y, oe, ø, u, o, ɔ/ sont arrondies. En revanche, en mandarin, seules deux voyelles arrondies sont confirmées, à savoir /y/ et /u/, tandis que l'arrondie des mi-voyelles est aussi discutable que l'inventaire même de ces mi-voyelles. Il est intéressant de noter que nos recherches précédentes (Fu et al. 2023) ont révélé que les apprenants mandarins ne prononçaient pas mieux les voyelles arrondies présentes dans les deux inventaires ; les voyelles arrondies en français ont tendance à présenter des qualités plus arrondies que celles du mandarin.

Une autre différence notable dans l'inventaire des voyelles entre les deux langues est la richesse des diphtongues en mandarin. Cependant, contrairement à Lee et Zee (2003) qui ont identifié onze diphtongues en mandarin, Duanmu (2007) n'inclut que quatre diphtongues dans l'inventaire phonétique du mandarin, soit /ai/, /au/, /əi/, et /əu/, et considère /uo/, /ie/, /ia/, /ye/, /ua/, /uə/, et /iu/ comme les combinaisons d'une semi-voyelle (ou glide) et d'une voyelle. Parmi les combinaisons de glides et de voyelles en français**,** notre étude précédente a montré que /ɥi/ tend à être particulièrement difficile pour les apprenants mandarins du niveau intermédiaire ou avancé. Précisément, comparé à la production des locuteurs natifs français, /i/ dans /ɥi/ était plus proche de /y/, indiqué par un F3 plus bas (qui est associé à l'arrondissement de voyelles) et une différence plus petite entre F3 et F2 que celle observée chez les locuteurs natifs français qui produisaient /i/ dans /ɥi/ sans différences significatives dans les trois premiers formants de /i/ non dans /ɥi/. L'assimilation de /ɥi/ à /y/ pour les apprenants mandarins peut être attribuée à l'absence de cette combinaison dans la langue maternelle des apprenants et à l'insuffisance des instructions et de l'accent mis sur la transition du mouvement articulatoire de la prononciation de /ɥ/ à /i/.

2 Objectifs de l'étude

En premier lieu, notre étude vise à aider les apprenants mandarins du français à améliorer la prononciation de sons préalablement identifiés comme particulièrement problematiues (Fu et al. 2023 ; 2022), sans empiéter sur leur temps d'apprentissage en classe. Dans cet article, nous nous concentrons sur les résultats de /ɥi/, souvent remplacé par /y/ selon nos recherches antérieures. Un autre objectif est de comparer les différents effets de l'imitation tardive et de la répétition immédiate (simplifiées à "imitation" et "shadowing" dans cette étude). Pour'optimiser la facilité d'accès et l'exhaustivité des tâches, notre programme d'entraînement comprend 1) des discours préenregistrés de locuteurs natifs français servant de modèles, 2) des scripts de matériel d'entraînement fournis à tous les participants, 3) des instructions phonétiques explicites avant la pratique de la production. Nous émettons l'hypothèse qu'après le programme d'entraînement, la distinction entre /ɥi/ et /y/ sera améliorée du point de vue perceptive, et que /i/ dans /ɥi/ sera mieux prononcé, indiqué par une différence plus grande entre le deuxième et le troisième formant (F2, F3). De plus, nous anticipons que les participants qui ont pratiqué avec succès le shadowing comme méthode principale de formation connaîtront une amélioration plus importante par rapport à ceux qui ont complété la tâche d'imitation tardive.

3 Méthodes

Figure 1 montre le déroulement de l'entraînement autonome dont l'efficacité était mesurée par la comparaison d'un prétest et d'un posttest en perception et en production.

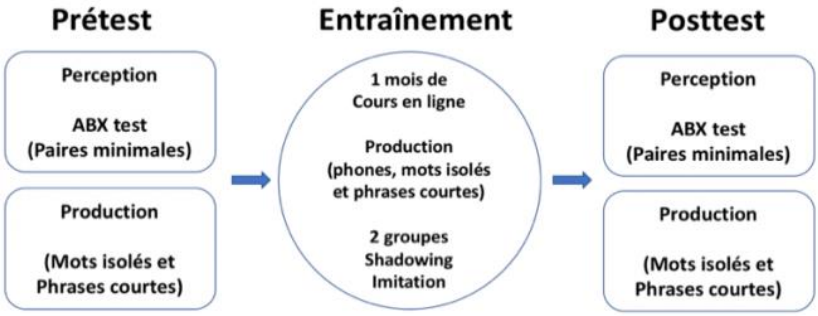


Figure 1 : Déroulement de l'expérience de l'entraînement à la prononciation

3.1 Participants

Trente étudiantes femmes du Département du français de l'Université des Langues et de Cultures de Pékin ont participé à l'ensemble de l'expérience, qui comprenait un questionnaire, deux prétests, quatre semaines d'entraînement à la prononciation et deux posttests. Toutes les étudiantes qui avaient passé les tests de base ont été répartis au hasard entre le groupe de shadowing et le groupe d'imitation. Parmi les 30 étudiantes dont les données ont été analysées dans le cadre de l'étude, 16 étudiantes faisaient partie du groupe de shadowing et 14 du groupe d'imitation. En outre, 17 étudiantes (âgées de 19 à 20 ans) étaient en deuxième année d'étude du français, ce qui correspondait à un niveau intermédiaire dans la présente étude, tandis que 13 étudiantes (âgées de 18 à 19 ans) étaient en première année, n'avaient étudié le français que pendant trois mois et demi, et se trouvaient donc à un niveau débutant au sens propre. Avant le début de l'expérience, les élèves ont été informés que ceux qui iraient jusqu'au bout de l'expérience recevraient 15 euros de compensation.

3.2 Tests de perception

Le test de perception est un test ABX créé à l'aide de jsPsycho et JATOS. Les stimuli utilisés dans le test proviennent d'une locutrice française native qui a également servi de modèle pour l'entraînement. Le test est composé de 48 paires minimales (par exemple *vœux, feu*), comprenant 5 paires opposant /ɥi/ et /y/ (*lui, lu*). Chaque paire minimale a été testée deux fois dans le cadre ABX totalisant ainsi 10 réponses par participant. Par exemple, "pluie" et "plu" ont été trouvés dans les triplets "pluie, plu, pluie" et "pluie, plu, plu. Avant de passer le test proprement dit, les participants ont fait une phase d'adaptation avec quatre triplets qui n'étaient pas analysés dans l'étude.

3.3 Tests de production

Le prétest de production comprenait la lecture de mots isolés et de phrases courtes, chacune d'entre elles contenant un mot isolé lu précédemment. Par exemple, la phrase correspondant au mot isolé "lui" était "C'est un texte à lui". Les mots cibles ont été placés en dernière position de la phrase afin d'éviter les influences potentielles de l'accentuation. Tous les mots isolés et toutes les phrases ont été lus deux fois de suite. Deux paires minimales (*lui, lu ; pluie, plu*) ont été entraînées au cours des semaines, tandis que deux triplets (*cuir, cure, kir ; suie, su, scie*) n'ont pas été entraînés. Cette conception nous a permis d'observer si l'effet de l'entraînement pouvait être généralisé aux nouveaux mots ou, au contraire, si l'effet restait spécifique aux mots entraînés.

Deux posttests, l'un de perception et l'autre de production, ont été complétés dans les deux semaines après le programme d'entraînement. Ces tests comprenaient le même corpus que les prétests, mais dans un ordre aléatoire différent.

3.4 Entraînement

Le programme d'entraînement, qui a débuté dans la semaine suivant les prétests, s'est étalé sur quatre semaines et a consisté en 12 séances, dont quatre étaient destinées à entraîner /ɥi/ et /y. Les sons contrastés ont été entraînés une fois par semaine, 20 minutes chacun en moyenne, avec un niveau de difficulté progressivement accru.

Les participants ont reçu des textes et des enregistrements comme matériel d'apprentissage pour leur formation. Chaque session est composée d'une illustration phonétique en mandarin, soulignant les différences importantes entre les deux sons contrastés, d'une pratique des sons individuels et d'une pratique de mots isolés ou de phrases contenant les mots cibles. Lorsque les élèves écoutaient l'instruction phonétique, on leur demandait de se référer au script dans lequel ils pouvaient voir les représentations graphiques. Les élèves ont ensuite écouté les démonstrations audios des sons isolés, d'abord les sons prolongés puis les sons réguliers, avant de commencer à s'exercer seuls. Toutes les démonstrations ont été enregistrées par deux professeurs de langue maternelle. Pour les mêmes sons contrastés, chaque professeur a enregistré deux sessions afin de s'assurer que les étudiants avaient les mêmes possibilités d'écouter la parole féminine et la parole masculine. Pour chaque son, mot et phrase cible, les étudiants ont été invités à s'entraîner cinq fois en imitant ou en shadowing, selon le groupe auquel ils appartiennent. Les participants ont soumis leur pratique enregistrée au cours de la semaine avant de recevoir le nouveau matériel d'apprentissage pour la semaine suivante. Bien que la pratique enregistrée n'ait pas été analysée dans l'étude, sa soumission a permis de garantir le temps global de pratique et d'éviter que les participants ne fassent plus d'une session de l'exercice par semaine.

La principale distinction entre les supports des deux groupes réside dans les instructions précisant les méthodes spécifiques à appliquer pendant l'entraînement à la prononciation automatisée. Dans le groupe pratiquant le shadowing, où les participants devaient répéter les éléments dès qu'ils entendaient le son de démonstration, une instruction typique était la suivante : "Maintenant, pratiquons ensemble la prononciation régulière (par opposition à la prononciation prolongée) de /ɥi/ ; veuillez répéter le son dès que vous entendez le professeur dire /y/". Dans la condition d'imitation, la deuxième phrase de la consigne a été modifiée en "Veuillez imiter /ɥi/ après le professeur". Afin de souligner davantage la différence entre les deux méthodes et de s'assurer que chaque participante comprenait sa tâche spécifique, un exercice de démonstration a été enregistré et envoyé aux étudiantes avant le programme d'entraînement. Les instructions ont été lues en mandarin et les scripts ont été fournis à la fois en mandarin et en français.

4 Résultats

4.1 Tests de perception

Dans le test de perception, 10 triplets ABX ont été créés pour tester la perception de /ɥi/ et /y/. Au total, 300 réponses ont été recueillies dans le prétest et le posttest respectivement. Dans le prétest, un total de 11 réponses incorrectes a été enregistré, représentant 3,7 % du total des réponses, ce qui suggère que les étudiants étaient bien capables de faire la distinction entre /ɥi/ et /y/ en général. Le nombre de réponses incorrectes a encore diminué pour atteindre 6 (2 %) après la formation. Si l'on examine les performances individuelles, seuls six étudiants ont commis des erreurs dans le contraste après l'entraînement, trois dans chaque condition et deux au niveau intermédiaire.

Au départ, deux participants du groupe de shadowing (Sujet 11 du niveau débutant et Sujet 8 du niveau intermédiaire ont donné trois mauvaises réponses sur dix tests, atteignant un pourcentage de mauvaises réponses de 30 %. Après l'entraînement, les deux participants n'ont plus commis d'erreur dans le posttest de perception. En revanche, le sujet 4 du niveau intermédiaire venait de commettre une erreur. D'autre part, les étudiants qui se sont entraînés avec la méthode de l'imitation au départ semblent avoir obtenu de meilleurs résultats que ceux qui pratiquaient le shadowing avant l'entraînement, avec quatre réponses erronées de la part de trois étudiants. Les sujets 10 et 13, tous deux du niveau intermédiaire, ont corrigé les réponses incorrectes après l'entraînement, tandis que davantage de sujets ont fait une seule erreur dans le posttest, qui serait également dû à une erreur de frappe au clavier. Dans l'ensemble, la distinction n'a pas posé de problèmes de perception graves dès le départ, et les participants qui commettaient initialement plus d'erreurs semblent avoir une amélioration satisfaisante. L'analyse statistique n'était pas mise en place lors du test de perception étant donné le faible nombre d'erreurs avant et après l'entraînement.

4.2 Tests de production

Toutes les données des enregistrements ont été traitées à l'aide du système d'alignement forcé du LIMSI (Gauvain et al. 2002) avant une correction manuelle. En tout 483 occurrences de /ɥi/, 481 occurrences de /y/ , et 238 occurrences de /i/ (pas dans /ɥi/) dans des mots ciblés ont été collectées à partir du prétest et du posttest de la production des mots isolés. Dans la production de phrases courtes, 506 occurrences de /ɥi/, 481 occurrences de /y/ et 238 occurrences de /i/ produites indépendamment ont été collectées et incluses dans les analyses. Nos analyses ont été spécifiquement porté sur la différence entre F2 et F3 (F3-F2) de /i/ dans /ɥi/. Pour chaque son, une seule valeur du paramètre a été obtenue en faisant la moyenne des mesures effectuées à 1/3, 1/2 et

2/3 du segment, les valeurs de formants ont été extraites à l'aide d'un script Praat (Boersma et Weenink, 2022). Les analyses statistiques ont été réalisées à l'aide d'un modèle linéaire mixte (lme4 package, Bates et al., 2015) dans l'environnement R (R Core Team 2024). La formule du modèle (F3-F2~Test* Condition* Entraînement du mot * Niveau* Tâche+ (1| Participant)+ (1| Mot)), comprenait cinq facteurs fixes codés par contraste : Test (prétest vs. posttest), Condition (imitation vs. shadowing), Niveau (débutant ou licence 1 vs. intermédiaire ou licence 2), Entraînement du mot (entraîné vs. non-entraîné), Tâche (mots vs. phrases) et toutes leurs interactions, avec deux intercepts aléatoires pour le Participant et le Mot individuel. La significativité de chacun des facteurs fixes contenus dans le modèle a ensuite été évaluée grâce à des ANOVA de type III. Les résultats montrent un effet de Tâche (χ^2 =4.8042, df = 1, $p < 0.05$), l'interaction Test Niveau (χ^2 =5.377, df = 1, $p < 0.05$), l'interaction Test x Condition (χ^2 =6.8731, df = 1, $p < 0.01$), l'interaction Niveau x Tâche (χ^2 =6.0454, df = 1, $p < 0.05$), les interactions Test x Niveau x Condition (χ^2 = 8.8165, df = 1, $p<0.01$), et les interactions Condition x Niveau x Tâche (χ^2 =7.3762, df = 1, $p < 0.01$). Les tests posthoc étaient réalisés grâce à emmeans package (Lenth, 2020)

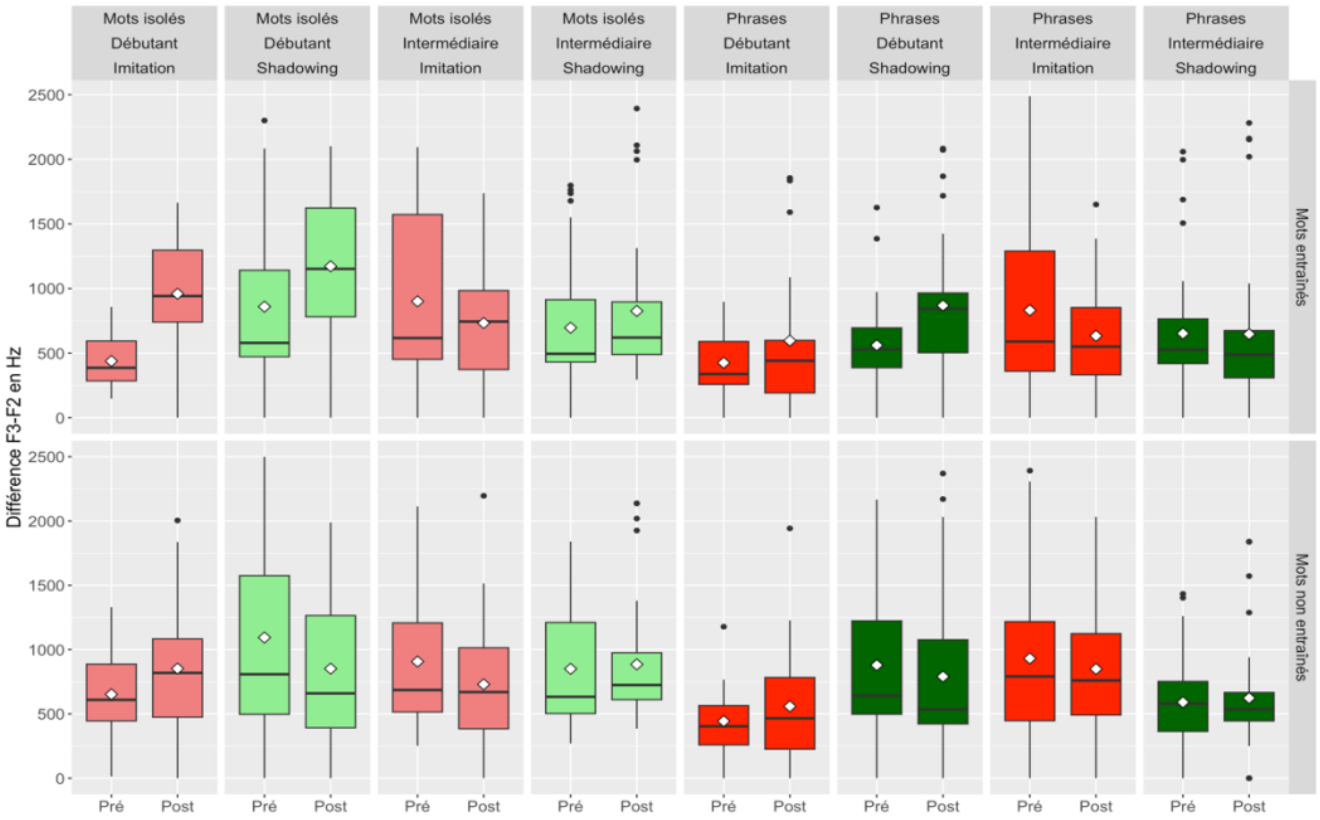


Figure 2 : Diagramme en boîte de la différence F3-F2 en Hz pour /i/ dans /qi/ produit par les apprenants mandarins selon la condition, le niveau et l'entraînement des mots dans la production des mots isolés et des phrases courtes (le carré blanc indique la moyenne)

Selon les tâches, le degré de l'augmentation de la différence F3-F2 est différent. Dans la tâche de lecture de mots isolés, la différence entre le pretest et le posttest est significative (t =2.002, $p<0.05$) : les étudiantes de deux niveaux dans les deux conditions ont augmenté la différence F3-F2 pour les mots entraînés après l'entraînement, bien que cette augmentation soit plus légère pour les étudiantes de niveau intermédiaire en imitation. L'effet positif de l'entraînement n'a pas pu être transmis aux mots non entraînés pour les étudiantes de niveau débutant pratiquant le shadowing et pour celles de niveau intermédiaire pratiquant l'imitation. Dans l'ensemble, l'amélioration n'est pas significative

dans la tâche de lecture de phrases ($t=1.027$, $p>0.1$). Toutefois, quand les phrases contiennent les mots entraînés, l'augmentation de la différence F3-F2 pour les étudiantes de niveau débutant dans les deux conditions reste toujours présente, ce qui n'est pas le cas pour les étudiantes de niveau intermédiaire. Pour celles qui ont complété l'entraînement de l'imitation, la différence F3-F2 a diminué dans le posttest. En revanche, pour les étudiants de ce niveau qui ont pratiqué le shadowing la différence F3-F2 a légèrement augmenté. Cette même augmentation a pu être observé pour les mots non entrainés compris dans les phrases.

Globalement, les étudiantes de niveau débutant ($t=3.873$, $p<0.001$), qui avaient initialement une prononciation moins précise du /i/ dans /ɥi/, se sont améliorés de manière plus marquante que les étudiantes de niveau intermédiaire. Pour eux, l'effet de l'imitation semble être plus solide, se manifestant même aux mots jamais entraînés. Cependant, la même méthode semble moins efficace pour les étudiantes ayant appris le français depuis plus d'un an. Malgré cela, les étudiantes de ce niveau ont néanmoins pu bénéficier du shadowing comme méthode de l'entraînement, avec un effet positif observé également dans les mots non entraînés et dans la tâche de lecture de phrase, qui est légèrement plus complexe que la lecture de mots isolés.

5 Discussion

Le défi de prononcer /ɥi/ réside principalement au niveau de la production du bigramme de phones (phonotactique), car la plupart des participants, quelle que soit leur compétence linguistique, étaient capables de percevoir la différence entre /y/ et /ɥi/ dans le test de perception, et que la voyelle /i/ n'est pas un son inconnu pour les apprenants mandarins. Dans ce contexte, l'acquisition de l'enchaînement parait plus difficile dans la production. Notre étude démontre certains effets positifs de l'utilisation du shadowing comme méthode d'entraînement à la prononciation. En écoutant attentivement et en suivant de près le modèle de parole des locuteurs natifs, les apprenants peuvent intérioriser les mouvements articulatoires nécessaires pour produire des sons difficiles comme /ɥi/, comblant ainsi le fossé entre la perception et la production. Néanmoins, comparée à l'imitation tardive, une tâche plutôt facile à compléter, la méthode shadowing semble offrir moins d'avantages supplémentaires aux apprenants de niveau débutants. Une des limites de notre étude reste dans les conditions de suivi restreintes ; pour les participantes utilisant des écouteurs pour écouter l'enregistrement modèle, nous n'avons pas pu vérifier avec quelle précision ils suivaient les instructions de shadowing. Certaines participantes, trouvant la tâche difficile, auraient pu inconsciemment la remplacer par l'imitation tardive. Dans l'ensemble, les participantes de niveau intermédiaire avaient une meilleure prononciation avant l'entraînement, mais l'amélioration était aussi moins évidente que les étudiantes de la première l'année. En même temps, elles ont bénéficié davantage de la pratique du shadowing que l'imitation, et les effets positifs se sont également étendus aux mots qui n'avaient pas fait l'objet d'un entraînement spécifique. D'ailleurs, ce groupe de participantes est également le seul qui a réduit la durée des toutes les voyelles entraînées dans la production des mots isolés et la voyelle /y/ dans la production des phrases. Cependant, pour obtenir une amélioration plus significative, notamment dans la lecture de textes, un style d'expression orale considéré comme très difficile pour les apprenants de L2, dans lequel les prononciations de L2 sont notablement différentes de celles de L1 (Fu et al., 2023) - le programme d'entraînement devrait être prolongé au-delà de la durée de notre étude, et le matériel fourni pourrait être simplifié davantage.

Références

BATES D., MACHLER M., BOLKER B., & WALKER S. (2015). Fitting linear mixed-effects models Usinglme4. Journal of Statistical Software, 67(1). DOI : 10.18637/jss.v067.i01

BOERSMA P. & WEENINK D. (2022). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

DARCY I. & ROCCA B. (2022). Comprehensibility improvements in integrated pronunciation instruction: A comparison of instructional methods and task effects. *Journal of Second Language Pronunciation*, *8*(3), 328-362. DOI : 10.1075/jslp.21035.dar

DARCY I., ROCCA B., & HANCOCK Z. (2021). A window into the classroom: How teachers integrate pronunciation instruction. *RELC Journal*, *52*(1), 110-127. DOI : 10.1177/0033688220964

DUANMU S. (2007). The phonology of standard Chinese. 2nd edn. New York: Oxford University Press.

FU W., ADDA-DECKER M., & KUHNERT B (2023). Characterization of Mandarin-accented French across three different speaking styles: a corpus-based study. *Proceedings of the 20th International Congress of Phonetic Sciences* (pp.2507-2511). Prague, Czech Republic. https://guarant.cz/icphs2023/341.pdf

FU W., ADDA-DECKER M., & KUHNERT B (2022). The production of oral vowels by Mandarin L2 learners of French: characterization for a training program.*10th International Symposium on the Acquisition of Second Language Speech (New Sounds 2022)* University of Barcelona, Spain. https://shs.hal.science/halshs-03984432

GAUVAIN J L., LAMEL L, & ADDA G (2002). *The Limsi Broadcast News Transcription System, Speech Communication,* 37(1-2):89-108. https://doi.org/10.1016/S0167-6393(01)00061-9

GOLDINGER S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105(2), 251-279. https://doi.org/10.1037/0033-295X.105.2.251

HU W., ZHAO K., ZHAO S., ZHI F., XIAO, C., & YANG, J. (2023). The Effect of Non-Native English Accent on Second Language Listening Comprehension. Open Access Library Journal, 10(4), 1-12. DOI : 10.4236/oalib.1110078

HAMADA Y. (2019). Shadowing: What is it? How to use it. Where will it go?. *RELC Journal*, *50*(3), 386-393. DOI : 10.1177/0033688218771380

HAMADA Y. & SUZUKI S. (2021). Listening to Global Englishes: Script-assisted shadowing. *International Journal of Applied Linguistics*, *31*(1), 31-47. DOI : 10.1111/ijal.12318

HIDA K. A. Z. U. K. I. (2020). The effectiveness of shadowing in English weak vowels acquisition: A study of Japanese junior high school students. *Dialogue*, *18*, 1-20. https://talk-waseda.net/dialogue/no18_2020/dialogue18_k1_hida.pdf

LAMBERT S. (1992). Shadowing. *Meta*, *37*(2), 263-273.

Lenth R. (2020). emmeans: Estimated marginal means, aka least- squares means. https://CRAN.R-project.org/package=emmeans

LEE W. S. & ZEE E. (2003). Standard chinese (beijing). *Journal of the International Phonetic Association*, *33*(1), 109-112. DOI: 10.1017/S0025100303001208

MORI T., YASUDA T., MAEDA T., MIZUBAYASHI W., LIU Y., SAKAMOTO K. & OTA H. (2011). Tunnel field-effect transistors with extremely low off-current using shadowing effect in drain implantation. *Japanese journal of applied physics*, *50*(6S), 06GF14. DOI : 10.1143/JJAP.50.06GF14

NGUYEN N. & DELVEUX V. "Role of imitation in the emergence of phonological systems." *Journal of Phonetics*53 (2015), 46-54. DOI: 10.1016/j.wocn.2015.08.004

SUMIYOSHI H. (2019). The effect of shadowing: exploring the speed variety of model audio and sound recognition ability in the japanese as a foreign language context. *Electronic Journal of Foreign Language Teaching*, *16*(1), 5-21. https://e-flt.nus.edu.sg/v16n12019/sumiyoshi.pdf

SATO M., GRABSKI K., GARNIER M., GRANJON L., SCHWARTZ J. L. & NGUYEN, N. (2013).

Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in psychology, 4,* 422. DOI : 10.3389/fpsyg.2013.00422

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Enseignement de l'intonation du français par une synthèse vocale contrôlée par le geste : étude de faisabilité

Xiao Xiao[1] Corinne Bonnet[2] Haohan Zhang[3] Nicolas Audibert[3] Barbara Kühnert[3] Claire Pillot-Loiseau[3]
(1) Léonard de Vinci Pôle Universitaire, Research Center, 12, avenue Léonard de Vinci, Paris – La Défense 92400, France
(2) DILTEC Didactique des langues, des textes et des cultures EA 2288, Sorbonne Nouvelle, 4, rue des Irlandais, 75005 Paris, France, et Université Toulouse Paul Sabatier, 118 Rte de Narbonne, 31062 Toulouse, France
(3) Laboratoire de Phonétique et Phonologie UMR 7018, Sorbonne Nouvelle, 4, rue des Irlandais, 75005 Paris, France
xiao.xiao@devinci.fr, {corinne.bonnet, haohan.zhang nicolas.audibert, barbara.kuhnert, claire.pillot}@sorbonne-nouvelle.fr

RESUME

Peut-on enseigner l'intonation française en classe avec une synthèse vocale contrôlée gestuellement sur une tablette ? La fréquence fondamentale et la durée de quatre phrases déclaratives, quatre questions polaires, quatre énoncés exprimant l'incrédulité (1 à 4 syllabes) de deux apprenantes ukrainiennes débutantes en français ont été comparées avant et après quatre entraînements hebdomadaires. Les apprenantes devaient écouter un enregistrement de référence, puis visualiser le modèle sur la tablette, tracer l'intonation manuellement, écouter le résultat synthétisé, et tracer et écouter leur tracé sans guide. Elles produisaient initialement des phrases déclaratives avec une intonation ascendante, et ont différencié les déclarations et les questions polaires après l'entraînement. L'expression de l'incrédulité s'est améliorée pour l'une. L'autre a montré quelques difficultés à maîtriser cette technologie. Cette première étude de cas utilisant la synthèse vocale contrôlée gestuellement est une approche prometteuse permettant plus de pratique de l'intonation en classe.

ABSTRACT

Teaching French intonation using gesture-controlled speech synthesis: a feasibility study.
Is learning French intonation with gesture-controlled speech synthesis on a tablet in the classroom feasible? The fundamental frequency and duration of four declarative sentences, four polar questions, four statements expressing incredulity (1 to 4 syllables) of two Ukrainian beginner learners of French were compared before and after four weekly training sessions. Learners were asked to listen to a reference recording, then look at the intonation guide on the tablet, trace the intonation manually, listen to the synthesized result, and trace and listen to their tracing without a guide. They initially produced declarative sentences with rising intonation and differentiated statements and polar questions after training. The expression of disbelief improved for one of them. The other showed some difficulty in mastering this technology. This first case study using gesture-controlled speech synthesis is a promising approach for more intonation practice in the classroom.

MOTS-CLES : synthèse vocale contrôlée par le geste, intonation, français, salle de classe
KEYWORDS : gesture-controlled vocal synthesis, intonation, French, classroom setting

Actes des 35èmes Journées d'Études sur la Parole, pages 342–350.

1 Introduction

L'acquisition de l'intonation peut s'avérer difficile pour les apprenants non natifs (Mennen, 2015). Plusieurs stratégies d'enseignement ont été proposées et testées, visant à attirer l'attention sur les changements de hauteur pour la perception et la production. L'association entre un son et une représentation visuelle ou gestuelle de son contour de fréquence fondamentale (f_o) peut aider un apprenant à percevoir un mouvement f_o non familier et à ancrer l'établissement de nouvelles catégories de mouvements de hauteur (Yuan et al., 2019). Les représentations visuelles des contours f_o peuvent également transmettre les différences entre un modèle et les productions des apprenants, pour les aider à comprendre et à corriger leurs propres erreurs (Taniguchi et Abberton, 1999). Pour la production, les gestes ont été utilisés pour renforcer kinesthésiquement les caractéristiques de l'intonation (entre autres : Baills et al., 2022).

Notre recherche explore comment la synthèse vocale contrôlée en temps réel par des gestes de la main, appelée Synthèse Vocale Performative (PVS, Locqueville et al., 2020 ; Xiao et al., 2023), peut être utilisée par des locuteurs non-natifs pour la pratique de l'intonation d'une langue étrangère. Des études pilotes de PVS avec des apprenants de L2 ont été menées sur des corpus français et anglais, à l'aide de l'interface Gepeto sur tablette mobile. Ces premières études ont permis de valider le potentiel de l'utilisation de PVS pour l'apprentissage de l'intonation (Xiao et al., 2021, 2023). Auparavant, le déploiement de Gepeto était limité à des études à session unique dans des conditions de laboratoire contrôlées. Le présent travail explore la manière dont Gepeto peut être incorporé de manière longitudinale pour compléter l'enseignement des langues étrangères en classe.

Dans cet article, nous présenterons une étude de cas sur l'acquisition du français par des apprenants ukrainiens dont l'une des difficultés de communication réside dans la compréhension et la production de la distinction prosodique entre les énoncés et les questions polaires (questions oui/non) en français. Certaines de ces difficultés pourraient être dues aux différences entre les systèmes prosodiques du français et de l'ukrainien : l'ukrainien est une langue slave avec une accentuation lexicale libre. Les phrases déclaratives et les questions "wh" sont caractérisées par un contour d'intonation descendant, tandis que les questions polaires sont produites avec un contour d'intonation ascendant. Cependant, l'ukrainien possède également des accents de hauteur qui signalent une focalisation large : dans ce cas, ils sont produits avec un accent prénucléaire ascendant (bas-haut) suivi d'un accent nucléaire descendant (haut-bas). Pour ces raisons, certaines phrases déclaratives peuvent présenter des schémas intonatifs ascendants (Pompino-Marschall et al., 2017). Le français est une langue romane avec un accent sur la dernière syllabe d'un groupe de mots (Fougeron et Smith, 1993). En français, sans expression d'une attitude particulière, les phrases déclaratives se terminent par un contour descendant, tandis que les questions polaires se terminent avec un contour ascendant (Di Cristo, 2016).

Cette étude de cas cherche à savoir si l'entraînement à l'utilisation de la PVS permet aux apprenants ukrainiens débutants de français d'améliorer leur production différenciée de phrases déclaratives, de questions polaires et d'un modèle d'intonation attitudinal (incrédulité) dans des énoncés courts.

2 Matériel et Méthodes

L'interface Gepeto a été utilisée pour les conditions gestuelles de l'étude. Elle consiste en une interface mobile personnalisée qui contrôle un synthétiseur vocal, Voks, qui permet le contrôle mélodique et rythmique en temps réel d'échantillons préalablement enregistrés ou de synthèse vocale par l'utilisation de gestes de la main (Locqueville et al., 2020). L'interface mobile fonctionne dans le navigateur d'une tablette ou d'un téléphone portable. Cette étude a utilisé les téléphones mobiles

personnels de chaque sujet, contrôlés par des mouvements du bout des doigts. Le tracé d'une courbe dans la zone de contrôle de l'interface du téléphone portable produit une resynthèse en temps réel de la phrase actuelle à partir de Voks. L'axe horizontal détermine la position temporelle de l'échantillon original à resynthétiser, et le moment de la resynthèse est déterminé par le moment du geste de l'utilisateur. L'axe vertical de la région de contrôle module la hauteur de la sortie. Il est régulièrement espacé sur une échelle de demi-tons (ST) avec une gamme de 24ST (2 octaves) calibrée autour du corpus de l'étude. Un panneau de boutons apparaît à gauche de la région de contrôle. Le bouton du haut permet de basculer entre le mode fondu et le mode maintenu pour la trace de l'utilisateur, où la trace disparaît après 1,5 seconde ou reste jusqu'à ce qu'elle soit effacée. En mode maintenu, trois autres boutons sont activés, permettant à l'utilisateur de lire, d'effacer ou d'enregistrer le tracé du geste en cours. Le bouton situé dans le coin inférieur gauche déclenche la lecture de l'enregistrement de référence pour la phrase en cours. L'enseignant dispose d'une interface de commande distincte qui lui permet de sélectionner la phrase en cours et d'activer ou de désactiver le guide visuel montrant la courbe d'intonation de la phrase de référence.

Le profil linguistique des apprenants a été documenté via un questionnaire initial : langue maternelle et autres langues apprises, durée du séjour en France et pratique du français, niveau de français, formation dans leur pays et en France, et utilisation actuelle du français et de la langue maternelle.

2.1 Contexte général de l’étude

Une classe d'étudiants ukrainiens adultes de niveau débutant a été sélectionnée pour l'étude. Les sujets ont été recrutés dans le cadre d'un cours d'introduction au français qui s'est déroulé au printemps 2023 à l'Université Paul Sabatier de Toulouse. Le cours d’1,5 heure a eu lieu pendant 14 semaines consécutives et a couvert la grammaire française de base et le vocabulaire lié à la vie quotidienne. La classe suivait un programme standard, dont la partie prononciation ne couvrait que la phonétique segmentale, mais pas l'intonation. Notre intervention s'est déroulée sur 6 séances hebdomadaires, dont 4 séances d'entraînement (20-30 minutes) à l'utilisation de Gepeto. Des enregistrements d’évaluation ont été effectués lors de la première et de la dernière séance. Les sessions de formation ont été supervisées par une formatrice en langues étrangères, qui a guidé les étudiants pendant les différents exercices, les a aidés avec les aspects techniques de l'interface et a pris des notes détaillées sur ses observations et ses échanges avec les apprenants. L'entraînement et les enregistrements d’évaluation ont utilisé un corpus d'énoncés courts (1-4 syllabes) produits avec des intonations différentes selon le contexte de communication (affirmation, question ou incrédulité).

Les interventions se sont déroulées sur 6 sessions de cours dans une salle séparée à côté de la salle de classe. La première et la dernière séance ont été consacrées aux pré-tests et aux post-tests, et les séances intermédiaires ont consisté en des activités destinées à entraîner les schémas d'intonation. Les sujets se rendaient individuellement dans la salle avant ou après le cours et réalisaient les activités prévues avec l'expérimentateur. Les enregistrements d’évaluation avaient pour but d'identifier les points de difficulté et de mesurer les améliorations apportées par l'entraînement.

2.2 Sujets

Les étudiants étaient des ressortissants ukrainiens installés en France entre 1,5 et 12 mois avant l'étude. 7 femmes et 3 hommes ont participé au pré-test (19-65 ans, moyenne 29,5 ans). L’ukrainien était leur première langue et l'anglais ou le russe (pour 7 sujets) leur deuxième ou troisième langue. Tous les sujets étaient débutants en français, avec un niveau entre A0 et A2 (Conseil de l’Europe, CECR 2001). Seules deux personnes ont participé à toutes les sessions d’entraînement : UF1 (femme,

37 ans, vivant en France depuis 1 an, niveau de français A1, chercheuse) et UF2 (femme, 65 ans, vivant en France depuis 6 mois, niveau de français A0, chimiste). Leurs deuxième et troisième langues apprises sont respectivement l'anglais et le russe. Les deux apprenantes n'ont aucune pratique musicale et n'ont jamais eu d'entraînement préalable à la mélodie du français parlé.

2.3 Corpus

Le corpus se compose de mots et de phrases françaises courtes de 1 à 4 syllabes enregistrées par un locuteur natif masculin avec trois modalités distinctes associées à des schémas d'intonation différents : affirmation (déclaration), question ou incrédulité. Le corpus a été divisé en trois groupes. Les énoncés du groupe A ont été utilisés pendant le pré-test ; ceux du groupe B lors des sessions de formation, et ceux du groupe C lors du post-test. Nous avons choisi des énoncés courts pour éviter les problèmes liés au rythme des phrases plus longues que nous avons rencontrés dans nos études précédentes (Xiao et al., 2021). De plus, les énoncés courts sont plus faciles à reproduire pour les apprenants de français de niveau débutant. Enfin, différents mots ont été enregistrés dans le pré- et le post-test afin de tester la capacité des apprenants à généraliser et à appliquer les modèles d'intonation qu'ils ont appris à d'autres exemples. Les enregistrements ont été réalisés sur un ordinateur portable à l'aide d'un microphone de bureau unidirectionnel à condensateur en utilisant le logiciel Audacity.

2.4 Analyse acoustique

La fréquence fondamentale (f_0) a été extraite à l'aide de scripts Praat (Boersma et Weenink 2023). Les valeurs minimales et maximales possibles ont été fixées en fonction de la portée observée du locuteur afin de minimiser les erreurs de détection. Les contours f_0 obtenus ont été inspectés pour éliminer les erreurs de détection. Après conversion en demi-tons, la position temporelle des points de mesure et les valeurs f_0 associées ont été extraites.

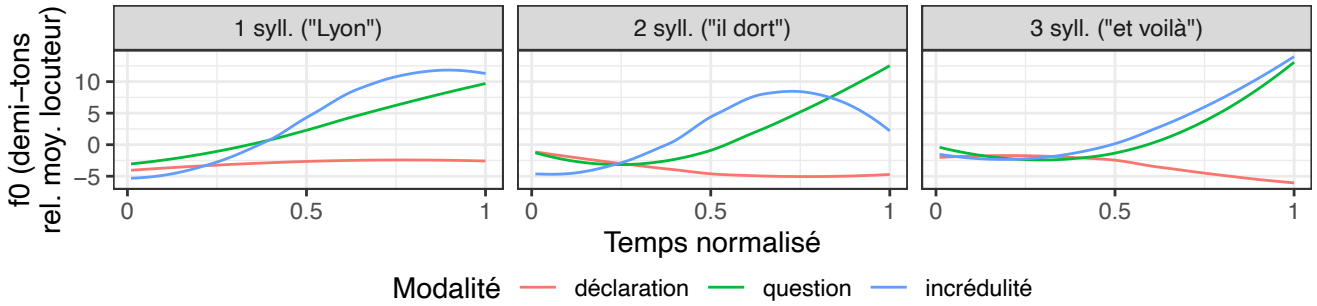


FIGURE 1 : Courbes de fréquence fondamentale (f_0) pour un énoncé monosyllabique, un disyllabique et un trisyllabique produit par le locuteur natif masculin dans chacune des trois modalités et utilisés dans les sessions d'entraînement (groupe B).

En outre, les enregistrements audio ont été segmentés en syllabes et en phones à l'aide de WebMAUS (Kisler et al., 2017). La segmentation obtenue a été corrigée manuellement, puis exportée au format TextGrid de Praat pour être affichée via l'interface Gepeto. La figure 1 montre des exemples de courbes f_0 pour chaque modalité du groupe B (énoncés utilisés dans les sessions d'entraînement) produite par le sujet natif, en fonction de la longueur en syllabes de l'énoncé.

L'analyse de f_0 de chaque énoncé produit par les apprenants lors du pré-test et du post-test a suivi la même procédure semi-automatique que celle décrite ci-dessus. La durée de chaque production a été extraite de la segmentation. Étant donné le nombre limité de sujets et d'exemples, l'analyse acoustique

se concentre principalement sur l'inspection visuelle de la forme du contour intonatif, considérée individuellement ou en moyenne sur les sujets, les conditions et les durées d'énonciation.

2.5 Prétest et Posttest

Des enregistrements ont été réalisés lorsque les sujets prononçaient les mots et les phrases pendant le pré-test et le post-test à l'aide d'un microphone d'ordinateur portable et du logiciel Audacity. Les instructions concernant les enregistrements étaient affichées sur des diapositives sur un écran d'ordinateur. Par exemple, pour le pré-test, les apprenants devaient produire le mot normalement comme s'il était à la fin d'une phrase ou question, en faisant attention à la ponctuation. Les modèles d'intonation des énoncés et des questions polaires étaient indiqués par des signes de ponctuation en rouge à la fin de chaque mot ou phrase. Les instructions ont été données en anglais, le niveau de français des apprenants débutants n'étant pas encore suffisant. Afin de fournir un contexte approprié pour l'intonation associée à l'incrédulité, une phrase d'introduction a été lue par l'expérimentateur pour chaque énoncé avant la production des stimuli par les sujets (exemple pour les stimuli « Lille ?! », « votre ami vous annonce qu'il déménage à Lille, à 900 km de Toulouse, et vous ne le croyez pas »).

2.6 Entraînement

Toutes les sessions de formation ont été supervisées par un professeur de langue et guidées par un jeu de diapositives d'instructions. Chaque session de formation utilisait un ensemble différent de 3 à 5 mots ou phrases. Les élèves ont participé aux séances de formation individuellement. Les séances d'entraînement commençaient par la lecture de chaque mot ou phrase avec les trois types d'intonation, suivie d'exercices utilisant l'interface Gepeto. Pour la tâche de lecture, les stimuli étaient présentés un par un sur une diapositive avec différents signes de ponctuation pour indiquer l'intonation appropriée (point pour les affirmations, point d'interrogation pour les questions polaires ou point d'interrogation et d'exclamation pour l'incrédulité). De plus, pour l'incrédulité, la même phrase de mise en contexte a été utilisée que dans les pré et post-tests.

Quatre types d'exercices ont été proposés avec Gepeto, sans limitation du nombre de répétitions, le guide visuel avec la production native de référence étant affiché pour les trois premiers :

1. Écouter l'enregistrement de référence tout en prêtant attention au guide visuel, puis imiter vocalement la référence.
2. Écouter la référence et la reproduire en synthèse vocale en traçant le guide visuel
3. Identique à la tâche 2 mais le geste tracé reste à l'écran et l'élève peut écouter son résultat.
4. Identique à la tâche 3, mais sans guide visuel.

Les participants écoutaient le modèle plusieurs fois avant d'essayer de reproduire les courbes. Seuls les exercices avec guides visuels ont été donnés aux élèves pour les sessions de formation 1 et 2. L'exercice sans guide a été ajouté pour les sessions de formation 3 et 4.

3 Résultats

3.1 Analyse acoustique des productions au prétest

La figure 2 montre les contours f_o moyennés et normalisés dans le temps des productions des 10 apprenants ukrainiens, pour chacun des énoncés produits pendant le pré-test (groupe A du corpus) et chacune des trois modalités. Les contours f_o correspondant aux productions des locuteurs natifs des

mêmes énoncés sont à droite de la figure. Les apprenants tendent à produire des contours intonatifs finaux ascendants pour les trois conditions dans le prétest, avec peu de distinction entre les modalités, bien que la variabilité entre locuteurs soit plus importante pour l'incrédulité (Figure 2). Inversement, alors que le modèle intonatif principalement produit par le locuteur natif pour exprimer l'incrédulité est proche de celui de la question sauf pour « Lille », ses énoncés déclaratifs sont systématiquement marqués par une chute intonative quel que soit le nombre de syllabes dans l'énoncé. L'incapacité apparente des apprenants ukrainiens débutants à produire la chute intonative caractéristique des énoncés déclaratifs en français peut expliquer les confusions fréquemment observées.

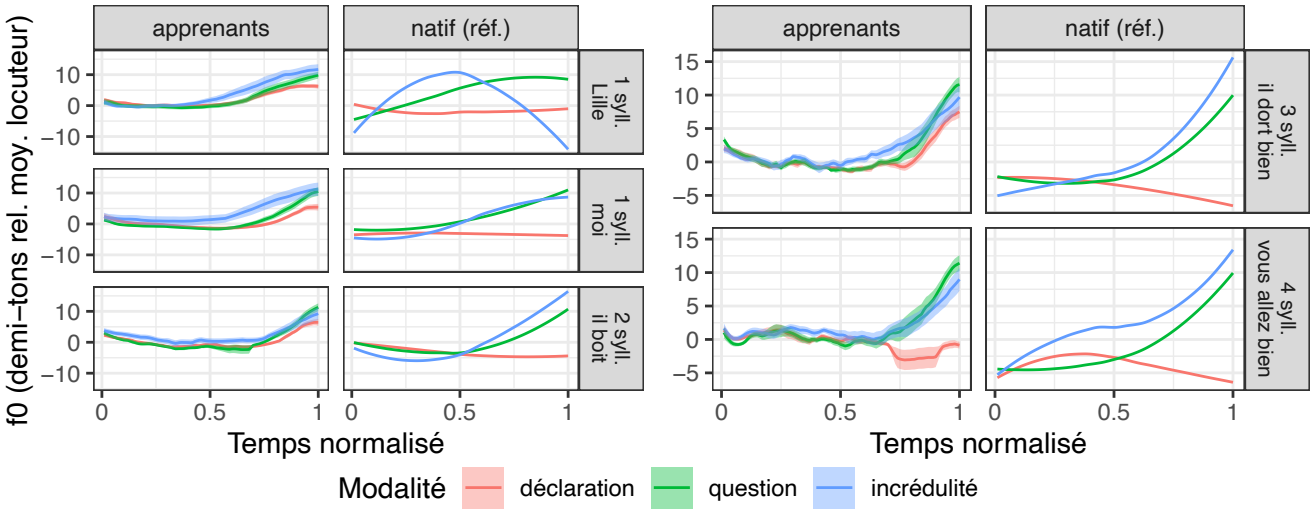


FIGURE 2 : Courbes de la production des cinq énoncés du groupe A par les 10 apprenants ukrainiens de la session de pré-test (à gauche, moyenne des apprenants), et par le locuteur natif masculin (à droite pour chaque longueur syllabique), pour chacune des trois modalités. Les enveloppes colorées autour des courbes moyennes des apprenants représentent l'erreur standard.

Une exception est l'énoncé de quatre syllabes « Vous allez bien », pour lequel les apprenants produisent en modalité déclarative des contours intonatifs plus proches de ceux du modèle natif, même s'ils présentent une trajectoire légèrement montante à la fin. Ces contours, produits avec peu de variabilité entre apprenants comme l'indique l'erreur standard modérée, est probablement influencé par les schémas prosodiques ukrainiens sur des énoncés plus longs.

3.2 Caractérisation acoustique des progrès des apprenantes

La figure 3 montre une comparaison entre le pré-test (groupe A) et le post-test (groupe C) des contours f_0 normalisés en fonction du temps dans les énoncés produits par les deux apprenantes ayant participé à l'ensemble du protocole, UF1 et UF2. Les contours f_0 des productions du locuteur natif (groupe A) sont reproduits à gauche. Les énoncés sont regroupés par nombre de syllabes pour la visualisation, qui comprend pour chaque locuteur et chaque condition deux énoncés monosyllabiques pour lesquels la variabilité est également représentée. Comme les énoncés de quatre syllabes n'étaient présents que dans les groupes A et B, seules les productions d'une à trois syllabes sont représentées.

La grande variabilité observée chez le locuteur natif pour les expressions d'incrédulité sur des énoncés monosyllabiques s'explique par les deux modèles intonatifs distincts produits par ce locuteur sur les énoncés « Lille » (Figure 2). Le schéma intonatif secondaire de montée et de descente se retrouve dans certaines productions d'incrédulité du post-test du locuteur UF1, qui semble avoir acquis la capacité de généraliser l'utilisation de ce schéma intonatif à des énoncés de longueur variable.

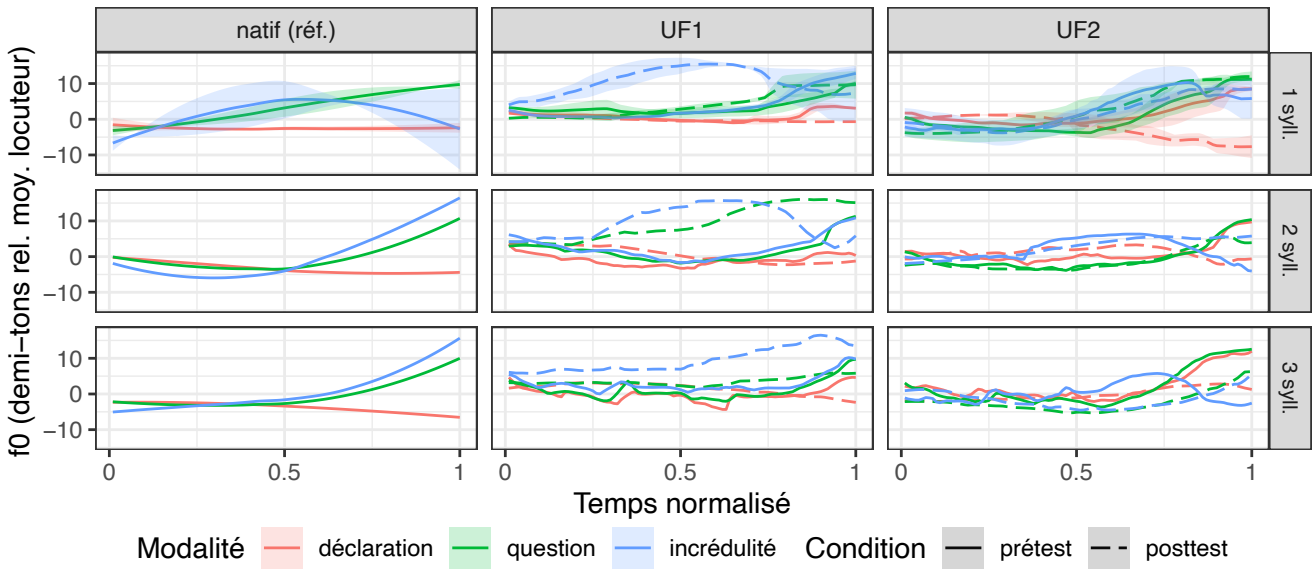


FIGURE 3 : Courbes de la production d'énoncés de 1 à 3 syllabes par les apprenants ukrainiens UF1 et UF2 qui ont participé au pré-test (énoncés du groupe A, trait plein) et au post-test (énoncés du groupe C, pointillés), pour chacune des trois modalités. Pour les énoncés monosyllabiques, des courbes moyennes sont présentées pour tenir compte des deux exemplaires inclus dans chaque condition, (enveloppes colorées : erreur standard).

En outre, les apprenants améliorent leur capacité à produire des énoncés déclaratifs et des questions de façon distincte et plus proche du natif, quel que soit le nombre de syllabes, d'où un écart plus important entre l'énoncé et la question dans la partie finale du même énoncé: l'écart moyen entre l'énoncé et la question dans le dernier quart de l'énoncé augmente entre le prétest et le post-test pour les énoncés d'une et de deux syllabes des deux apprenants (augmentation moyenne de cet écart de 9,1 demi-tons). Bien que plus modérée (3,4 demi-tons), cette augmentation de la distinction entre énoncé et question est également observée pour les énoncés de trois syllabes produits par UF1.

Enfin, la légère réduction par UF2 (-2,2 demi-tons) de l'écart entre déclaration et question (énoncés de trois syllabes), est principalement due à la baisse de f_o sur la syllabe médiane lors de la production de questions dans le post-test. De plus, la forme des contours de l'énoncé et de la question produits par UF2 en fin de l'énoncé est beaucoup plus proche du natif dans le post-test que dans le pré-test.

La normalisation temporelle des courbes de f_o pourrait masquer certaines distinctions entre modalités marquées par la durée et pas seulement par les variations intonatives. Si chez le natif, la distinction entre modalités n'est pas marquée par la durée, en pré-test les apprenants adoptent une stratégie de distinction par la durée (énoncés déclaratifs de 2 et 3 syllabes plus longs que les questions). En post-test, ces différences de durée sont réduites et la distinction entre les modalités s'améliore sur le plan mélodique. De plus, pour les trois modalités, quelle que soit la longueur des énoncés, les apprenants augmentent leur débit de parole entre pré- et post-test, pour les déclarations et les questions.

4 Discussion et conclusion

Malgré une notable réduction de la population due à des circonstances imprévues et indépendantes de notre volonté, l'étude a été utile et bénéfique pour tester la configuration expérimentale globale et la faisabilité de l'utilisation d'un système de synthèse vocale contrôlé par le geste. Elle a dévoilé plusieurs points inattendus liés à la production de l'intonation du français et à la prise en main de l'outil.

La montée finale observée pour certains des énoncés déclaratifs plurisyllabiques produits par UF2 pourrait s'expliquer par la présence éventuelle d'un accent de hauteur en ukrainien (Pompino-Marschall et al., 2017). L'intonation moyenne du pré-test pour les 10 sujets a montré un f_0 ascendant pour les trois types d'intonation en français, la question et l'incrédulité présentant une pente plus raide que pour la déclaration. Cependant, un schéma différent a été trouvé pour les énoncés déclaratifs de 4 syllabes, qui avaient un contour d'intonation plat dans le prétest, se terminant par une légère montée finale. Il existe donc des variations dans les modèles de production intonative en fonction de la longueur de l'énoncé. Par la suite, il serait souhaitable de tester l'outil sur des énoncés plus longs afin d'évaluer son utilité pour améliorer l'intonation de ces énoncés.

Dans le post-test, la principale amélioration s'est produite dans la production d'énoncés déclaratifs, qui ont montré un contour descendant plus fort par rapport au pré-test pour les deux apprenants. Dans l'ensemble, UF1 a réalisé des contours déclaratifs finaux plus proches de ceux du locuteur natif que UF2. UF1 semble également généraliser l'incrédulité mieux que UF2, et ses productions ont augmenté à la fin de la formation. Les différences entre les deux apprenants peuvent être en partie liées à la plus grande satisfaction d'UF1 vis-à-vis de l'expérience. De plus, UF1, plus jeune et ayant un meilleur niveau de français, a vécu plus longtemps en France que le sujet UF2. UF2 a indiqué qu'elle « parle parfois français » : on peut donc supposer que son exposition au français en dehors de la classe est limitée. Le sujet UF2 a également exprimé à plusieurs reprises sa gêne à utiliser la technologie liée à l'interface Gepeto au début des sessions de formation. Elle a été plus hésitante que UF1.

Cette étude est la première à tester une interface de contrôle gestuel dans un cadre autre que les conditions contrôlées d'un laboratoire. Parmi les difficultés rencontrées, citons les problèmes de stabilité de la connexion Internet dans la salle de classe ; la surface du téléphone en silicone, qui a entraîné une gêne pour l'un des participants qui avait les mains humides ; la latence dans l'ouverture du logiciel ; les problèmes de compatibilité entre les différents téléphones et navigateurs ; la qualité de la synthèse vocale, qui s'est parfois révélée peu familière aux participants ; et le fait que les participants n'étaient pas familiarisés avec les paramètres techniques de leur téléphone au début de l'entraînement. Dans les études futures, nous visons donc à fournir les appareils mobiles.

Malgré les différences de performance entre les deux apprenantes, toujours rencontrées en contexte de classe, et malgré le peu de sujets, cette étude exploratoire a montré qu'il est important d'enseigner l'intonation en français langue étrangère dès le début de l'apprentissage. Dans cette expérience, même s'il manque un groupe contrôle bénéficiant d'un enseignement intonatif plus « traditionnel » en comparaison avec notre pédagogie (prévu mais non réalisé en raison du blocage de l'université), l'interface Gepeto a aidé les apprenants à utiliser plus d'un modèle d'intonation en français et, surtout, à différencier les phrases déclaratives des questions polaires, une distinction qui, bien qu'essentielle à des fins de communication en français, n'est pas évidente au début pour les apprenants ukrainiens.

Remerciements

Cette recherche a été financée par l'ANR "GEPETO" (ANR-19-CE28-0018-01), et le LabEx EFL (ANR-10-LABX-0083) qui contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001). Nous remercions les enseignants et les apprenants qui ont rendu possible cette étude.

Références

BAILLS F., ALAZARD-GUIU C. & PRIETO P. (2022). Embodied prosodic training helps improve accentedness and suprasegmental accuracy. *Applied Linguistics* 43, 776–804. https://doi.org/10.1093/applin/amac010

BOERSMA P. & WEENINK D. (2023). Praat: doing phonetics by computer. Version 6.3.06, en ligne : http://www.praat.org/

CONSEIL DE L'EUROPE. (2001). *CECRL (Cadre européen commun de référence pour les langues)*. Strasbourg : Éditions Didier.

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*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110479645

FOUGERON, C. & L. SMITH C.L. (1993). Illustrations of the IPA: French. *Journal of the International Phonetic Association 23*, 73–76.

KISLER T., REICHEL U. & SCHIEL F. (2017). Multilingual Processing of Speech Via Web Services. *Computer Speech and Language* 45, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

LOCQUEVILLE G., D'ALESSANDRO C., DELALEZ S., DOVAL B. & XIAO X. (2020). Voks: Digital Instruments for chironomic control of voice samples. *Speech Communication* 125, 97–113. https://doi.org/10.1016/j.specom.2020.10.002

MENNEN I. (2015). Beyond segments: towards a L2 intonation learning theory. In: Delais-Roussarie, E., Avanzi, M. & Herment, S. (Eds.). *Prosody and Languages in Contact: L2 acquisition, attrition, languages in multilingual situations*, Springer Verlag, pp. 171–188.

POMPINO-MARSCHALL B., STERIOPOLO E. & ŻYGIS M. (2017). Ukrainian. *Journal of the International Phonetic Association* 47, 349–57. https://doi.org/10.1017/S0025100316000372

TANIGUCHI M. & ABBERTON E. (1999). Effect of interactive visual feedback on the improvement of English intonation of Japanese EFL learners. *Speach, Hearing, and Language: work in progress* 11, 77–89.

XIAO X., AUDIBERT N., LOCQUEVILLE G., D'ALESSANDRO C., KÜHNERT B., & PILLOT-LOISEAU C. (2021). Prosodic Disambiguation Using Chironomic Stylization of Intonation with Native and Non-Native Speakers. In *Proc Interspeech 2021*, pp. 516–520. https://doi.org/10.21437/Interspeech.2021-182

XIAO X., KÜHNERT B., AUDIBERT N., LOCQUEVILLE G., PILLOT-LOISEAU C., ZHANG H. & D'ALESSANDRO C. (2023). Performative Vocal Synthesis for Foreign Language Intonation Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp.1–9. https://doi.org/10.1145/3544548.3581210

YUAN C., GONZÁLEZ-FUENTE S., BAILLS F., & PRIETO P. (2019). Observing Pitch Gestures Favors the Learning of Spanish Intonation by Mandarin Speakers. *Studies in Second Language Acquisition* 41, 5–32. https://doi.org/10.1017/S0272263117000316

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Entraînement de la coordination respiration-parole en apprentissage de la lecture assistée par ordinateur

Delphine Charuau[1] Andrea Briglia[1] Erika Godde[2] Gérard Bailly[1]
(1) GIPSA-Lab, Univ. Grenoble-Alpes, 11, rue des Mathématiques 38400 St Martin d'Hères
(2) LEAD, Univ. Bourgogne Franche-Comté, 11, Esplanade Erasme 21000 Dijon
delphinecharuau1@gmail.com, gerard.bailly@gipsa-lab.fr,
erika.godde@u-bourgogne.fr

RÉSUMÉ

Cette étude vise d'une part, à identifier les indices respiratoires pouvant être considérés comme la signature de l'amélioration de la fluence, et d'autre part, à examiner les effets de l'entraînement de lecture assistée par ordinateur sur la progression de la coordination respiration/parole. 66 élèves (CE2-CM2) ont été répartis en trois groupes selon le mode d'entraînement suivi : contrôle, entraînement avec surlignage par mot et entraînement avec soulignage par groupe de souffle. Tous ont été enregistrés avant (pré-test) et après trois semaines d'entraînement de lecture assistée (post-test) lors de la lecture d'un texte entraîné et d'un autre non-entraîné. Les résultats indiquent que la planification respiratoire et la gestion des pauses est améliorée sur un texte entraîné. Toutefois, il n'y a pas de transfert significatif de ces améliorations sur le texte non-entraîné.

ABSTRACT

Breathing-speech coordination training in computer-assisted reading

This study aims, on one hand, to identify respiratory indicators that could be considered as the hallmark of fluency improvement, and on the other hand, to examine the effects of computer-assisted reading training on the progression of respiratory/speech coordination. 66 students (grades 3-5) were divided into three groups according to the training mode they followed : control, training with word highlighting, and training with breath group highlighting. All students were recorded before (pre-test) and after three weeks of computer-assisted reading training (post-test) while reading both a trained and an untrained text. The results of this study indicate that respiratory planning and pause management are improved on a trained text. However, there is no significant transfer of these improvements to the untrained text.

MOTS-CLÉS : Pauses, Coordination respiration-parole, Prosodie, Lecture assistée par ordinateur.

KEYWORDS: Pauses, Speech breathing coordination, Prosody, Computer-assisted reading.

1 Introduction

La fluence en lecture est traditionnellement mesurée comme le nombre de mots correctement lus par minute, faisant de la capacité de décodage et de la vitesse de parole les indicateurs-clés de la compétence en lecture (Breznitz, 2012). Or cette approche de la fluence fait totalement abstraction d'un aspect crucial de la lecture : la prosodie. La prosodie, qui englobe le rythme, le phrasé, les variations d'intonation, joue un rôle significatif dans la compréhension du texte. Des études ont montré que la prosodie en lecture, non seulement facilite la compréhension du texte par l'auditeur,

mais reflète également la bonne compréhension du texte par le lecteur lui-même, soulignant ainsi l'importance de la manière dont les mots sont lus (Rasinski, 2004; Schwanenflugel *et al.*, 2004).

Dans cette perspective, notre étude se concentre sur le développement de la coordination respiration-parole chez l'enfant, et à son impact sur le phrasé des enfants dans un contexte de lecture assistée par ordinateur basée sur le principe du karaoké. Cette méthode, combinant les bénéfices de la lecture chorale et de la lecture répétée, a déjà démontré son efficacité dans l'apprentissage d'une langue seconde (Luo *et al.*, 2008; Webb & Chang, 2012).

L'objectif de notre travail est double. Il s'agit, d'une part, d'analyser la coordination entre la respiration et la parole dans ce contexte particulier de la lecture assistée par ordinateur, afin d'identifier des indices respiratoires qui pourraient être considérés comme la signature d'une amélioration de la fluence, et d'autre part, d'examiner les effets d'un entraînement spécifique de la coordination respiration/parole par lecture assistée. En examinant la manière dont les enfants respirent pendant la lecture karaoké, et la manière dont cela influence leur phrasé, nous cherchons à comprendre les bénéfices de ce mode de lecture sur le développement des *patterns* respiratoires et leur contribution à l'amélioration de la compétence en lecture.

2 État de l'art

Lors de la lecture à haute voix, le contrôle de la respiration influe fortement sur le rythme et le phrasé.

La délimitation des unités du discours dépend du placement des pauses et de leur durée, au même titre que l'intonation ou l'allongement de certaines syllabes. En lecture, les prises de souffle sont majoritairement réalisées aux frontières des unités syntaxiques (Winkworth *et al.*, 1994). L'organisation des pauses dépend de la structure du texte : les pauses respiratoires sont plutôt réalisées à la fin des paragraphes, puis des phrases, tandis que les pauses non-respiratoires sont plutôt localisées aux frontières des syntagmes (Grosjean & Collins, 1979; Conrad *et al.*, 1983). Les textes fournissent des indices visuels favorisant la planification de la respiration et des pauses, notamment via les espaces et ponctuations. La lecture en ligne fait donc appel à la capacité des lecteurs à anticiper les signes de ponctuation comme indicateur de pauses, et à leur maîtrise de la structure syntaxico-sémantique du texte, pour la réalisation des pauses.

Compte tenu du lien étroit entre la respiration et le phrasé en lecture, il est important de prendre en considération le développement de ces aspects chez les apprenants. L'étude transversale récente menée par Godde *et al.* (2022) révèle une évolution significative dans la gestion des pauses pendant les premières années d'apprentissage de la lecture. Les résultats indiquent que les enfants de CE1 et CE2 ont un ratio plus élevé de pauses agrammaticales par rapport au nombre total de pauses, en comparaison à ceux des groupes de niveaux scolaires plus avancés (CE1 = 32% ; CE2 = 44% ; CM1-5^{e} = 20%). Cette proportion diminue progressivement à partir du CM1 jusqu'à la 5^{e}, bien qu'elle reste tout de même supérieure à celle observée chez les adultes (6%), suggérant un processus de développement continu. Ainsi, la maîtrise du placement des pauses semble être un aspect dynamique de l'apprentissage de la lecture, évoluant progressivement vers des modèles similaires à ceux des adultes au fil du temps. Ces résultats soulignent que, bien que la majorité des pauses chez l'enfant se situent aux frontières des unités syntaxiques (Lalain *et al.*, 2012), les enfants débutant l'apprentissage de la lecture produisent un nombre significatif de pauses agrammaticales, dues à des erreurs de décodage et à une absence de planification de la respiration et des pauses. Toutefois, avec l'amélioration des compétences en lecture, une diminution de ces pauses agrammaticales devrait être observée, reflétant ainsi une plus grande fluidité dans la lecture au fur et à mesure du développement de ces compétences.

Le placement des pauses est également déterminé par la coordination entre la respiration et la parole. Cette coordination se développe progressivement entre 4 et 10 ans (Hoit *et al.*, 1990; Boliek *et al.*, 2009), marquant une transition significative vers des modèles respiratoires plus matures vers l'âge de 7 à 8 ans (Hoit *et al.*, 1990). Les enfants de 7 ans et moins reprennent leur souffle à des intervalles plus fréquents que les enfants plus âgés, compensant ainsi de plus petites capacités respiratoires (Russell & Stathopoulos, 1988). Ils ont également recours à leur volume expiratoire de réserve pour conclure les groupes de souffle (Boliek *et al.*, 2009; Stathopoulos & Sapienza, 1997), ce qui suggère un manque de maturité de la planification de la respiration pour la parole. De plus, la dépense d'air par syllabe est particulièrement élevée chez les jeunes enfants, ce qui, combinée à un rythme de parole plus lent, conduit à un nombre réduit de syllabes par groupe de souffle. Ces caractéristiques ont un impact sur la fluidité de la lecture, car les interruptions fréquentes dans le flux de parole peuvent compromettre la fluence de la lecture. Toutefois, il a été montré que les enfants exploitent davantage les frontières entre les séquences syntaxiques pour reprendre leur souffle, là où les adultes produiraient une pause silencieuse démarcative (Charuau *et al.*, 2022). À mesure que les enfants grandissent, leurs stratégies respiratoires s'affinent progressivement, se rapprochant des modèles observés chez les adultes entre 10 et 14 ans (Hoit *et al.*, 1990). Ce développement de la coordination entre la respiration et la parole se produit en parallèle du développement des compétences en lecture, suggérant que certains paramètres respiratoires pourraient servir de solides indicateurs de la fluence.

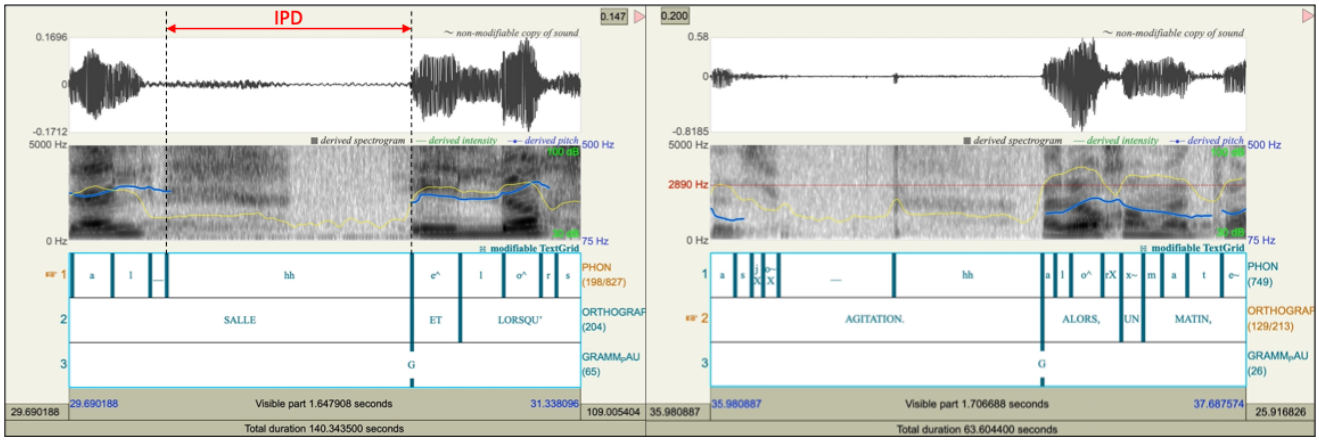


FIGURE 1 – Illustrations d'un IPD typique de l'enfant primo-lecteur (à gauche), et de celui d'un adulte (à droite).

L'un des indices de la coordination respiration/parole auquel nous portons une attention particulière est l'*inhalation-to-phonation delay* (IPD). Cette mesure, qui représente l'intervalle entre le début du bruit d'inspiration et le début de la phonation, évolue de manière significative au cours de la croissance. La pause respiratoire n'est pas entièrement dédiée à l'inspiration : cette dernière est précédée d'une phase de pré-inspiration, représentant environ 28% de la pause, et parfois suivie d'une phase post-inspiration (Grosjean & Collins, 1979). Chez l'adulte, il n'y a pas de post-inspiration : l'inspiration intervient généralement à la fin de la pause (Fig. 1), *juste à temps* avant de reprendre la phonation sans maintenir une pression sous-glottique élevée trop longtemps. L'IPD moyen d'un adulte est d'environ 400 ms (Godde *et al.*, 2022). À la différence des adultes, les enfants tendent à reprendre leur souffle juste après la fin de phonation, suggérant une respiration *en urgence*, se caractérisant par une pré-inspiration très courte, et une phase post-inspiratoire significativement longue. L'étude de Godde *et al.* (2022) montre que l'IPD est significativement plus élevé chez les élèves de CE1 (durée moyenne = 529 ms), par rapport aux adultes et aux enfants de CE2 à la 5^{e}. À partir du CE2, l'IPD décroît significativement et progressivement en fonction du niveau scolaire, jusqu'à se rapprocher

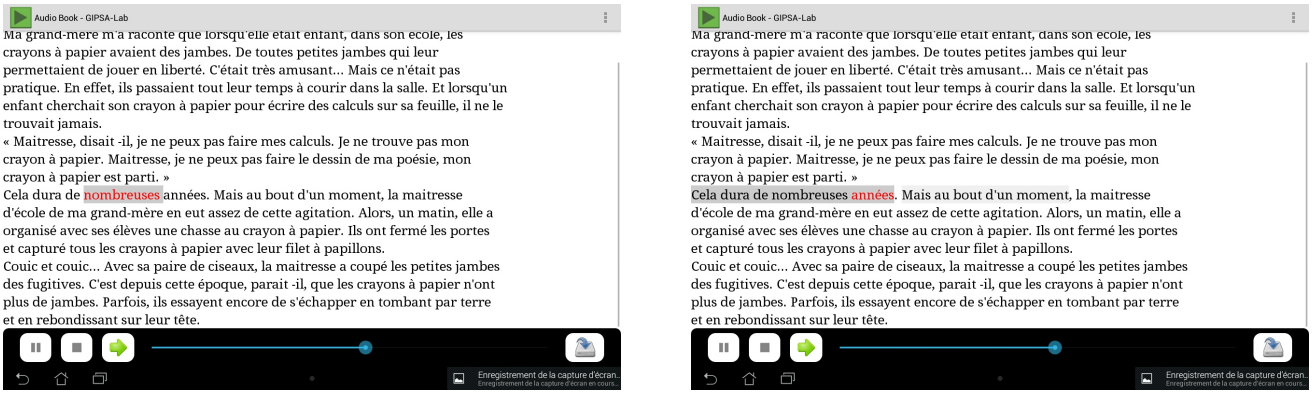


FIGURE 2 – Illustration de la lecture assistée par ordinateur avec un surlignage par mot (à gauche) et un surlignage par mot et groupes de souffle courant et suivant (à droite).

d'un IPD similaire à celui des adultes. L'un des objectifs de la lecture assistée par ordinateur vise à aider les enfants à planifier leur respiration, et à décaler la prise d'inspiration vers la fin de la pause.

3 Évaluation subjective de la fluence dans un contexte de lecture assistée par ordinateur (Godde *et al.*, 2019)

3.1 Lecture assistée par ordinateur

L'expérience a été menée à l'aide de l'application de lecture karaoké RAKE (Reading Assistance by KaraokE). Cette application offre un entraînement audio-visuel à la lecture, permettant à l'utilisateur de lire tout en écoutant (Bailly & Barbour, 2011). Équipé d'un casque, le locuteur entend un lecteur expert lire le texte affiché à l'écran, tout en voyant le texte mis simultanément en surbrillance. Deux niveaux de surbrillance sont étudiés ici : le mot seul *vs*. le mot et les groupes de souffle courant et suivant (Fig. 2). L'entraînement consiste à lire le texte à haute voix tout en se synchronisant avec la voix du lecteur expert.

L'apprenti lecteur bénéficie d'un entraînement visuel et auditif explicite, avec un surlignage progressif du texte pour guider la respiration et le placement des pauses et un modèle prosodique fourni par le lecteur expert. Le surlignage par groupe de souffle vise à aider les apprentis lecteurs à anticiper les prises de souffle et à les placer aux frontières syntaxiques. Le modèle donné par le lecteur expert permet la progression en termes de phrasé, d'expressivité et de vitesse, en imitant une distribution experte des pauses et une intonation appropriée.

Ce type d'entraînement favorise un apprentissage implicite grâce à une lecture chorale (*close-shadowing*), où l'apprenti suit de près le lecteur expert, avec un délai très court, typiquement une centaine de ms (Bailly, 2001). Ceci vise à un transfert implicite de compétences motrices, notamment en ce qui concerne la respiration. De plus, la lecture répétée et variée favorise l'amélioration de la fluence et l'apprentissage des compétences de lecture telles que la vitesse et la prosodie (Rasinski, 1990).

3.2 Protocole expérimental

97 élèves de classes de CE2, CM1 et CM2, provenant de 9 classes réparties dans 2 écoles élémentaires de l'agglomération de Grenoble, ont été sélectionnés pour participer à cette étude. À la suite de pré-tests, 8 élèves de chaque classe ont été sélectionnés selon leur profil de lecture. Ces élèves ont atteint un niveau de décodage satisfaisant et affichent tous une vitesse de lecture comprise entre 90 et 130 mots par minute. Leur prosodie a été évaluée selon l'échelle de fluence multidimensionnelle de Rasinski (2004), adaptée au français par Godde *et al.* (2021), avec un score de 2 sur 4 en vitesse, et inférieur à 3 en phrasé. Les élèves ayant manqué deux séances d'entraînement ou plus, ainsi que ceux ayant manqué la phase de pré- ou de post-test, ont été exclus de l'étude. Au total, 66 élèves (âge : 9 ans 3±17 mois) ont été inclus dans notre étude.

Ces 66 élèves ont été répartis de manière aléatoire en 3 groupes selon le mode d'entraînement suivi :

— Un groupe contrôle ne suivant pas d'entraînement avec une lecture assistée par ordinateur **(C)** ;
— Un groupe utilisant uniquement un surlignage des mots **(M)** ;
— Un groupe bénéficiant d'un surlignage par groupe de souffle **(S)**.

Chaque élève des groupes expérimentaux (M et S) participe à 9 séances de 20 minutes sur l'application RAKE en utilisant le type de surlignage correspondant à son groupe. Les séances d'entraînement se déroulent en petits groupes de 8 participants, pendant les heures de classe, sous la supervision d'un chercheur. Au cours de chaque séance, les élèves lisent à haute voix le même texte à 3 reprises. Le programme d'entraînement se déroule sur 3 semaines, avec 3 séances par semaine, et chaque semaine, un nouveau texte (A, B, C) est introduit. En pré-test et post-test, les enfants sont enregistrés sur 4 textes, 3 sur lesquels ils ont été entraînés (A, B, C), et un pour lequel ils ne l'ont pas été (D). Nos analyses portent sur la lecture d'un texte entraîné (A) et le texte non entraîné (D).

3.3 Résultats de l'évaluation subjective par l'échelle multidimensionnelle de fluence adaptée au français

L'évaluation subjective de la fluence a été menée par trois évaluateurs, tous chercheurs, qui ont respecté le protocole décrit dans l'article de Godde *et al.* (2019). L'évaluation a été réalisée durant la première minute d'écoute de la lecture. Quatre critères ont été évalués à l'aide de l'échelle multidimensionnelle de fluence adaptée au français (Godde *et al.*, 2021) : l'expressivité, le phrasé, le décodage et la vitesse.

Pour le texte A, le score total de fluence sur 16 est significativement amélioré entre les pré- et post-tests pour les groupes S et M (pre : M = 2.7±0.5 ; S = 2.7±0.5 ; post : M = 3.4±0.4 ; S = 3.6±0.5 ; $p < 0.001$), mais pas pour le groupe C, dont le score reste approximativement à 3 dans les deux phases (pre = 3±0.6 ; post = 3.2±0.5). En entrant dans le détail des dimensions, les résultats révèlent une augmentation significative des scores de phrasé pour les mêmes groupes (pre : M = 10±2 ; S = 9.7±1.5 ; post : M = 13.2±1.2 ; S = 13.7±1.6 ; $p < 0.001$).

Lors de la lecture du texte D, une progression significative du score de fluence total est observée pour tous les groupes, y compris le groupe contrôle (C : pre = 10.4±2.3 ; post = 12.1±1.8 ; $p < 0.05$; M : pre = 9.8±2.1 ; post = 11.8±1.8 ; $p < 0.01$; S : pre = 9.4±1.9 ; post = 12±1.3 ; $p < 0.001$). Pour ce même texte, le phrasé s'améliore entre les phases pré- et post-test pour les groupes S et M, passant respectivement d'un score de 2.6±0.6 à 3.1±0.3 ($p < 0.01$) et de 2.8±0.6 à 3.1±0.5 ($p < 0.05$). En

revanche, aucun effet n'est observé pour le groupe C, dont le score stagne autour de 3 (pre = 3 ± 0.5 ; post = 3.2 ± 0.4). Il convient de noter qu'une baisse de la variabilité est observée sur les groupes entraînés durant la lecture post-test. Enfin, la différence de niveau de surbrillance (M et S) ne présente d'effet significatif ni sur le phrasé, ni sur le score de fluence, aussi bien pour le texte entraîné que pour le texte non-entraîné.

Au regard de ces résultats, il apparaît que la lecture karaoké a un effet bénéfique sur la fluence, notamment sur le phrasé, aussi bien pour le texte A que pour le texte D. Ainsi, nous nous attendons à observer également une amélioration de certaines mesures respiratoires prises depuis le signal acoustique de parole entre les phases de pré- et de post-test.

4 Analyse des indices respiratoires

4.1 Traitement des données

Extraction des données. L'analyse objective porte sur les enregistrements pré-tests et post-tests des textes A et D. Pour chaque enregistrement, nous avons calculé le taux de phonation comme suit : durée totale de phonation/durée totale de lecture. Cette mesure rend compte du temps de phonation et de pause par rapport à la durée totale de lecture. Au sein des pauses respiratoires, nous avons procédé à la segmentation manuelle des IPD. Contrairement à l'étude présentée précédemment (Godde *et al.*, 2022), nous n'avons pas fait usage de ceintures respiratoires pour mesurer la variation des mouvements thoraciques et abdominaux. En l'absence de données respiratoires complémentaires, la délimitation des IPD repose exclusivement sur des caractéristiques acoustiques. L'étiquetage syntaxique des pauses a été réalisé de manière automatique, puis vérifié manuellement. Toute pause réalisée aux frontières des unités de rection et des séquences syntaxiques est considérée comme grammaticale (G). Les pauses localisées en dehors de ces frontières sont considérées comme agrammaticales (NG).

Analyses statistiques. Une analyse de variance est conduite suivie d'un test de Tukey-Kramer (Multcompare sous Matlab). Les tests significatifs sont indiqués avec les seuils de significativité habituels sur les figures le cas échéant.

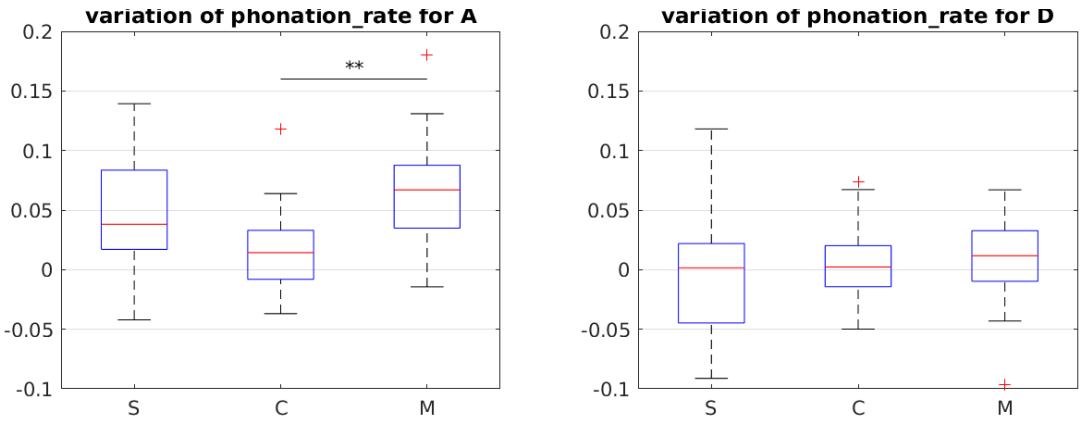


FIGURE 3 – Variation du taux de phonation entre pré- et post-test pour les textes A et D, selon les groupes de locuteurs. Une variation positive signe une proportion de phonation plus importante.

4.2 Résultats

Taux de phonation. Lors de la lecture du texte A, le taux de phonation s'améliore significativement chez les enfants du groupe M (Fig. 3). Si le groupe S suit la même tendance, la variation du taux de phonation n'est pas significative. En revanche, aucune évolution notable du taux de phonation n'est constatée lors de la lecture du texte D, quel que soit le groupe de locuteurs. Toutefois, notons la hausse de la variabilité pour le groupe S.

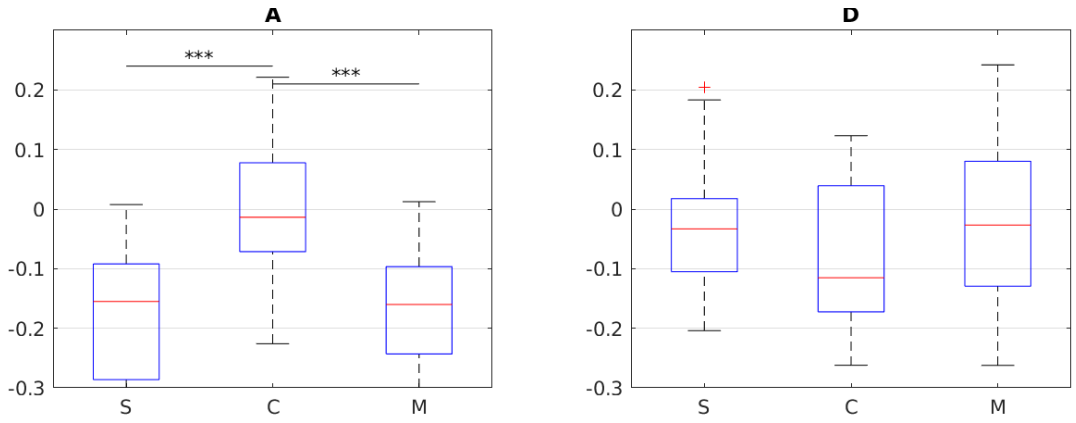


FIGURE 4 – Variation du pourcentage du nombre de pauses agrammaticales entre pré- et post-test pour les textes A et D, selon les groupes de locuteurs. Une moyenne négative signe une amélioration.

Ratio pauses grammaticales vs. non grammaticales. La figure 4 met en évidence les effets bénéfiques de l'entraînement en lecture karaoké sur le placement syntaxique des pauses. En effet, lors de la lecture du texte A, nous constatons une diminution significative du pourcentage de pauses agrammaticales pour les groupes S et M, par rapport au groupe contrôle C. Bien qu'une légère amélioration du placement des pauses soit observée pour les trois groupes lors de la lecture du texte D, aucune différence significative n'est relevée. Le mode de surbrillance du texte se semble pas non plus avoir d'impact sur la progression du placement des pauses.

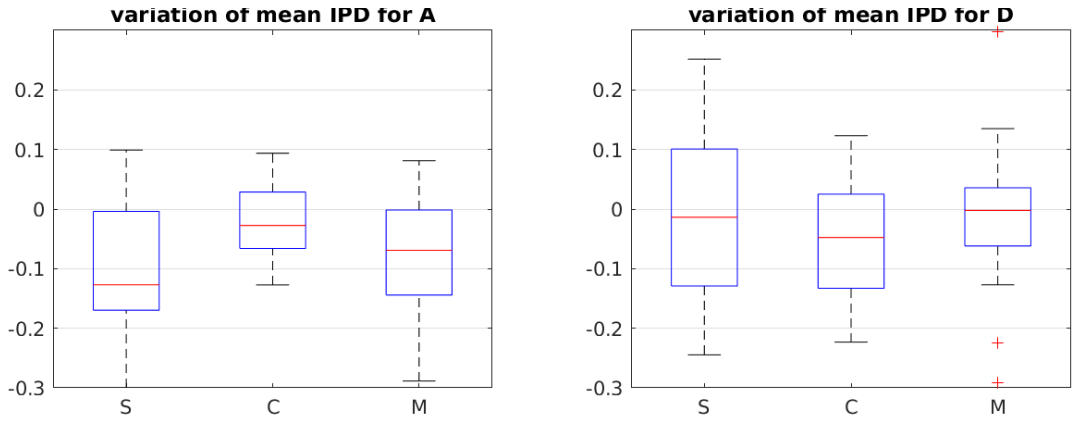


FIGURE 5 – Variation des IPD moyens entre pré- et post-test pour les textes A et D, selon les groupes de locuteurs. Une valeur négative signe une meilleure coordination respiration/parole.

IPD. Au regard des résultats présentés par la figure 5, nous observons, pour le texte A, une diminution des IPD chez les groupes S et M, avec un gain d'environ 100 ms pour le groupe S. Cependant, ces diminutions ne se distinguent pas significativement des valeurs du groupe C. Pour ce même texte, les

groupes S et M montrent une hausse de la variabilité de l'IPD par rapport au groupe C. À l'instar des précédents résultats pour le texte D, aucune amélioration significative de l'IPD n'est observée, quel que soit le groupe de locuteurs. Malgré l'absence d'effet sur l'IPD lors de la lecture du texte D, nous constatons une forte variabilité pour le groupe S, en comparaison avec les autres groupes.

5 Discussion

Cette étude portait sur l'identification d'indicateurs respiratoires pouvant témoigner d'une amélioration de la fluence en lecture, ainsi que sur l'observation des effets d'un entraînement spécifique de la coordination respiration/parole par la lecture assistée.

Nous constatons un impact positif de l'entraînement par lecture assistée sur la coordination respiration/parole, bien que cela ne se confirme que tendanciellement pour certains paramètres (IPD et taux de phonation). Alors que l'évaluation subjective effectuée sur le même corpus montrait une progression évidente du phrasé avec l'entraînement par lecture assistée (Godde *et al.*, 2019), nous nous attendions à retrouver davantage de significativité dans nos résultats. Ce décalage peut s'expliquer par la différence de durée d'analyse : l'évaluation subjective ne portait que sur la première minute de lecture, tandis que l'analyse des indices respiratoires couvrait l'ensemble de l'enregistrement. Au début de la lecture, une planification plus aisée est observée, suivie par des difficultés croissantes à planifier à mesure que la fatigue s'installe, ce qui peut nuire à la coordination respiration/parole.

Les résultats obtenus indiquent l'absence de transfert des améliorations de la coordination respiration/parole lors de la lecture du texte D. Cette absence de transfert pourrait être due à la complexité de ce texte, avec peu de ponctuations pour aider à identifier la structure syntaxique des phrases. De plus, la présence de mots rares et difficiles pour les enfants a conduit à un nombre conséquent d'erreurs de lecture et d'hésitations. Pour optimiser ce type d'entraînement, introduire davantage de marques de ponctuation pourrait réduire les difficultés lors de la lecture. À l'instar de ce qui a été observé lors de l'évaluation subjective de la fluence, les différents modes de surlignage ne semblent pas impacter la coordination respiration/parole. Le surlignage par groupe de souffle et mot était censé offrir des repères visuels pour aider à la planification de la respiration, mais la complexité de cette méthode, générant une surcharge cognitive pour les enfants, peut expliquer l'absence de résultats significatifs.

En somme, cette étude offre un aperçu encourageant des effets de l'entraînement par lecture assistée pour la planification de la respiration chez l'enfant, en particulier pour le placement des pauses. Bien que les améliorations de l'IPD et du taux de phonation ne soient pas statistiquement significatives, une forte variabilité a été observée chez les groupes expérimentaux, indiquant une progression chez certains lecteurs, mais pas chez d'autres. Des recherches supplémentaires pourraient aider à comprendre les raisons de ces différences de progression. Il est important de souligner que cette étude se base sur un entraînement de seulement trois semaines. Pour une progression plus significative de la coordination respiration/parole et un transfert efficace de ces compétences à des textes non-entraînés, un entraînement plus long pourrait être nécessaire. Dans le cadre de travaux futurs, cette étude sera étendue à un panel d'élèves plus important (≈ 1000). L'entraînement sera déployé sur une période de 10 semaines et portera sur la lecture de 35 textes plus homogènes.

Remerciements

Ce travail a bénéficié du soutien de l'ANR TRANS3 (ANR-22-FRAN-0008) et du projet Fluence (e-Fran, PIA2) financé par la CDC. Un grand merci aux élèves et à leurs enseignants.

Références

BAILLY G. (2001). Close shadowing natural vs. synthetic speech. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, p. 87–90, Perthshire - Scotland.

BAILLY G. & BARBOUR W.-S. (2011). Synchronous reading : learning french orthography by audiovisual training. In *Interspeech 2011-12th Annual Conference of the International Speech Communication Association*, p. 1153–1156.

BOLIEK C. A., HIXON T. J., WATSON P. J. & JONES P. B. (2009). Refinement of speech breathing in healthy 4- to 6-year-old children. *Journal of Speech, Language, and Hearing Research*, **52**(4), 990–1007. DOI : 10.1044/1092-4388(2009/07-0214).

BREZNITZ Z. (2012). *Fluency in Reading Synchronization of Processes*. Psychology Press, 2nd édition.

CHARUAU D., VAXELAIRE B. & SOCK R. (2022). L'organisation spatio-temporelle de la respiration chez l'enfant. *SHS Web of Conferences*, **138**, 08005. Publisher : EDP Sciences, DOI : 10.1051/shsconf/202213808005.

CONRAD B., THALACKER S. & SCHÖNLE P. (1983). Speech respiration as an indicator of integrative contextual processing. *Folia Phoniatrica*, **35**(5), 220–225. DOI : 10.1159/000265766.

GODDE E., BAILLY G. & BOSSE M.-L. (2019). Un Karaoké pour Entraîner Prosodie et Compréhension en Lecture. In *EIAH 2019 - Environnements Informatiques pour l'Apprentissage Humain*, Paris, France. HAL : hal-02141164.

GODDE E., BAILLY G. & BOSSE M.-L. (2022). Pausing and breathing while reading aloud : Development from 2nd to 7th grade in french speaking children. **35**(1), 1–27. DOI : 10.1007/s11145-021-10168-z.

GODDE E., BOSSE M.-L. & BAILLY G. (2021). Échelle Multi-Dimensionnelle de Fluence : nouvel outil d'évaluation de la fluence en lecture prenant en compte la prosodie, étalonné du CE1 à la 5ème. *L'année Psychologique/ Trends in Cognitive Psychology*, **121**(2), 19–43. DOI : 10.3917/anpsy1.212.0019, HAL : hal-02954060.

GROSJEAN F. & COLLINS M. (1979). Breathing, pausing and reading. *Phonetica*, **36**(2), 98–114.

HOIT J. D., HIXON T. J., WATSON P. J. & MORGAN W. J. (1990). Speech breathing in children and adolescents. *Journal of Speech, Language, and Hearing Research*, **33**(1), 51–69. DOI : 10.1044/jshr.3301.51.

LALAIN M., MENDONCA-ALVES L., ESPESSER R., GHIO A., LOOZE C. D. & REIS C. (2012). Lecture et prosodie chez l'enfant dyslexique, le cas des pauses. p. 41–48.

LUO D., MINEMATSU N., YAMAUCHI Y. & HIROSE K. (2008). Automatic assessment of language proficiency through shadowing. *2008 6th International Symposium on Chinese Spoken Language Processing*, p. 1–4.

RASINSKI T. V. (1990). Effects of repeated reading and listening-while-reading on reading fluency. *Journal of Educational Research*, **83**(3), 147–50. ERIC Number : EJ406326.

RASINSKI T. V. (2004). Assessing reading fluency.

RUSSELL N. K. & STATHOPOULOS E. (1988). Lung volume changes in children and adults during speech production. *Journal of Speech and Hearing Research*, **31**(2), 146–155. DOI : 10.1044/jshr.3102.146.

SCHWANENFLUGEL P. J., HAMILTON A. M., WISENBAKER J. M., KUHN M. R. & STAHL S. A. (2004). Becoming a fluent reader : Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, **96**(1), 119–129. DOI : 10.1037/0022-0663.96.1.119.

STATHOPOULOS E. T. & SAPIENZA C. M. (1997). Developmental changes in laryngeal and respiratory function with variations in sound pressure level. *Journal of Speech, Language, and Hearing Research*, **40**(3), 595–614. DOI : 10.1044/jslhr.4003.595.

WEBB S. & CHANG A. C.-S. (2012). Vocabulary learning through assisted and unassisted repeated reading. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, **68**, 267 – 290.

WINKWORTH A. L., DAVIS P. J., ELLIS E. & ADAMS R. D. (1994). Variability and consistency in speech breathing during reading. *Journal of Speech, Language, and Hearing Research*, **37**(3), 535–556. DOI : 10.1044/jshr.3703.535.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Erreurs de prononciation en L2 : comparaison de méthodes pour la détection et le diagnostic guidés par la didactique

Romain Contrain[1] Julien Pinquier[1] Lionel Fontan[2] Isabelle Ferrané[1]
(1) IRIT, 118 Route de Narbonne, F-31062 Toulouse CEDEX 9, France
(2) Archean Labs, 20 place Prax-Paris, 82000 Montauban, France
romain.contrain@irit.fr, julien.pinquier@irit.fr, lfontan@archean.tech, isabelle.ferrane@irit.fr

RÉSUMÉ

La détection et diagnostic d'erreurs de prononciation nécessite des systèmes adaptés aux spécificités de la parole non-native. Élaborer de tels systèmes reste difficile à cause de la rareté des corpus dédiés incluant des annotations expertes. Dans cet article, nous proposons et comparons deux approches, l'une basée sur une transcription phonétique et l'autre sur l'alignement de signaux audio, élaborées dans le but de servir dans un programme d'entraînement à la prononciation assisté par ordinateur (EPAO). Nous les évaluons sur un corpus de parole non-native annoté selon des considérations didactiques, et nous trouvons que l'approche basée sur l'alignement a des propriétés préférables pour l'EPAO, dépassant la précision de l'autre approche de 31,1 % et 3,8 % en absolu sur deux erreurs communes des apprenants japonais du français.

ABSTRACT

L2 mispronunciations : a comparison of didactics-guided detection and diagnosis methods

Mispronunciation detection and diagnosis requires systems that are adapted to the specificities of non-native speech. Developing such models remains challenging due to the scarcity of non-native speech corpora and expert annotations. In this work, we propose and compare two approaches, one based on phonetic transcription and the other based on audio-to-audio alignment, meant to be used in computer-assisted pronunciation training (CAPT) software. We evaluate them on a corpus of non-native speech that was annotated following didactic considerations, and find that the alignment-based approach has preferable properties for CAPT, surpassing the precision of the other approach by 31.1 % and 3.8 % absolute on two common mispronunciations of Japanese learners of French.

MOTS-CLÉS : Entraînement à la Prononciation Assisté par Ordinateur, détection et diagnostic d'erreurs de prononciation, parole non-native, apprentissage profond.

KEYWORDS: Computer-Assisted Pronunciation Training, Mispronunciation Detection and Diagnosis, non-native speech, deep learning.

1 Introduction

La majorité des apprenants n'ayant pas accès à un professeur particulier pour travailler leur compétence orale, les outils d'entraînement à la prononciation assisté par ordinateur (EPAO) offrent alors un support pédagogique intéressant. Il est nécessaire que ces outils soient capables de détecter et de diagnostiquer les erreurs de prononciation avec suffisamment de fiabilité et de précision pour pouvoir

fournir à l'apprenant des retours pertinents vis-à-vis des difficultés qu'il rencontre.

Dans ce domaine, nombre de travaux s'appuient sur un alignement forcé du signal de parole avec la prononciation canonique, ce qui permet d'évaluer les sons produits en connaissant les phones canoniques auxquels ils correspondent. Dans cette approche, la méthode la plus utilisé est le GOP (Witt & Young, 2000) employé par exemple par Laborde *et al.* (2016) qui utilisent des scores issus de la variante F-GOP (combiné à d'autres informations) dans une régression logistique. Des méthodes plus récentes emploient des classifieurs basés sur des réseaux de neurones profonds (DNN) ou des représentations comme wav2vec 2.0 (Baevski *et al.*, 2020). Dans Sancinetti *et al.* (2022), un DNN utilisé dans un pipeline GOP est *fine-tuné* pour produire directement des probabilités de mauvaise prononciation. Pour détecter les erreurs de prononciation, Xu *et al.* (2021) se basent sur des frontières obtenues par un alignement forcé avec l'énoncé cible et sur des représentations issues de wav2vec 2.0 fournies à un réseau convolutionnel (CNN). Ces travaux se limitent cependant à de la détection d'erreur sans fournir de diagnostic.

D'autres méthodes se basent sur une transcription phonétique suivie d'une comparaison avec la prononciation canonique, ce qui permet de fournir un diagnostic d'erreur. La méthode de comparaison peut être un alignement de Needleman-Wunsch (Needleman & Wunsch, 1970), comme dans Leung *et al.* (2019), mais n'est pas toujours spécifiée. Lin & Wang (2022) entraînent conjointement un modèle sur la détection d'erreurs de prononciation et la reconnaissance de phones (apprentissage multi-tâches) pour produire des transcriptions phonétiques associées à des probabilités de mauvaise prononciation. Bien que les auteurs choisissent de se concentrer sur la détection, une telle approche pourrait servir également au diagnostic. Wu *et al.* (2021) expérimentent avec deux architectures basées sur des *transformers* pour effectuer la phase de reconnaissance de phones, une utilisant comme fonction de coût l'entropie croisée et une autre utilisant le coût CTC (Connectionist Temporal Classification).

Les travaux décrits précédemment reposent sur des modèles phonétiques appris sur de la parole non-native, parfois en relativement grande quantité (environ 30 heures pour Wu *et al.* (2021)). Cependant, les corpus de parole non-native dédiés à la détection et au diagnostic d'erreurs restent rares et peu volumineux par rapport aux corpus de parole native. Ceci est dû à la rareté relative des apprenants de langue et de l'expertise nécessaire pour annoter cette parole au niveau phonétique. Récemment, Korzekwa *et al.* (2022) propose de palier ce manque de données en générant de nouveaux exemples de mauvaises prononciations à partir d'exemples existants, tandis que Xu *et al.* (2021) apprend les caractéristiques de la parole non-native de manière auto-supervisée sur des données non annotées avant d'entraîner son modèle de détection d'erreurs sur une quantité plus faible de données annotées.

Les erreurs faites par les non-natifs sont souvent dues à l'influence de leur langue maternelle (L1) sur la langue apprise (L2) (Detey & Racine, 2016). Par conséquent, certains travaux se concentrent sur une paire L1/L2 donnée comme Sancinetti *et al.* (2022) et Laborde *et al.* (2016), ce dernier se concentrant sur la paire japonais/français, spécifiquement sur les phonèmes /ʁ/ et /v/, difficiles pour les japonophones (Kamiyama *et al.*, 2016). Certains travaux incorporent même des connaissances sur les erreurs communes pour cette paire, comme Ghosh *et al.* (2017) ou Harrison *et al.* (2009) qui utilise un *extended recognition network* qui modélise les schémas d'erreur probables pour la paire cantonnais/anglais.

Dans cet article, nous présentons deux approches de détection et diagnostic d'erreurs de prononciation appliquées à la paire japonais/français. Après avoir décrit le corpus de productions orales d'apprenants japonophones du français à notre disposition, nous présentons les deux approches développées. L'une

se base sur un système de transcription phonétique, tandis que l'autre s'appuie sur une méthode d'alignement entre signaux pour s'affranchir de l'apprentissage de modèles phonétiques. Nous les évaluons sur un sous-ensemble de phonèmes et d'erreurs cibles dont le choix a été guidé par des connaissances didactiques spécifiques à la paire de langues considérée. Enfin, nous comparons et discutons les résultats de chaque approche dans la perspective de leur intégration dans un système d'EPAO proposant une tâche de répétition de stimuli.

2 Corpus d'apprenants et catégories didactiques

Dans le cadre du laboratoire commun ALAIA, un corpus regroupant des productions d'étudiants japonais apprenant le français a été constitué (ci-après APPR). Il a été collecté auprès de 67 apprenants dans le cadre de tâches de répétition. Chaque apprenant devait répéter plusieurs stimuli parmi les 199 possibles, les stimuli présentés ayant été prononcés par le même locuteur français natif. Chaque stimulus est composé d'un seul mot ou d'une courte phrase (1 à 6 syllabes) choisis pour faire travailler certains sons de la langue apprise. Ce corpus de français L2 totalise 7112 enregistrements.

Un expert en interphonologie japonais/français a transcrit phonétiquement chaque production et mis en correspondance chaque phone produit avec le phonème attendu. Un second expert a ensuite annoté les mêmes productions et aligné temporellement les transcriptions phonétiques avec le signal audio correspondant. Nous disposons ainsi des segments correspondant à la réalisation de 47541 phonèmes, dont 20 % sont des erreurs de prononciation. Les annotateurs sont d'accord sur 82 % des énoncés et sur 96,7 % des phonèmes. Plusieurs réalisations ont pu être identifiées comme ambiguës (environ 4500 réalisations touchant 3639 enregistrements), parce qu'un annotateur a hésité ou parce que les annotateurs sont en désaccord (par exemple sur un phone qui serait entre [y] et [ɯ]). En retirant ces réalisations ambiguës, il reste environ 43000 réalisations dont 17 % sont des erreurs de prononciation.

Les réalisations ont été regroupées suivant une approche guidée par la didactique. Chaque catégorie didactique, définie du point de vue de la perception d'un locuteur natif, correspond à un type de difficulté rencontrée par les apprenants qui appelle à un type d'exercice de remédiation en particulier. Certaines réalisations sont trop peu nombreuses et trop atypiques pour justifier la création d'une catégorie propre et sont considérées comme « autres ». Le français compte environ une douzaine de phonèmes difficiles pour les japonophones, source d'erreurs fréquentes et importantes pour leur intelligibilité (Kamiyama *et al.*, 2016), et qui sont donc ceux qui nous intéressent pour notre étude. Le travail de définition de ces catégories, mené par un expert de l'enseignement du FLE, a été réalisé pour deux phonèmes d'intérêt : la voyelle /y/ et la consonne /ʒ/. La table 1 présente les catégories définies pour ces deux phonèmes. Le corpus compte respectivement 1182 et 1540 réalisations non-ambiguës de/y/ et /ʒ/. Les catégories sont déséquilibrées en terme d'effectifs, avec par exemple 74 % des réalisations de /y/ qui sont correctes, contre seulement 13 % de « perçu comme [ø;œ] ».

3 Systèmes de détection et diagnostic d'erreurs de prononciation

Nous comparons deux approches : l'une repose sur une transcription phonétique de la production de l'apprenant, et l'autre sur un alignement des segments audio correspondant à la prononciation attendue et à sa réalisation. La production de l'apprenant est analysée en connaissant la transcription phonétique du stimulus à répéter (prononciation canonique), et le phone cible, celui sur lequel nous

Catégorie	Exemples de réalisations	Effectif
/y/ correct	[y], [yh]	869
/y/ perçu comme [ø;œ]	[ø], [œ], [ɯ], [ə], [əh]	151
/y/ perçu comme [j+Voyelle]	[jy], [jɯ], [ju], [jə], [je], [jø]	87
autre réalisation de /y/	[u], [a], [o], [ɔ], [i], délétion	75
/ʒ/ correct	[ʒ]	647
/ʒ/ perçu comme [dʒ]	[dʒ], [tʃ]	862
/ʒ/ perçu comme [ʃ]	[ʃ]	19
autre réalisation de /ʒ/	[d], [t]	12

TABLE 1 – Classement des réalisations en catégories didactiques

cherchons une erreur de prononciation. Le système prédit la catégorie didactique de la réalisation, qui peut être « correct » ou une des catégories d'erreurs spécifiques au phonème (voir table 1).

3.1 Approche basée transcription

Dans l'approche basée sur la transcription phonétique, un modèle de reconnaissance de phones fournit une transcription phonétique de la production de l'apprenant. Ensuite, un alignement de Needleman-Wunsch avec la prononciation canonique permet d'identifier la partie de la transcription qui correspond au phone cible. Enfin, nous recherchons quelle catégorie didactique concorde le plus avec ce qui a été transcrit. Pour cela, nous calculons une mesure de similarité entre le(s) phone(s) transcrit(s) et des réalisations typiques de chaque catégorie. Les deux dernières étapes utilisent une matrice de similarité entre les phones basée sur la matrice de distances proposée par Ghio *et al.* (2018). La figure 1 donne un exemple du fonctionnement de ce système dans le cas d'une répétition de « je chante » où nous nous intéressons au phonème /ʒ/.

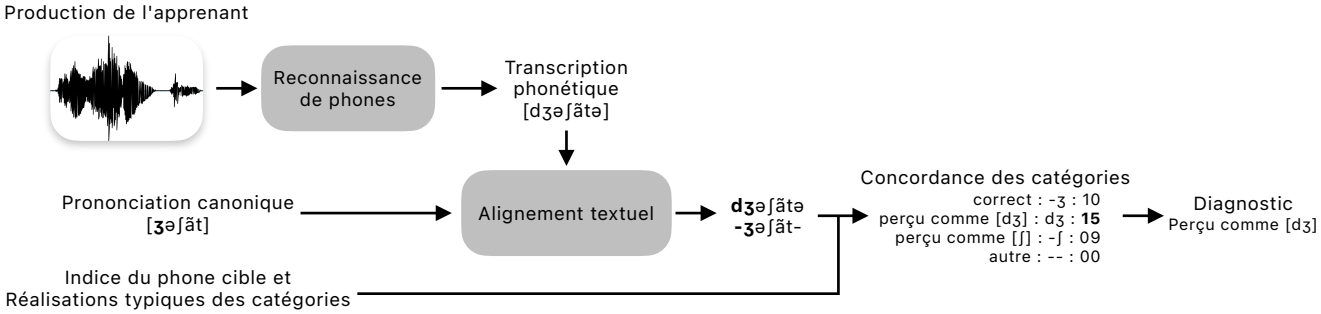


FIGURE 1 – Fonctionnement de l'approche basée sur la transcription phonétique

Le modèle de reconnaissance de phones employé se base sur Allosaurus [1], un modèle multilingue présenté par Li *et al.* (2020). Ce modèle utilise des unités BLSTM et une *loss* CTC. Il est conçu pour fonctionner avec n'importe quelle langue grâce à une étape de reconnaissance phonétique indépendante de la langue suivie d'une interprétation phonologique des phones transcrits. Cela permet de choisir n'importe quel jeu de phones pour la transcription. Pour l'inférence nous exploitons notre connaissance de l'énoncé à répéter pour restreindre le jeu de phones à ceux présents dans le mot et aux

1. https://github.com/xinjli/allosaurus v1.0

erreurs communes sur les phonèmes d'intérêt (s'ils sont présents). Par exemple, si la prononciation attendue est [ʒəʃãt], le jeu de phones sera restreint à [ʒ], [dʒ], [ʃ], [ə], [ã], [t].

Bien qu'il soit présenté comme universel, le modèle Allosaurus de base n'a jamais été entraîné sur du français et fonctionne mal sur cette langue : sur le corpus APPR il obtient un taux d'erreur phone (PER) d'environ 60 % (contre 25,0 % en moyenne sur ses langues d'entraînement) et il est notamment incapable de transcrire les voyelles nasales correctement. Nous avons donc dû pré-entraîner le modèle sur une quantité plus grande de parole native avant de l'adapter sur la petite quantité de parole non-native à notre disposition (voir section 2). Nous avons débuté avec un modèle monolingue pré-entraîné sur 150h de français natif issu du corpus Common Voice (ci-après Al-fr), puis nous avons expérimenté avec un modèle hybride pré-entraîné sur 150h de français et 65h de japonais natif issus du même corpus (ci-après Al-frjp). En effet, les apprenants ont tendance à produire des phones issus du système phonologique de leur L1 même s'ils sont absents de celui de la L2, et un modèle monolingue pourrait mal les gérer. Lors de l'adaptation à la parole non-native, chacun des modèles précédents a été entraîné sur le sous-ensemble des réalisations (issues d'APPR) correspondant aux phones qu'ils supportent (ces sous-ensembles sont nommés APPR-fr et APPR-frjp dans la section 4).

Common Voice (Ardila *et al.*, 2020) est un corpus multilingue de phrases lues, enregistrées par les contributeurs d'une plate-forme en ligne ouverte [2]. Pour le français, nous avons repris le sous-ensemble de 148,9 heures employé par Gelin *et al.* (2021) pour l'entraînement, et utilisé les 9,6 heures mises de côté par les auteurs comme ensemble de validation. Pour le japonais, 68,7 heures étaient exploitables dans la version 13.0. L'ensemble de validation a été construit pour représenter 6% du total en maximisant le nombre de locuteurs différents et en ayant le même ratio homme/femme que dans l'ensemble d'entraînement, aboutissant à 64,4 heures pour l'entraînement et 4,3 heures pour la validation. Les textes français ont été phonétisés à l'aide d'un dictionnaire de prononciation, et les japonais à l'aide de l'outil pykakasi [3] et de règles de prononciation.

3.2 Approche basée alignement

Un alignement temporel est réalisé entre le stimulus à répéter et de la production de l'apprenant au moyen de l'algorithme fastDTW (Salvador & Chan, 2007) et de représentations issues de wav2vec 2.0. Cela permet de faire correspondre la réalisation du phonème cible dans le stimulus, dont les frontières sont connues, avec sa réalisation dans la production. Ensuite, le segment correspondant est isolé, représenté avec un nouveau jeu de paramètres et des modèles de classification spécifiques au phonème cible déterminent la catégorie didactique de la réalisation. La figure 2 illustre ce principe de fonctionnement avec l'exemple d'une répétition de « le bus » où nous nous intéressons au phonème cible /y/.

Les deux étapes utilisent les mêmes représentations (wav2vec 2.0 [4]), mais la classification y ajoute des paramètres plus classiques, notamment 20 MFCC avec leurs dérivées premières et secondes, le Zero Crossing Rate et divers paramètres spectraux. Étant donné que la classification a besoin de données de dimension fixe, c'est la moyenne et l'écart-type sur le segment de chaque caractéristique qui sont fournis en entrée, ce qui donne une représentation à 252 dimensions.

La classification est réalisée par des *Random Forest* (Breiman, 2001) organisés en une architecture

2. https://commonvoice.mozilla.org
3. https://codeberg.org/miurahr/pykakasi
4. modèle pré-entraîné *wav2vec2-base-960h*

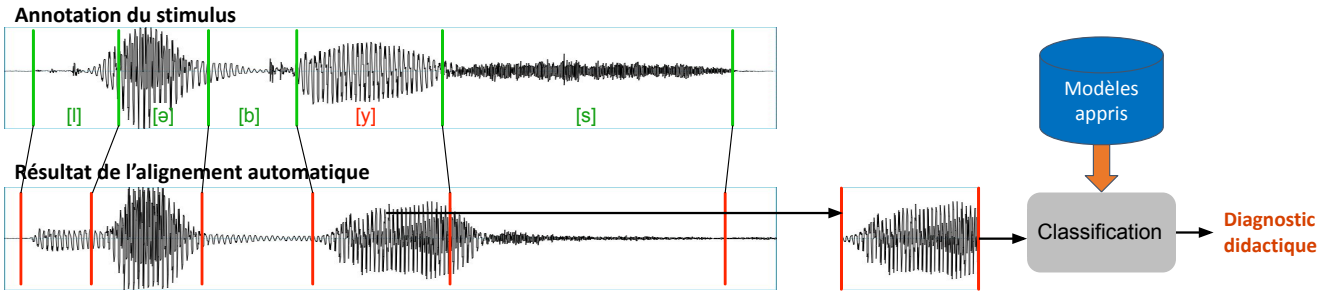


FIGURE 2 – Fonctionnement de l'approche basée alignement

hiérarchique binaire. Chacun est spécialisé dans l'appartenance à une catégorie didactique précise. Si le segment à classer n'appartient pas à la classe modélisée par le premier classifieur, alors nous regardons s'il appartient à la classe du deuxième, etc. jusqu'à atteindre le dernier niveau qui discrimine entre la dernière classe d'erreur et une classe « autre » utilisée comme une classe de rejet.

Les classifieurs pour chaque phonème d'intérêt ont été entraînés sur les réalisations de ce phonème dans le corpus APPR. Pour compenser le déséquilibre des classes, nous employons la technique de sur-échantillonnage SVM SMOTE (Tang *et al.*, 2008). À l'inférence, nous appliquons un seuil supplémentaire sur la probabilité renvoyée par les classifieurs (un seuil par classifieur), ce qui nous fournit un contrôle supplémentaire sur le compromis entre précision et rappel. Les seuils que nous utilisons ont été réglés par échantillonnage aléatoire pour maximiser la précision tout en maintenant le nombre d'exemples rejetés inférieur à 70%.

4 Méthode d'évaluation

Au vu de la petite taille des sous-ensembles utilisables pour l'entraînement, nous avons choisi de mettre en œuvre une validation croisée, en suivant un schéma *leave-one-speaker-out* pour mesurer l'adaptabilité des systèmes à de nouveaux locuteurs.

Comme les modèles de reconnaissance de phones utilisés dans l'approche transcription sont limités au jeu de phones qui a servi à leur pré-entraînement, nous les avons entraînés sur des sous-ensembles issus de notre corpus APPR : celui des énoncés ne comportant que des phones du français (APPR-fr) et celui des énoncés ne comportant que des phones du français ou du japonais (APPR-frjp). La table 2 donne des informations sur ces ensembles. Le nombre d'énoncés de APPR-frjp paraît petit par rapport aux 7100 énoncés du corpus, mais c'est parce que les 3600 énoncés contenant des réalisations ambiguës ne sont pas utilisables pour entraîner la reconnaissance de phones.

Sous-ensemble	Durée (h)	#énoncés	#phones
APPR-fr	1,22	2500	17000
APPR-frjp	1,51	3200	21000

TABLE 2 – Taille des données d'entraînement pour l'approche transcription

Les classifieurs utilisés dans l'approche alignement sont entraînés sur les réalisations de /y/ ou de /ʒ/ selon le classifieur. Les seuils des modèles hiérarchiques sont réglés après la validation croisée, à partir des probabilités prédites par les classifieurs pendant la phase de validation.

Nous analysons les résultats de nos systèmes en terme de performances sur la tâche de prédiction de la catégorie didactique. Comme nous sommes limités à deux phonèmes d'intérêt, /y/ et /ʒ/, nous pouvons regarder les performances classe par classe, que nous présentons en terme de précision et de rappel. Pour les systèmes d'EPAO, il est plus dommageable de marquer comme erronée une prononciation correcte que de manquer une prononciation incorrecte (Witt, 2012). Par rapport à nos métriques, cela veut dire que la précision sur les erreurs a plus d'importance que le rappel. Nous nous fixons un objectif de 85 % de précision pour dire si un système a des performances suffisantes pour une catégorie donnée. Par ailleurs, nous considérons qu'avoir des performances insuffisantes sur certaines catégories ayant peu de représentants n'est pas forcément dommageable. Nous ne nous intéressons plus aux catégories « autre » car elles n'ont pas d'intérêt didactique.

5 Résultats

Après avoir évalué les deux modèles de reconnaissance de phones au sein du système basé transcription, et le système basé alignement, nous avons obtenu les résultats consignés dans la table 3. Ils nous informent tout d'abord que l'approche alignement atteint des scores de précision en général plus élevés que l'approche transcription (par exemple 92,7% contre 83,3% sur « /ʒ/ correct »), mais des rappels plus faibles (23,5 % sur cette même classe).

Métrique →	Précision (%)			Rappel (%)		
Approche →	Transcription		Alignement	Transcription		Alignement
Modèle →	Al-fr	Al-frjp		Al-fr	Al-frjp	
/y/ correct	88,5	87,9	**95,0**	**74,1**	73,5	32,5
/y/ perçu comme [ø;œ]	27,3	30,2	**61,3**	39,7	**49,0**	12,6
/y/ perçu comme [j+V]	59,6	52,4	**84,6**	**32,2**	25,3	12,8
/ʒ/ correct	61,1	83,3	**92,7**	**81,5**	75,7	23,5
/ʒ/ perçu comme [dʒ]	83,3	84,4	**88,2**	60,6	**87,6**	31,2
/ʒ/ perçu comme [ʃ]	**25,0**	21,6	0,0	**47,4**	42,1	0,0

TABLE 3 – Résultats des deux approches sur les différentes catégories didactiques

Pour l'approche transcription, les différences entre les deux modèles sont notables. Sur le phonème /ʒ/, l'apprentissage hybride fait diminuer le nombre de confusions de [dʒ] pour [ʒ], ce qui se traduit par des gains absolus de 22,2% de précision sur « /ʒ/ correct » et de 27,0% de rappel sur « /ʒ/ perçu comme [dʒ] ». Sur le phonème /y/, « /y/ perçu comme [ø;œ] » gagne en précision et en rappel (+2,9% et +9,3%) tandis que « /y/ perçu comme [j+Voyelle] » perd en performances (-7,2% et -6,9%), de même que « /y/ correct » dans une moindre mesure, du fait d'une tendance à plus prédire « /y/ perçu comme [ø;œ] ».

L'approche alignement a, sur une majorité de classes, une précision plus élevée et un rappel plus faible que l'approche transcription. Pour les classes où c'est le cas, l'écart absolu de précision va de +3,8% sur « /ʒ/ perçu comme [dʒ] » à +31,1% sur « /y/ perçu comme [ø;œ] », tandis que l'écart de rappel va de -19,4% sur « /y/ perçu comme [j+Voyelle] » à -58,0% sur « /ʒ/ correct ».

Le nombre de représentants a un certain impact sur les résultats : toutes les approches testées ont de meilleurs résultats sur les catégories majoritaires (/y/ et /ʒ/ bien prononcés, « /ʒ/ perçu comme [dʒ] ») que sur les catégories avec moins d'exemples (« /y/ perçu comme [ø;œ] » par exemple).

6 Discussion

Notre objectif de précision de 85 % est atteint pour le phonème /ʒ/ avec le système basé sur l'alignement : 92,7 % et 88,2 % respectivement sur « correct » et « perçu comme [dʒ] ». Le système basé sur la transcription avec Al-frjp est tout juste en dessous de l'objectif (83,3 % et 84,4 %). Prédire avec suffisamment de précision le diagnostic « perçu comme [ʃ] », qui ne représente de toute façon que 1,2 % des réalisations, semble hors de portée des approches évaluées. Pour le phonème /y/, dépasser l'objectif ailleurs que sur la classe « correct », majoritaire, semble difficile. Le système basé sur l'alignement l'atteint presque sur « perçu comme [j+V] » avec 84,6 % de précision, mais de manière surprenante les résultats sont moins bons sur « perçu comme [ø;œ] » (61,3 % de précision), alors que ce diagnostic a 74 % de représentants en plus que « perçu comme [j+V] ».

Notre choix de régler les seuils de nos classifieurs afin de maximiser la précision se voit bien dans les résultats de l'approche alignement. En effet, nous rejetons beaucoup d'exemples (d'où un faible rappel). Si nous déployons ce système dans un programme d'EPAO, un apprenant devra donc réaliser plusieurs exercices avant d'obtenir un diagnostic juste, voire obtenir un diagnostic tout court. C'est particulièrement gênant pour /y/ où le rappel sur les erreurs avoisine $^1/_8$.

Nos résultats montrent que l'apprentissage hybride permet au modèle de reconnaissance de phones de s'améliorer sur la tâche de diagnostic d'erreurs de prononciation. La baisse des confusions entre [ʒ] et [dʒ] s'explique par le fait que le corpus de japonais contient autant d'exemples de ces deux sons. De même, l'intégration du phone [ɯ] du japonais, qui représente une bonne partie des réalisations de /y/ perçues comme [ø;œ], permet de mieux détecter cette catégorie. Dans les deux cas, cette amélioration s'accompagne cependant d'une augmentation des confusions entre les autres classes et la classe qui s'améliore, amoindrissant le gain de précision et diminuant le rappel des autres classes. Si le diagnostic d'erreur s'améliore nettement sur /ʒ/, l'amélioration est moins sensible sur /y/.

7 Conclusions

Cet article compare deux approches de détection et diagnostic d'erreurs de prononciation, l'une basée sur une transcription phonétique et l'autre sur un alignement de signaux audio. Nous les évaluons sur un corpus d'apprenants japonais du français dont les annotations sont guidées par des connaissances didactiques, ce qui rend les diagnostics pertinents pour l'apprentissage des langues. Nous trouvons que l'approche basée sur l'alignement est plus adaptée pour l'EPAO, avec des précisions plus élevées que l'approche basée sur la transcription, imputables au contrôle qu'elle fournit sur le compromis précision/rappel.

Les travaux présentés ici se focalisent sur deux phonèmes particuliers pour une paire L1/L2 donnée, mais la méthodologie mise en œuvre pourrait aisément être appliquée à d'autres phonèmes ou paires de langues, pourvu qu'un corpus de parole non-native intégrant des connaissances didactiques soit disponible. Aussi, il serait intéressant de mener plus d'expérimentations dans ce sens.

Remerciements

Ces travaux ont été financés par l'ANR dans le cadre du LabCom ALAIA (ANR-18-LVC3-001).

Références

ARDILA R., BRANSON M., DAVIS K., HENRETTY M., KOHLER M., MEYER J., MORAIS R., SAUNDERS L., TYERS F. M. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 4211–4215.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 12449–12460 : Curran Associates, Inc.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**, 5–32.

DETEY S. & RACINE I. (2016). L'apprentissage de la prononciation d'une langue étrangère : le cas du français. In S. DETEY, J. EYCHENNE, Y. KAWAGUCHI & I. RACINE, Éds., *La prononciation du français dans le monde : du natif à l'apprenant*, chapitre 14, p. 84–96. Paris : CLE International.

GELIN L., DANIEL M., PINQUIER J. & PELLEGRINI T. (2021). End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, **134**, 71–84.

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 *XXXIIe Journées d'Etudes sur la Parole*, p. 285–293 : ISCA.

GHOSH S., FAUTH C., LAPRIE Y. & SINI A. (2017). End-to-End Acoustic Feedback in Language Learning for Correcting Devoiced French Final-Fricatives. In *Proc. Interspeech 2017*, p. 349–353. DOI : 10.21437/Interspeech.2017-1031.

HARRISON A. M., LO W.-K., QIAN X.-J. & MENG H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *Proc. Speech and Language Technology in Education (SLaTE 2009)*, p. 45–48.

KAMIYAMA T., DETEY S. & KAWAGUCHI Y. (2016). Les japonophones. In S. DETEY, J. EYCHENNE, Y. KAWAGUCHI & I. RACINE, Éds., *La prononciation du français dans le monde : du natif à l'apprenant*, chapitre 24, p. 155–161. Paris : CLE International.

KORZEKWA D., LORENZO-TRUEBA J., DRUGMAN T. & KOSTEK B. (2022). Computer-assisted pronunciation training—speech synthesis is almost all you need. *Speech Communication*, **142**, 22–33. DOI : https ://doi.org/10.1016/j.specom.2022.06.003.

LABORDE V., PELLEGRINI T., FONTAN L., MAUCLAIR J., SAHRAOUI H. & FARINAS J. (2016). Pronunciation Assessment of Japanese Learners of French with GOP Scores and Phonetic Information. In *Proc. Interspeech 2016*, p. 2686–2690. DOI : 10.21437/Interspeech.2016-513.

LEUNG W.-K., LIU X. & MENG H. (2019). Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8132–8136. DOI : 10.1109/ICASSP.2019.8682654.

LI X., DALMIA S., LI J., LEE M., LITTELL P., YAO J., ANASTASOPOULOS A., MORTENSEN D. R., NEUBIG G., BLACK A. W. *et al.* (2020). Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8249–8253 : IEEE.

LIN B. & WANG L. (2022). Phoneme mispronunciation detection by jointly learning to align. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6822–6826. DOI : 10.1109/ICASSP43922.2022.9746727.

NEEDLEMAN S. B. & WUNSCH C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453. DOI : https ://doi.org/10.1016/0022-2836(70)90057-4.

SALVADOR S. & CHAN P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, **11**(5), 561–580.

SANCINETTI M., VIDAL J., BONOMI C. & FERRER L. (2022). A transfer learning approach for pronunciation scoring. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6812–6816. DOI : 10.1109/ICASSP43922.2022.9747727.

TANG Y., ZHANG Y.-Q., CHAWLA N. V. & KRASSER S. (2008). Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**(1), 281–288.

WITT S. M. (2012). Automatic error detection in pronunciation training : Where we are and where we need to go. In *International Symposium on automatic detection on errors in pronunciation training*, p. 1–8.

WITT S. M. & YOUNG S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, **30**(2-3), 95–108.

WU M., LI K., LEUNG W.-K. & MENG H. (2021). Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2021*, p. 3954–3958. DOI : 10.21437/Interspeech.2021-1467.

XU X., KANG Y., CAO S., LIN B. & MA L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. In *Proc. Interspeech 2021*, p. 4428–4432. DOI : 10.21437/Interspeech.2021-777.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude IRM de la production des /l/ de l'anglais par des locuteurs francophones

Alice Léger[1] Coline Caillol[1] Emmanuel Ferragne[1] Hannah King[1] Sylvain Charron[2] Clément Debacker[2, 3] Maliesse Lui[2, 3] Catherine Oppenheim[2, 3]
(1) Université Paris Cité, CLILLAC-ARP, F-75013 Paris, France
(2) Université Paris Cité, Inserm, Institute of Psychiatry and Neurosciences of Paris, F-75014 Paris, France
(3) GHU-Paris Psychiatrie et Neurosciences, Hôpital Sainte-Anne, F-75014 Paris, France
alice.leger@etu.u-paris.fr, coline.caillol@etu.u-paris.fr

RÉSUMÉ

Cette étude analyse l'articulation des allophones clairs et sombres du /l/ de l'anglais par trois locuteurs francophones et une locutrice native d'anglais britannique. Nous examinons en imagerie par résonance magnétique si les apprenants développent un /l/ sombre (absent du français), avec plus de rétraction en coda qu'en attaque comme attendu en anglais standard. Nous mesurons également si les apprenants acquièrent la corrélation observée chez les natifs entre longueur de la rime et degré de rétraction du /l/ sombre. L'effet de l'antériorité théorique et empirique de la voyelle est aussi analysé. Nos résultats indiquent que les participants ont acquis la distribution allophonique attendue avec une influence de l'antériorité, mais pas de la longueur de la voyelle. Notre étude contribue ainsi à caractériser les gestes articulatoires complexes acquis par des apprenants avancés de l'anglais à travers une technique d'imagerie permettant de visualiser l'intégralité des zones articulatoires pertinentes pour le /l/ de l'anglais.

ABSTRACT

An MRI study of the production of English /l/ by French speakers.

This study investigates the articulation of the dark and light allophones of /l/ in English by three French native speakers and one native speaker of British English. Using magnetic resonance imaging, we examine whether the learners develop a dark /l/ (absent from French), with more retraction in coda than in onset, as observed in standard Englishes. We also measure whether learners acquire the correlation found in native speakers between the length of the rhyme and the degree of retraction of the dark /l/. The effect of theoretical and empirical vowel frontness is also considered. Our results indicate that the participants acquired the expected allophonic distribution with an influence of frontness but not vowel length. Our study thus contributes to the characterisation of complex articulatory gestures acquired by advanced learners of English using an imaging technique which allows us to visualise the entire tongue areas relevant to English /l/.

MOTS-CLÉS : /l/ sombre, coarticulation, IRM, anglais L2.

KEYWORDS: dark /l/, coarticulation, MRI, L2 English.

1 Introduction

L'acquisition de la prononciation d'une langue étrangère implique souvent la production de gestes articulatoires qui sont absents de la langue maternelle. C'est par exemple le cas du /l/ anglais pour des apprenants francophones. En effet, les variétés d'anglais standard présentent un allophone plus « sombre » du /l/ en coda de syllabe, et un allophone plus « clair » en attaque (Wells, 1982). Le /l/ sombre se distingue du /l/ clair par sa double articulation : il implique, en plus d'une élévation de l'apex commune avec le /l/ clair, une rétraction du dos de la langue. Traditionnellement qualifiée de « vélarisation », cette rétraction peut s'étendre de la région uvulaire à pharyngale (Browman & Goldstein, 1995; Narayanan *et al.*, 1997) ; plus elle est importante, plus le /l/ est dit « sombre ». De plus, l'articulation acquise doit être conforme à la variété d'anglais parlée. En effet, les allophones du /l/ produits en attaque et en coda se distinguent davantage par leur degré de rétraction de la langue que par la simple présence ou absence de rétraction. Par exemple, le /l/ de l'anglais américain implique une rétraction de la langue quelle que soit la position dans la syllabe (Sproat & Fujimura, 1993), mais qui reste plus importante en coda qu'en attaque (Proctor *et al.*, 2019). La distinction est plus marquée en anglais britannique standard, avec une rétraction de la langue en coda uniquement (Bladon & Al-Bamerni, 1976; Turton, 2017). L'acquisition de cette allophonie reste toutefois peu étudiée chez les apprenants francophones (à l'exception de Colantoni *et al.* (2023) et King & Ferragne (2015)). Dans le cadre d'expériences pilotes pour l'étude de la prononciation d'apprenants d'anglais langue seconde (L2), la présente étude analyse l'articulation du /l/ de l'anglais par des locuteurs avancés d'anglais L2 ayant le français pour langue maternelle (L1).

Nous examinons plus particulièrement un aspect du /l/ de l'anglais observé à l'origine chez des locuteurs natifs par Sproat & Fujimura (1993) : une corrélation entre la longueur de la rime dans la syllabe et le degré de rétraction de la langue. Puisqu'on dispose de plus de temps pour déplacer le dos de la langue dans le contexte d'une rime longue, ils proposent que le /l/ sombre est produit avec une rétraction plus importante dans les rimes longues que les rimes courtes (par ex. dans *peel* [piːɬ] par rapport à *pill* [pɪɬ]). Une corrélation entre rétraction et longueur de la rime a également été observée par Yuan & Liberman (2011). Dans leur étude acoustique de 20 000 tokens de /l/ du corpus anglais américain SCOTUS (Supreme Court Justice of the United States corpus), ils observent un effet dans le contexte du /l/ sombre, mais pas du /l/ clair. Turton (2017) trouve un résultat similaire auprès de locuteurs de cinq variétés d'anglais parlées au Royaume-Uni. Ses données échographiques indiquent un effet de la durée de la rime sur les /l/ sombres dans les variétés montrant un /l/ clair en attaque et sombre en coda. Il est intéressant de noter que dans la variété parlée à Manchester, où seul un variant sombre existe, la rétraction du /l/ en coda est plus importante en fin d'énoncé qu'en cours d'énoncé (par ex. dans *peel* par rapport à *peel bananas*). Comme l'explique Turton (2017), cette position finale implique une rime plus longue et donne au geste dorsal plus de temps pour atteindre sa constriction maximale. L'influence de la longueur de la rime sur le degré de rétraction du /l/ sombre n'a à notre connaissance pas été examinée auprès d'apprenants de l'anglais. Cette étude vise à déterminer si des locuteurs francophones d'anglais L2 développent naturellement ce schéma articulatoire.

Nous analysons la production de trois apprenants avancés de français L1-anglais L2 ayant un accent américain ou britannique, et d'une locutrice native d'anglais britannique. Nous examinons si les apprenants développent un allophone sombre du /l/ absent dans leur L1, ainsi qu'une articulation cohérente avec leur variété d'anglais acquise. Nous attendons une rétraction du /l/ en coda quel que soit l'accent, et du /l/ en attaque uniquement en anglais américain. Dans un deuxième temps, nous étudions si les apprenants développent un effet de la longueur de la rime sur le degré de rétraction du /l/. D'après les résultats de Sproat & Fujimura (1993), nous prédisons que le /l/ sombre sera produit

avec une rétraction du dos de la langue plus importante après une voyelle longue qu'après une voyelle courte. Les contextes vocaliques inclus dans notre étude étant plus variés que ceux de Sproat & Fujimura (1993) (/iː, ɪ/), nous considérons également l'influence de l'antériorité de la voyelle sur le degré de rétraction du /l/ sombre. Nous prédisons un effet de coarticulation du /l/ sombre en coda avec la voyelle précédente, c'est-à-dire une rétraction du /l/ moins importante dans le contexte d'une voyelle antérieure que d'une voyelle postérieure.

Les données analysées sont issues d'enregistrements pilote pour l'étude de la prononciation des apprenants en imagerie par résonance magnétique à haute résolution temporelle (IRM-HRT). Parmi les méthodes qui permettent d'observer la coordination des articulateurs pour la production de la parole, nous disposons de l'échographie de la langue (ex. Léger *et al.*, 2023) et de l'IRM-HRT. Ces 15 dernières années, l'IRM-HRT a permis l'étude détaillée des structures physiologiques impliquées dans la production de la parole (Kochetov, 2020; Lim *et al.*, 2021; Narayanan *et al.*, 2004; Carey & McGettigan, 2017). Elle permet notamment d'imager ce qui reste invisible par échographie : la surface complète de la langue, sa face inférieure et les structures pertinentes autres que la langue telles que les lèvres, le palais dur, le voile du palais ou encore le pharynx. En effet, alors que l'apex et la racine de la langue sont essentiels à l'étude de la double articulation du /l/ sombre, ils sont invisibles à l'échographie en raison de l'ombre projetée par la mandibule et l'os hyoïde à l'avant et, à l'arrière, du conduit vocal (comme illustré par la Figure 1). La visualisation offerte par IRM-HRT nous permet ainsi une description détaillée des stratégies articulatoires des apprenants pour le /l/ sombre de l'anglais.

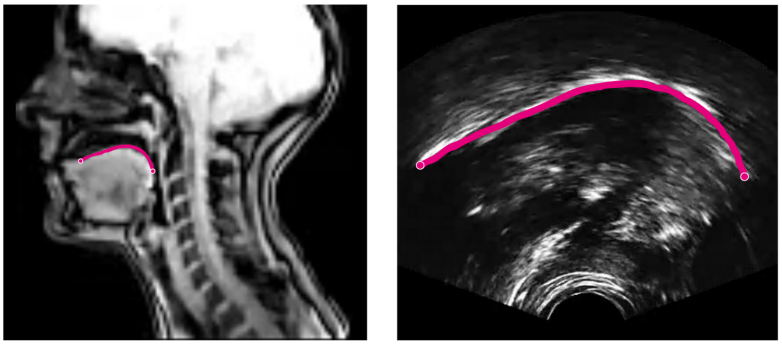


FIGURE 1 – Comparaison de la visibilité de la langue pour la locutrice FFB avec l'IRM (gauche) et l'échographie (droite). Le contour tracé manuellement en rose recouvre approximativement la même zone dans les deux images

2 Méthode

2.1 Participants

Les participants sont quatre locuteurs et locutrices de l'anglais, tous enseignants-chercheurs en phonétique. Le groupe se compose d'une locutrice native d'anglais britannique (AFB) et de trois locuteurs de langue maternelle française présentant un niveau avancé d'anglais. Parmi ces trois participants, on compte une locutrice d'anglais américain (FFA) et deux locuteurs d'anglais britannique (FFB, FHB). Les participants ont été nommés selon le code suivant : 1ère lettre, langue maternelle (F : français, A :

anglais) ; 2e lettre, genre (F : femme ; H : homme) ; 3e lettre, accent (A : américain ; B : britannique). À noter que les trois locuteurs d'anglais L2 de cette étude ont acquis l'anglais dans des contextes différents. La locutrice FFA a vécu et étudié aux États-Unis de ses 10 à ses 15 ans. Les locuteurs FFB et FHB ont eux acquis l'anglais en France dans un cadre scolaire, sans immersion dans une communauté de locuteurs natifs avant leur majorité.

2.2 Matériel linguistique

Chaque participant a lu une liste de 18 monosyllabes contenant un /l/ en attaque (ex. *leap*) ou en coda (ex. *peel*), précédant ou suivant une monophtongue de l'anglais. La Table 1 présente les stimuli utilisés pour chaque position syllabique et chaque contexte vocalique. Nous faisons référence à chaque contexte vocalique par son ensemble lexical (Wells, 1982), c'est-à-dire le « mot-clé » représentant la voyelle cible (ex. LOT pour /ɒ, ɑ ː/). Une seconde liste de mots avec ces mêmes voyelles en contexte /b/ + voyelle (+ consonne /t, d, k/) a été produite pour obtenir l'articulation de chaque voyelle par chaque locuteur dans un contexte indépendant du /l/. Trois répétitions de la liste des /l/ ont été enregistrées par participant, et une seule de la liste des voyelles, en raison de contraintes de temps.

Contexte vocalique		**Mot test**	
Mot-clé	**BR / AM**	**/l/ Attaque**	**/l/ Coda**
FLEECE	**iː**	leap	peel
KIT	**ɪ**	lip	pill
DRESS	**e**	let	tell
TRAP	**æ**	lap	pal
LOT	**ɒ / ɑː**	lock	col
STRUT	**ʌ**	luck	cull
THOUGHT	**ɔː / ɑː**	law	all
GOOSE	**uː**	loop	pool
FOOT	**ʊ**	look	pull

TABLE 1 – Liste des mots test par contexte vocalique, selon la réalisation attendue par accent d'anglais (BR = britannique, AM = américain)

2.3 Acquisition des données

Les données ont été acquises au moyen d'une IRM Vantage Galan 3T XGO de Canon Medical Systems. Une coupe sagittale médiane de l'appareil phonatoire a été acquise avec une séquence 2D en écho de gradient rapide avec les paramètres suivants : TR = 2,8 ms, TE = 1,2 ms, BW = 78,25 Hz, FOV = 24×24 cm, angle de bascule = 5°. La résolution dans le plan était de $2,5 \times 2,5$ mm pour une épaisseur de coupe de 10 mm. Les images ont été acquises avec une résolution temporelle de 10 images par seconde par l'intermédiaire d'une antenne tête/cou de 16 canaux combinée avec une antenne flexible de 16 canaux placée au-dessus de la bouche et du cou. Les données IRM ont été générées par une reconstruction en deep learning incluant un débruitage des images (Advanced intelligent Clear-IQ Engine, Canon Medical Systems). Les données audio ont été capturées à l'aide du microphone FOMRI III+ d'Optoacoustics et débruitées avec le logiciel iZotope RX [1].

1. Voir un exemple des vidéos que nous obtenons, avec la locutrice AFB ici : https://www.youtube.com/shorts/wxSpz_NY0us

2.4 Analyses

Pour chacune des trois occurrences de chaque mot test comportant un /l/ en attaque ou en coda, les contours de la langue ont été tracés manuellement par le premier auteur à trois instants différents : la constriction maximale du geste antérieur, la constriction maximale du geste postérieur et l'articulation de la voyelle. La trame correspondant à chacune de ces constrictions a été choisie manuellement, en examinant trame par trame les images IRM des mots test produits. Pour l'unique occurrence de chaque mot de type /b/ + voyelle (+ consonne), le contour de la langue a été tracé à la trame montrant l'articulation cible de la voyelle, avant la transition vers la consonne suivante ou la position de repos. Les deux points aux extrémités de chaque contour ont été positionnés de façon systématique, pour la partie antérieure, au point de la face inférieure de la langue se trouvant à hauteur de la mandibule et, pour la partie postérieure, au dernier point de jonction identifiable entre la racine de la langue et l'épiglotte. Afin de corriger les différences interindividuelles d'inclinaison de la tête, une droite des moindres carrés a été ajustée au contour moyen de chaque participant. Puis l'angle entre cette droite et l'axe des x a été employé pour opérer une rotation de tous les contours de chaque participant. Ensuite, les contours ont été recalés pour chaque participant par rapport à la valeur de x minimale pour ce participant (c.-à-d. la valeur la plus antérieure de la pointe de la langue) et par rapport à la valeur de y minimale. Puisque dans le cas du /l/ sombre, nous nous intéressons à une éventuelle élévation du dos de la langue vers le voile du palais ou à une constriction pharyngale, nos variables dépendantes sont les coordonnées du point le plus haut de la langue dans sa moitié postérieure ainsi que la valeur de x maximale pour chaque contour (reflétant le degré de constriction postérieure). Nous avons également pris une mesure plus globale de la hauteur et de l'antériorité de la langue en calculant le barycentre du polygone formé par les contours tracés.

3 Résultats

Afin de déterminer si les locuteurs d'anglais L2 de nos données parviennent à produire des allophones du /l/ distincts en attaque et en coda de syllabe, nous avons calculé un modèle linéaire mixte avec la position (attaque vs coda) et le locuteur comme facteurs fixes, la répétition et le mot comme effet aléatoire, et la valeur horizontale des barycentres comme variable dépendante. On trouve un effet significatif du locuteur ($F_{(3,211)} = 120,77 \quad p < 0,001$) et un effet significatif de la position ($F_{(1,211)} = 15,375 \quad p < 0,001$) dans le sens prédit (position plus postérieure du /l/ en coda).

La comparaison visuelle des trames de constriction postérieure maximale du /l/ révèle incidemment des schémas articulatoires individuels sensibles à la variété d'anglais parlée. La Figure 2, qui montre l'articulation du /l/ pour *leap* et *peel*, reflète les tendances d'articulation observées sur l'ensemble des mots test. On voit chez les locuteurs d'anglais britannique (AFB, FFB, FHB) une différence nette entre le /l/ en attaque et en coda, marquée par l'absence de constriction du dos de la langue en attaque. En revanche, la locutrice d'anglais américain (FFA) présente en attaque une rétraction du dos de la langue plus importante que les trois autres locuteurs. Cette réalisation sombre du /l/ semble visuellement plus antériorisée que le /l/ sombre qu'elle produit en coda, conformément à la tendance observée chez les locuteurs natifs d'anglais américain.

Pour tester l'effet de la longueur phonologique sur l'articulation du geste postérieur du /l/ en coda, cinq modèles linéaires mixtes (un pour chacune de nos variables dépendantes) ont été calculés, avec pour facteur fixe la longueur théorique de la voyelle (longue vs courte) et pour facteurs aléatoires le

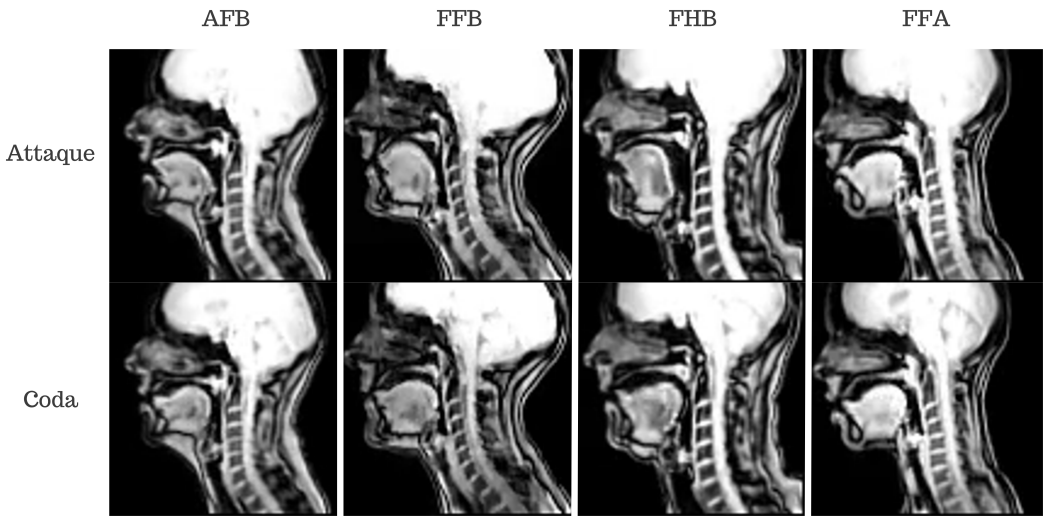


FIGURE 2 – Rétraction maximale du /l/ en attaque et en coda dans une répétition de *leap* et *peel* par chaque locuteur

locuteur, la répétition, et la longueur emboîtée dans le facteur mot. La Table 2 résume les modèles testés. On n'observe d'effet significatif de la longueur phonologique sur aucune des cinq variables. Nos résultats suggèrent que la longueur de la voyelle précédant le /l/ sombre n'impacte pas le degré de rétraction ou d'élévation de la langue.

Variable dépendante	**Statistiques**
$maxX$	$F_{(1,106)} = 1,3769 \quad p = 0,24326$
$maxYSecondHalf$	$F_{(1,106)} = 1,955 \quad p = 0,16497$
$xOfMaxYSecondHalf$	$F_{(1,106)} = 0,3969 \quad p = 0,53016$
$xCentroid$	$F_{(1,106)} = 0,31609 \quad p = 0,57516$
$yCentroid$	$F_{(1,106)} = 4,0097 \quad p = 0,52795$

TABLE 2 – Résumé des modèles testant l'effet de la longueur phonologique. Le code des effets aléatoires dans chaque modèle est le suivant : $(1|spkr) + (1|rep) + (1 + vLength|word)$

Variable dépendante	**Statistiques**
$maxX$	$F_{(1,106)} = 1,9422 \quad p = 0,16635$
$maxYSecondHalf$	$F_{(1,106)} = 1,6392 \quad p = 0,20323$
$xOfMaxYSecondHalf$	$F_{(1,106)} = 0,39083 \quad p = 0,53321$
$xCentroid$	$F_{(1,106)} = 44,707 \quad p < 0,001$
$yCentroid$	$F_{(1,106)} = 1,6801 \quad p = 0,19773$

TABLE 3 – Résumé des modèles testant l'effet du caractère antérieur ou postérieur (théorique) de la voyelle. Le code des effets aléatoires dans chaque modèle est le suivant : $(1|spkr) + (1|rep) + (1 + frontness|word)$

Nous avons ensuite testé (Table 3) l'effet du caractère antérieur ou postérieur de la voyelle sur nos cinq variables dépendantes en commençant par une répartition « théorique » des voyelles : FLEECE,

KIT, DRESS et TRAP ont été classées comme antérieures ; LOT, FOOT, GOOSE, THOUGHT et STRUT, comme postérieures. Seule la valeur x du centre de gravité de la langue montre un effet significatif : la voyelle postérieure entraîne une valeur de x plus élevée, reflétant le caractère rétracté de la langue pour le /l/ suivant une voyelle postérieure. La hauteur de la langue pour le /l/ ne présente cependant pas d'influence de l'antériorité de la voyelle qui précède.

Afin de comparer l'effet de l'antériorité théorique à celui de l'antériorité empirique, nous avons déterminé quelles voyelles étaient phonétiquement antérieures ou postérieures en fonction du locuteur. La Figure 3 montre pour chaque locuteur le point le plus haut de la langue lors de la production des neuf voyelles issues des mots-tests. Pour la classification empirique des voyelles comme antérieure ou postérieure, une séparation a été appliquée pour chaque locuteur à partir de la valeur sur l'axe horizontal qui est à mi-chemin entre le point de la voyelle la plus postérieure et celui de la voyelle la plus antérieure.

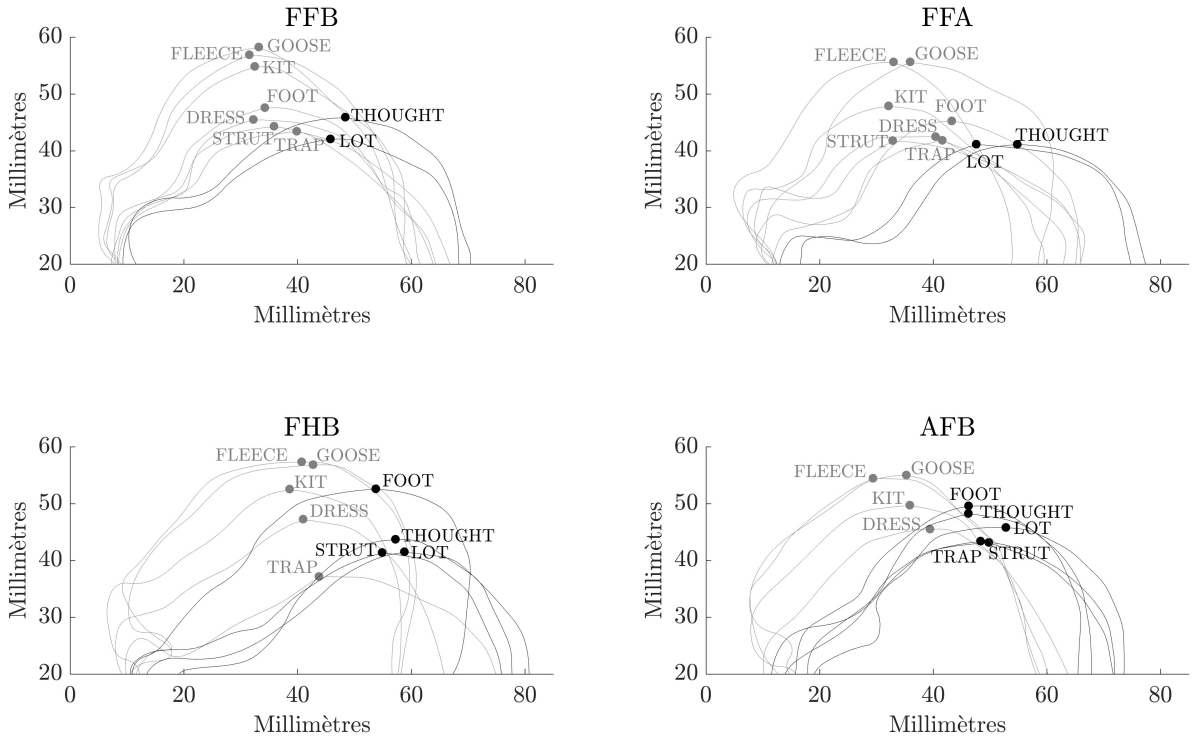


FIGURE 3 – Caractère antérieur (gris) ou postérieur (noir) de la voyelle déterminé par le point le plus haut de la langue pour chaque participant. Code participants : 1ère lettre, langue maternelle (F : français, A : anglais) ; 2e lettre, genre (F : femme ; H : homme) ; 3e lettre, accent (A : américain ; B : britannique)

Le seul modèle qui présentait un effet significatif dans la Table 3 était celui qui avait pour variable dépendante la valeur x du centre de gravité de la langue. En prenant le même modèle, mais en remplaçant le caractère antérieur-postérieur théorique par celui que nous venons de déterminer empiriquement grâce au point le plus haut de la langue, le nouveau modèle renvoie un même effet significatif du caractère antérieur-postérieur ($F_{(1,106)} = 9,96854 \quad p < 0,005$). Le meilleur modèle est celui qui a pour facteur le caractère antérieur-postérieur théorique ($AIC = 423, BIC = 444$; comparé au modèle avec antériorité empirique : $AIC = 436, BIC = 457$).

4 Discussion

À la question de savoir si des locuteurs non-natifs de l'anglais peuvent acquérir un geste articulatoire absent de leur L1 et n'ayant qu'un rôle allophonique dans la L2, nos résultats indiquent que oui ; et cela, y compris pour des apprenants tardifs. En dépit des limites imposées par une supposée période critique pour l'acquisition de la phonologie d'une L2 (Birdsong, 2018), une certaine plasticité reste disponible à l'adolescence, y compris pour des détails phonétiques dont l'absence n'entraverait pas la compréhension, comme le /l/ sombre de notre étude ou le VOT du français par des anglophones dans l'étude de Birdsong (2003). Les /l/ produits en coda montrent également un effet de coarticulation selon l'antériorité de la voyelle qui précède : le geste dorsal du /l/ sombre est antériorisé si la voyelle est antérieure, et sa rétraction plus marquée si la voyelle est postérieure. Les locuteurs d'anglais L2 de cette étude ne présentent cependant pas le même effet de la longueur phonologique de la voyelle sur le geste postérieur du /l/ que celui rapporté chez les natifs (Sproat & Fujimura, 1993). Toutefois, nos résultats sont à généraliser avec prudence : les locuteurs de notre étude sont des apprenants avancés, qui plus est spécialistes de phonétique de l'anglais. Examiner ces mêmes effets de coarticulation auprès d'apprenants de niveau intermédiaire aurait un intérêt pédagogique certain pour l'enseignement de la prononciation en anglais L2. Dans une étude acoustique, Chung & Kim (2021) trouvent par exemple que /ɔl, ɑl/ sont des contextes qui favorisent l'acquisition du /l/ sombre pour les apprenants coréen L1-anglais L2. Dans le même ordre d'idée, en laissant plus de temps pour la réalisation du /l/ sombre, une rime longue pourrait également constituer un contexte de production plus accessible par les apprenants de l'anglais. Nous avons utilisé comme facteur la dichotomie « théorique » entre voyelles phonologiquement longues et brèves en anglais britannique ; un prolongement logique de cet aspect consistera à examiner le signal acoustique pour calculer des durées phonétiques précises des voyelles et mesurer leur impact sur l'articulation du /l/. Notre analyse pourra aussi bénéficier d'une description plus précise du caractère antérieur-postérieur des voyelles. La meilleure performance du modèle utilisant l'antériorité théorique plutôt qu'empirique suggère que le point le plus haut de la langue, pourtant utilisé comme critère discriminant pour les voyelles dans les manuels de phonétique (par ex. Roach (2007)), est un piètre descripteur. Par exemple, dans la Figure 3 pour la locutrice FFA, la proximité du point le plus haut pour GOOSE et FLEECE occulte la configuration de la langue plus postérieure pour GOOSE que FLEECE. Utiliser par exemple le centre de gravité saurait mieux tenir compte de cette différence.

5 Conclusion

Notre étude visait à analyser l'articulation du /l/ sombre de l'anglais chez des locuteurs avancés français L1-anglais L2. Nous avons confirmé que nos participants présentaient l'alternance allophonique attendue : /l/ clair en attaque de syllabe, et sombre en coda. Nous avons ensuite déterminé que l'articulation du /l/ sombre de l'anglais était influencée par l'antériorité de la voyelle précédente, mais pas par la longueur de la rime, contrairement au schéma observé chez les natifs. Nos résultats indiquent une influence de l'antériorité de la voyelle sur le /l/ en coda, avec un geste dorsal plus marqué après une voyelle postérieure. En autorisant la visualisation de l'intégralité de la langue tout en rendant possible l'analyse de la succession des gestes articulatoires du /l/ anglais, l'IRM-HRT nous a permis de caractériser en détail la réalisation d'un son complexe de l'anglais par une locutrice anglophone et trois francophones locuteurs avancés d'anglais L2.

Références

BIRDSONG D. (2003). Authenticité de prononciation en français L2 chez des apprenants tardifs anglophones : Analyses segmentales et globales. *Acquisition et interaction en langue étrangère*, (18), 17–36. DOI : 10.4000/aile.1150.

BIRDSONG D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, **9**. DOI : 10.3389/fpsyg.2018.00081.

BLADON R. A. W. & AL-BAMERNI A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics*, **4**(2), 137–150. DOI : 10.1016/S0095-4470(19)31234-3.

BROWMAN C. P. & GOLDSTEIN L. (1995). Gestural syllable position effects in American English. In F. BELL-BERTI & L. RAPHAEL, Éds., *Producing Speech : Contemporary Issues. For Katherine Safford Harris*, p. 19–33. New York : AIP Press.

CAREY D. & MCGETTIGAN C. (2017). Magnetic resonance imaging of the brain and vocal tract : Applications to the study of speech production and language learning. *Neuropsychologia*, **98**, 201–211. DOI : 10.1016/j.neuropsychologia.2016.06.003.

CHUNG H. & KIM Y. (2021). Acoustic characteristics of Korean-English bilingual speakers'/l/ and the relationship to their foreign accent ratings. *Journal of Communication Disorders*, **94**, 106157. DOI : 10.1016/j.jcomdis.2021.106157.

COLANTONI L., KOCHETOV A. & STEELE J. (2023). Articulatory insights into the L2 acquisition of English-/l/ allophony. *Language and Speech*. DOI : 10.1177/00238309231200629.

KING H. & FERRAGNE E. (2015). The dark side of the tongue : The feasibility of ultrasound imaging in the acquisition of English dark /l/ in French learners. In *Ultrafest VII*.

KOCHETOV A. (2020). Research methods in articulatory phonetics II : Studying other gestures and recent trends. *Language and Linguistics Compass*, **14**(6), e12371. DOI : 10.1111/lnc3.12371.

LÉGER A., FERRAGNE E. & KING H. (2023). Is rhoticity on the tip of your tongue ? Investigating tongue shapes for English /r/ in French learners with ultrasound. In R. SKARNITZL & J. VOLÍN, Éds., *Proceedings of the 20th International Congress of Phonetic Sciences*, p. 2741–2745 : Guarant International.

LIM Y., TOUTIOS A., BLIESENER Y., TIAN Y., LINGALA S. G., VAZ C., SORENSEN T., OH M., HARPER S., CHEN W. *et al.* (2021). A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. *Scientific data*, **8**(1), 187. DOI : 10.1038/s41597-021-00976-x.

NARAYANAN S., NAYAK K., LEE S., SETHY A. & BYRD D. (2004). An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, **115**(4), 1771–1776. DOI : 10.1121/1.1652588.

NARAYANAN S. S., ALWAN A. A. & HAKER K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *The Journal of the Acoustical Society of America*, **101**(2), 1064–1077. DOI : 10.1121/1.418030.

PROCTOR M., WALKER R., SMITH C., SZALAY T., GOLDSTEIN L. & NARAYANAN S. (2019). Articulatory characterization of English liquid-final rimes. *Journal of Phonetics*, **77**, 100921. DOI : 10.1016/j.wocn.2019.100921.

ROACH P. (2007). *English phonetics and phonology : A practical course*. Cambridge (GB) : Cambridge university press, third ed édition.

SPROAT R. & FUJIMURA O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of phonetics*, **21**(3), 291–311. DOI : 10.1016/S0095-4470(19)31340-3.

TURTON D. (2017). Categorical or gradient ? An ultrasound investigation of /l/-darkening and vocalization in varieties of English. *Laboratory Phonology*, **8**(1). DOI : 10.5334/labphon.35.

WELLS J. C. (1982). *Accents of English. The British Isles*, volume 2. Cambridge : Cambridge University Press.

YUAN J. & LIBERMAN M. (2011). /l/ variation in American English : A corpus approach. *Journal of Speech Sciences*, **1**(2), 35–46. DOI : 10.20396/joss.v1i2.15025.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Evaluation de la dysarthrie parkinsonienne en lecture par la mesure de la déviation phonologique perçue : effets de la sévérité et du traitement dopaminergique

Alain Ghio[1], Muriel Lalain[1], Cindy Defais[1], Alexia Brevet[1], Manon Jayr[1], Danielle Duez[1], Marie Rebourg[1], Corinne Fredouille[2], Virginie Woisard[3], François Viallet[1,4]

(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(2) Laboratoire Informatique d'Avignon, Univ. Avignon
(3) UT2J, LNPL URI EA 4156 & CHU Toulouse
(4) Service de neurologie, CH du Pays d'Aix, Aix-en-Provence, France
alain.ghio@univ-amu.fr, muriel.lalain@univ-amu.fr

RESUME

La perte d'intelligibilité chez des patients atteints de troubles de la production de la parole est un élément important du bilan orthophonique. Nous proposons un test fondé sur des séquences délexicalisées de type Voyelle-Consonne-Voyelle ainsi que sur des voyelles isolées extraites automatiquement d'un corpus de lecture. 12 locuteurs contrôles et 30 patients atteints de la maladie de Parkinson ont participé à l'expérience. Pour chaque locuteur, nous avons extrait automatiquement 50 séquences VCV et 50 voyelles isolées qui ont été soumises à l'identification par des auditeurs. La mesure de l'intelligibilité est fondée sur le comptage du nombre de traits phonémiques mal perçus par les auditeurs (Perceived Phonological Deviation = PPD). Nos résultats montrent une différence significative entre le groupe contrôle et les patients. Nous n'observons aucun effet lié au traitement dopaminergique. En revanche, nous observons une augmentation du PPD en fonction de la sévérité de la dysarthrie évaluée cliniquement par le neurologue.

ABSTRACT

Assessment of Parkinsonian dysarthria in reading by measuring perceived phonological deviation: effects of severity and dopaminergic treatment

The loss of intelligibility in patients with speech production disorders is an important element of the speech therapy assessment. We propose a test based on delexicalized sequences as Vowel-Consonant-Vowel type as well as on isolated vowels automatically extracted from a reading corpus. 12 control speakers and 30 patients suffering from Parkinson's disease participated in the experiment. For each speaker, we automatically extracted 50 VCV sequences and 50 isolated vowels which were submitted for identification by listeners. The measurement of intelligibility is based on counting the number of phonemic features misperceived by listeners (Perceived Phonological Deviation = PPD). Our results show a significant difference between the control group and the patients. We do not observe any effect linked to dopaminergic treatment. On the other hand, we observe an increase in PPD depending on the severity of dysarthria assessed clinically by the neurologist.

MOTS-CLES : phonétique clinique, maladie de Parkinson, dysarthrie, intelligibilité

KEYWORDS: clinical phonetics, Parkinson's disease, dysarthria, intelligibility

1 L’intelligibilité dans la maladie de Parkinson

1.1 La maladie de Parkinson, ses symptômes, ses traitements

La maladie de Parkinson (MDP) est une pathologie neurodégénérative qui se caractérise par la disparition progressive de cellules dans les ganglions de la base (substance noire des noyaux gris centraux). La principale conséquence de cette perte neuronale est la réduction de la production de dopamine, un neurotransmetteur dans une région essentielle au contrôle des mouvements (Kalia et al., 2015). La maladie de Parkinson est donc avant tout une maladie qui affecte les fonctions motrices. D’après le (GBD, 2016), environ 120 000 personnes sont atteintes par cette maladie en France en 2016 et 6 millions à travers le monde.

Parmi les symptômes moteurs de la maladie, on observe fréquemment une dysarthrie, qui est un trouble de la réalisation motrice de la parole dont l’origine est une lésion du système nerveux central ou périphérique (Pinto et al., 2010). (Pahwa et al. 2007) estiment que 90% des patients parkinsoniens sont touchés par un trouble de la parole, hypokinétique dans le cas de la MDP. La parole parkinsonienne est souvent caractérisée par une insuffisance prosodique avec une faible modulation de l’intensité et de la hauteur et une diminution de l’accentuation, un débit variable avec des ralentissements et accélérations paroxystiques, une imprécision des consonnes notamment occlusives, et une voix à la fois soufflée et éraillée (Pinto et al., 2010).

La perte d’intelligibilité présente dans la maladie peut avoir un fort impact sur la qualité de vie des malades. Ainsi, Miller et al. (2011) ont mis en évidence une auto perception négative chez les patients parkinsoniens atteints de dysarthrie, exprimant leur frustration et un sentiment de compétences amoindries dans leurs capacités de communication.

Le traitement pharmacologique par lévodopa est considéré comme le traitement de référence pour la maladie de Parkinson (De Letter et al., 2005). Il agit essentiellement sur les symptômes moteurs de la maladie en atténuant la bradykinésie, la rigidité et les tremblements. Concernant l’intelligibilité de la parole, l’effet du traitement dopaminergique reste variable d’un sujet à l’autre (De Letter et al., 2005).

1.2 La mesure de l’intelligibilité en pratique clinique

(Kent et al., 1989) distinguent deux approches différentes pour évaluer l'intelligibilité de la parole :
- une évaluation subjective faite par des cliniciens qui écoutent le patient et proposent une note sur une échelle standardisée (ex : 5/10).
- une approche décrite comme plus « objective » d’identification perceptive d'éléments (des mots) qui se mesure généralement par le pourcentage d'éléments correctement reconnus par un auditeur

Les évaluations subjectives sont répandues en pratique clinique essentiellement parce qu’elles sont faciles à réaliser. Cette approche est similaire à celle utilisée dans l'évaluation de la qualité de la voix où les cliniciens jugent subjectivement la performance (0 = normal, 1 = léger trouble, 2 = trouble modéré, 3 = trouble sévère). Il s’agit plus d’un jugement « esthétique » que d’un processus linguistique. La subjectivité de l'auditeur introduit une variable très difficile à contrôler. Ce phénomène a été largement étudié dans les évaluations de la qualité vocale, où la fiabilité inter-évaluateurs est notoirement faible. Il en va de même pour l'intelligibilité où l'évaluation subjective est imparfaite, notamment en raison du faible accord entre les auditeurs (McHenry, 2011).

Dans le cas de l'identification d'items, il s'agit pour le patient de lire une liste de mots issus de tests standardisés pendant que le clinicien retranscrit ce qu'il a compris (Enderby 1983 ; Auzou et al., 2006). Les transcriptions des cliniciens sont ensuite confrontées à la liste initiale ; un comptage des bonnes réponses permet d’obtenir une note globale. Les limitations de ce type de test résident dans la capacité des auditeurs à restaurer les séquences distordues. Cet effet est d’autant plus fort que les auditeurs ont une connaissance forte des mots utilisés dans le test et que ces mots sont peu ambigus et donc fortement prédictibles. C’est généralement le cas des orthophonistes qui peuvent faire un usage si

important de ces listes qu'elles finissent par les connaître par cœur. La BECD par exemple ne comporte que 50 mots (Auzou et al., 2006). Les travaux de Rebourg (2022) ont mis en évidence des biais d'apprentissage par les auditeurs lors de l'utilisation de listes courtes de mots avec au final des résultats d'évaluation pouvant varier du simple au double. Le biais lié à cette connaissance et donc à la forte influence des mécanismes perceptifs descendants est un score d'intelligibilité surévalué car la restauration phonémique de l'auditeur rend « transparentes » les distorsions de production (Lalain et al, 2020). Il en résulte une sensibilité insuffisante au changement à cause de cette restauration phonémique chez les auditeurs.

Afin de limiter ces biais, nous proposons donc un test fondé sur des éléments délexicalisés sous la forme de séquences de type Voyelle-Consonne-Voyelle (VCV) où seule la consonne est évaluée ainsi que sur des voyelles isolées (VOY) extraites automatiquement d'un corpus de lecture. Nous nous positionnons donc sur une définition restreinte de l'intelligibilité entendue comme la quantité de parole comprise uniquement à partir d'informations issues du signal acoustique (Ghio et al., 2021). Cette définition permet de distinguer cette notion de la compréhensibilité qui elle, intègre toutes les informations possibles à la disposition de l'auditeur, notamment indépendantes du signal acoustique. Cette évaluation est proche de ce qui a pu être fait à partir de pseudomots.(Lalain et al., 2020). Le travail exposé ici présente l'avantage de se fonder sur des séquences de parole continue, plus naturelle que la production de pseudomots artificiels. Nous faisons l'hypothèse que la mesure de la déviation phonologique perçue (Lalain et al., 2020) dans ce cadre de parole continue rendra compte de la sévérité du trouble de la parole et donc de la perte d'intelligibilité.

2 Matériel et méthode

2.1 Locuteurs

Dans cette étude, nous avons sélectionné les locuteurs à partir de la base de données AHN du service de neurologie du Centre Hospitalier du Pays d'Aix (Ghio et al., 2012). Nous avons sélectionné 12 locuteurs contrôles et 30 patients atteints de la maladie de Parkinson (MDP). L'âge moyen du groupe contrôle était de 58 ans (de 43 à 68 ans, écart-type de 6.9 ans).

L'âge moyen du groupe MDP était de 63.5 ans (de 49 à 81 ans, écart-type de 8.8 ans). Chez les patients, la durée de la maladie était en moyenne de 11 ans (de 2 à 30 ans, écart-type de 5.1 ans). Tous les patients étaient traités par L-dopa de façon usuelle. Pour observer de façon plus nette les effets de la MDP, tous les patients avaient été sevrés de médicament pendant plus de 12 heures, délai usuel pour annuler les effets pharmacologiques. Il s'agissait de la situation OFF dopa. Une fois l'examen clinique et les enregistrements effectués en situation OFF dopa, les patients recevaient leur dose usuelle de dopa et le protocole était répété une heure après en situation ON dopa.

Avant l'enregistrement, l'évaluation motrice de chaque patient avait été menée par un neurologue à l'aide de l'échelle UPDRS (Unified Parkinson's Disease Rating Score). Ce score UPDRS moteur (section III) sert surtout de mesure pour quantifier la progression de la maladie. L'item n°18 de cette échelle est particulièrement informatif car il indique la sévérité de la dysarthrie dans une approche subjective clinique avec les conventions suivantes: 0=normal; 1=légère perte d'expression, de diction et / ou de volume; 2=monotone, flou, mais compréhensible, modérément altéré; 3=déficience marquée, difficile à comprendre; 4=inintelligible.

Sur nos données, le score UPDRS moyen de déficit moteur en situation OFF était de 33.1 (σ=8.8), ce qui correspond à une maladie installée. Une heure après la prise de L-dopa (situation ON), le score UPDRS moyen était de 14.0 (σ=8.5), ce qui correspond à une amélioration de plus de 50% par rapport à la situation OFF, amélioration notable qui valide la dépendance à la dopa des patients sélectionnés. Au niveau de la sévérité de la dysarthrie, nous disposions de l'item 'parole' (n°18) de l'UPDRS, qui était en moyenne de 1.4 (σ=0.66) en situation OFF et de 0.98 (σ=0.87) en situation ON. Nos locuteurs avaient donc une altération légère de la parole, significativement plus dégradée en OFF ($p < 0.05$). Il

est important de préciser que cette évaluation n'est pas aveugle, le neurologue connait le patient ainsi que son état pharmacologique.
Parallèlement, nous disposions d'évaluations subjectives récoltées auprès d'un jury de 4 phoniatres/orthophoniste du CHU Toulouse. La tâche de cette évaluation était de donner en aveugle un score entre 0 (parole inintelligible) et 10 (parole normale) en écoutant la lecture des patients et des sujets contrôles. Le score de sévérité ainsi obtenu était en moyenne à 7.85 (σ=1.25) pour le groupe MDP-OFF, 7.46 (σ=1.47) pour MDP-ON et 9.56 (σ=0.44) pour le groupe Contrôle.

2.2 Corpus

Tous les locuteurs ont lu les premiers paragraphes du texte de la « chèvre de M. Seguin » d'Alphonse Daudet à hauteur, intensité et vitesse confortable. Les enregistrements se sont déroulés dans une pièce calme du Centre Hospitalier du Pays d'Aix avec un microphone AKG C520 alimenté par le dispositif EVA2 (Ghio et al., 2012). Le format des fichiers audio est un codage en 16 bits avec une fréquence d'échantillonnage de 25 kHz. L'objectif ensuite était d'extraire de ces lectures des séquences Voyelle-Consonne-Voyelle où la consonne est la cible, ainsi que des voyelles isolées.
Un travail préalable de sélection des items a été effectué par Defais (2021). Les consonnes cibles ont été choisies dans des contextes où elles étaient en position intervocalique. Pour les voyelles isolées, nous avons exclu les schwas. Nous nous sommes limités au début du texte car certains centres hospitaliers effectuent la lecture avec une version raccourcie du document. Nous avons décidé de sélectionner 50 séquences VCV et 50 voyelles isolées dans cette partie de texte.
La Table 1 fournit la liste des items retenus. La ligne « Ortho » est la transcription orthographique syllabique du texte. La ligne « Phono » est la transcription phonétique de la syllabe (norme de codage du LIA, Avignon). La ligne « Label VCV » est l'identifiant de la séquence consonantique sélectionnée, la ligne « Label VOY » est l'identifiant de la voyelle isolée. Ainsi, l'item « euGGin-02» cible la consonne [g] en contexte [Ø] et [ɛ̃]. Cet élément est la 2ième séquence VCV sélectionnée et se situe sur le mot « Seguin ». De même, l'item « in-02 » correspond au [ɛ̃] de « Seguin ». C'est la 2ième voyelle sélectionnée.

Ortho	Mon	sieur	Se	guin	n'a	vait	ja	mais	(z) eu	de	bo	nheur	a	vec	ses	chè	vres.	ll	les	per
Phono	mm eu	ss yy eu	ss eu	gg in	nn aa	vv ai	jj aa	mm ei	zz uu	dd eu	bb oo	nn oe rr	aa	vv ai kk	ss ai	ch ai	vv rr	ii ll	ll ei	pp ai rr
LabelVCV	-	-	euSSeu-01	euGGin-02	inNNaa-03	aaVVai-04	aiJJaa-05	aaMMei-06	eiZZuu-07	uuDDeu-08	euBBoo-09	ooNNoe-10	-	aaVVai-11	-	aiCHai-12	-	-	-	eiPPai-13
LabelVoy	-	eu-01	-	in-02	-	-	-	-	uu-03	-	oo-04	oe-05	aa-06	-	ai-07	-	-	ii-08	-	ai-09

Ortho	dait	tou	tes	de	la	mê	me	fa	çon;	un	beau	ma	tin,	e	lles	ca	ssaient	leur	cor	de,
Phono	dd ei	tt ou	tt	dd eu	ll aa	mm ai	mm eu	ff aa	ss on	in	bb au	mm aa	tt in	ai	ll eu	kk aa	ss ai	ll oe rr	kk oo rr	dd
LabelVCV	-	-	-	-	-	aaMMai-14	-	euFFaa-15	aaSSon-16	-	inBBau-17	auMMaa-18	aaTTin-19	-	-	euKKaa-20	aaSSai-21	aiLLoe-22	-	-
LabelVoy	-	ou-10	-	-	aa-11	ai-12	-	-	on-13	in-14	au-15	aa-16	in-17	ai-18	-	-	-	oe-19	oo-20	-

Ortho	s'en	n_a	llait	dans	la	mon	tagne	et	là	- haut	le	loup	les	man	geait.	Ni	les	ca	resses	de
Phono	ss an	nn aa	ll ei	dd an	ll aa	mm on	tt aa nn yy	ei	ll aa	au	ll eu	ll ou	ll ei	mm an	jj ai	nn ii	ll ai	kk aa	rr ai ss	dd eu
LabelVCV	-	-	aaLLei-23	eiDDan-24	anLLaa-25	aaMMon-26	onTTaa-27	-	eiLLaa-28	-	-	euLLou-29	-	eiMMan-30	anJJai-31	aiNNii-32	-	aiKKaa-33	aaRRai-34	-
LabelVoy	an-21	-	-	an-22	-	on-23	-	-	aa-24	au-25	eu-26	ou-27	-	an-28	-	ii-29	-	aa-30	-	-

Ortho	leur	maî	tre	ni	la	peur	du	loup,	rien	ne	les	re	te	nait.	C'é	tait	pa	raît	t-il,	des
Phono	ll oe rr	mm ai	tt rr	nn ii	ll aa	pp oe rr	dd uu	ll ou	rr yy in	nn eu	ll ei	rr eu	tt eu	nn ai	ss ai	tt ai	pp aa	rr ai	tt ii ll	dd ei
LabelVCV	-	-	-	-	-	aaPPoe-35	-	-	-	inNNeu-36	-	eiRReu-37	-	euNNai-38	-	aiTTai-39	aiPPaa-40	aaRRai-41	aiTTii-42	-
LabelVoy	oe-31	-	-	ii-32	-	oe-33	uu-34	ou-35	in-36	-	-	-	-	-	ai-37	-	aa-38	-	ii-39	-

Ortho	chè	vres	in	dé	pen	dan	tes,	vou	lant	à	tout	prix	le	grand	air	et	la	li	ber	té.
Phono	ch ai	vv rr in	in	dd ei	pp an	dd an	tt eu	vv ou	ll an	aa	tt ou	pp rr ii	ll eu	gg rr an	tt ai rr	ai	ll aa	ll ii	bb ai rr	tt ei
LabelVCV	eiCHai-43	-	-	inDDei-44	eiPPan-45	anDDan-46	-	-	-	-	aaTTou-47	-	-	-	anTTai-48	-	-	aaLLii-49	iiBBai-50	-
LabelVoy	ai-40	-	in-41	ei-42	an-43	-	-	ou-44	an-45	-	ou-46	ii-47	-	an-48	-	-	-	ii-49	-	ei-50

Table 1 : *Sélection des séquences VCV et des voyelles isolées dans le début du texte de la « chèvre de Monsieur Seguin » d'Alphonse Daudet (Defais, 2021)*

L'aspect raccourci du texte ne permet pas d'avoir un inventaire équilibré des phonèmes (Table 2) avec notamment un déficit de [g], [f] et [z], qui n'apparaissent qu'une fois dont une fois en contexte de liaison non systématique pour le [z].

p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	R
4	6	2	3	4	1	1	3	2	2	1	2	5	5	6	3

i	Y	e	ø	ɛ	œ	a	o	ɔ	u	ɑ̃	ɛ̃	ɔ̃
6	2	2	2	6	4	6	2	2	5	6	5	2

Table 2 : *nombre d'occurrences des consonnes et des voyelles dans le corpus*

2.3 Préparation des stimuli

Dans l'étude de (Duez et al., 2020), les stimuli avaient été repérés manuellement. Cette approche bien que précise et fiable nécessite un très long travail pour espérer ensuite un usage courant en orthophonie. Nous avons donc pris le parti d'utiliser des techniques d'alignement automatique forcé ne nécessitant que peu de manutention. Un tel principe pourrait être utilisé en situation clinique. Le principe est le suivant :

- un opérateur réécoute le texte lu et annote si nécessaire les disfluences par rapport à une lecture totalement congruente au texte. Il s'agit principalement de répétitions ou d'élisions.
- le texte est ensuite phonétisé et les frontières des phonèmes sont posées automatiquement par le biais de l'aligneur du laboratoire d'informatique d'Avignon
- les séquences cibles sont repérées dans le flux et extraites de façon automatique

Pour les séquences VCV où seule la consonne va être évaluée, le critère est de prendre comme début le milieu de la voyelle précédant la consonne et comme fin le milieu de la voyelle suivant la consonne (FIGURE 1). Du fait de la présence de pauses, nous avons décidé d'intégrer aussi des séquences de type Pause + Consonne + moitié de Voyelle. Le début de la séquence est alors le début de la consonne. Pour les voyelles isolées, le stimulus est contraint par les bornes de la voyelle. Pour éviter les effets de bords des débuts et fin des stimuli, nous avons appliqué un fade-in de 10 ms et un fade-out de 10 ms avec le logiciel Sox.

Une vérification manuelle de l'alignement auto a été menée sur 4 locuteurs contrôles et 4 Patients MDP; nous avons relevé 1% d'erreurs sur les sujets contrôles et 5% sur les patients, pourcentages que nous considérons comme acceptable au regard de la quantité de données traitées.

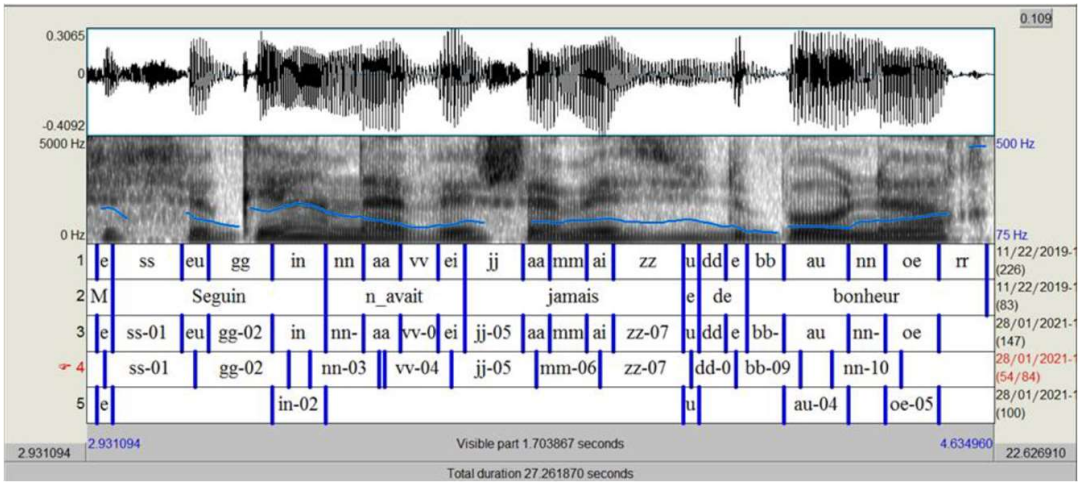


FIGURE 1: Résultat de l'alignement phonétique de la lecture et mise en évidence des séquences VCV cibles et des voyelles isolées

2.4 Les tests de perception

Pour chaque locuteur, nous avons extrait automatiquement 50 séquences VCV et 50 voyelles isolées. L'ensemble des 3600 séquences [50 x (12 ctrl + 30 on-dopa + 30 off-dopa)] a été soumis à l'identification perceptive de la cible par 18 auditeurs naïfs francophones natifs ne présentant aucun trouble auditif ni visuel. L'expérience portait d'abord sur les 3341 consonnes puis sur les 3549 voyelles, certaines séquences étant non disponibles pour certains locuteurs (omission de mots, réduction de syllabes…) Ces tests ont eu lieu au Centre d'Expérimentation sur la parole (http://cep.lpl-aix.fr/) au Laboratoire Parole et Langage à Aix-en-Provence. La présentation des stimuli et le recueil des réponses étaient automatisés grâce au dispositif Perceval-Lancelot (André et al., 2003). Chaque auditeur, portant un casque audiophonique Superlux HD 681B, a transcrit 3 blocs de 185 éléments, soit 555 stimuli. L'intensité de lecture du son a été préréglée par l'auditeur pour être confortable et

optimale pour la tâche. La réponse était en choix forcé et l'auditeur cliquait sur le phonème qu'il estimait avoir perçu. Chaque test a commencé avec quatre stimuli d'entraînement. Chaque élément était présenté une fois automatiquement mais l'auditeur pouvait répéter la lecture deux fois. L'auditeur a eu une pause de 5 minutes entre les blocs. Au final, chaque stimulus a été écouté par 3 auditeurs différents afin d'augmenter la fiabilité de la tâche d'identification.

2.5 Prétraitement des données

Une fois les réponses recueillies, les phonèmes perçus ont été examinés par rapport au phonème cible. Nous appelons « score de déviation phonologique perçue » (Perceived Phonological Deviation, PPD) le nombre de traits phonétiques qui diffèrent entre le phonème attendu et la réponse donnée par l'auditeur. Un score de 0 signifie que le phonème a été correctement identifié. Un score de *N* signifie qu'il y avait *N* traits phonétiques mal identifiés. La décomposition en traits que nous avons choisie est celle publiée dans (Lalain et al., 2020). Plus le score PPD est faible, meilleure peut être considérée l'intelligibilité du locuteur. En effet, nous partons du postulat que nos auditeurs sont normo-entendants, que les conditions d'écoute sont optimales et donc que toute erreur de perception est liée à un problème de production de la parole chez le locuteur.

3 Résultats

Le score PPD d'un locuteur est la moyenne des scores obtenus sur toutes les écoutes des séquences relatives à ce locuteur (50 items x 3 auditeurs = 150 valeurs). Nous avons distingué le score obtenu sur les consonnes et celui obtenu sur les voyelles. Tous les tests statistiques ont été effectués dans l'environnement logiciel R version 4.2.1 Des modèles linéaires à effets mixtes (fonction lmer) ont été utilisés pour analyser les scores PPD considérés comme une variable continue. Etant donnée la structure 'within' chez les patients, que l'on retrouve dans le groupe DOP et OFF, le 'locuteur' a été pris comme random effect.

3.1 Effets du groupe et du traitement pharmacologique

Les distributions du PPD par groupe suivent des lois normales (test de Shapiro). Sur les consonnes, nos résultats montrent des différences significatives entre le groupe contrôle (moyenne $PPD_{control}$=0.31) et les patients avec traitement ON dopa (moyenne $PPD_{MDP\text{-}DOP}$=0.73) ou en situation de sevrage OFF dopa (moyenne $PPD_{MDP\text{-}OFF}$=0.70). Nous n'observons aucun effet lié au traitement dopaminergique (FIGURE 2, gauche). Pour les voyelles, les tendances sont identiques mais avec des valeurs de PPD plus élevées (moyenne $PPD_{control}$=0.86 ; moyenne $PPD_{MDP\text{-}DOP}$=1.13 ; moyenne $PPD_{MDP\text{-}OFF}$=1.09). Nous n'observons là aussi aucun effet lié au traitement dopaminergique (FIGURE 2, droite).

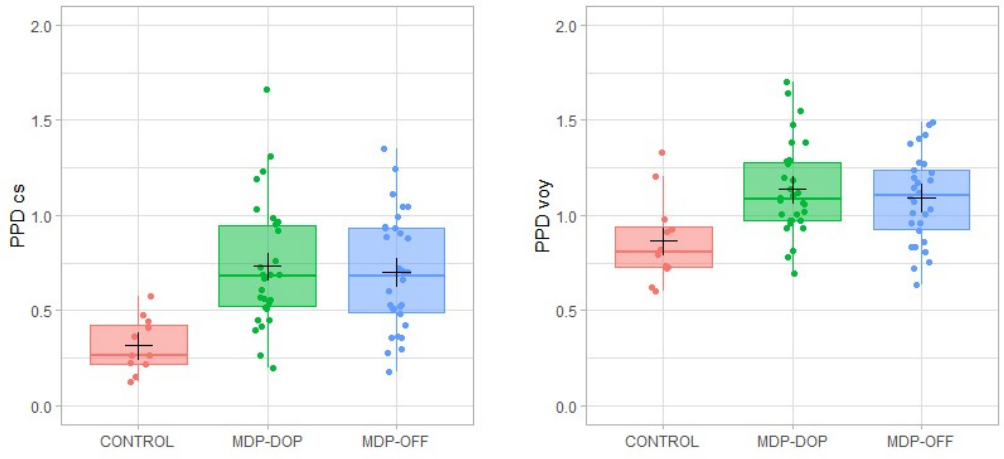


FIGURE 2: score PPD par groupe pour les consonnes (à gauche) et pour les voyelles (à droite)

3.2 Effets liés à la sévérité de la dysarthrie

Le nombre de patients ayant une dysarthrie sévère (grade 3 pour l'item parole de l'UPDRS) étant insuffisant (N=3), nous avons regroupé les patients de grade 2 et 3. Ce regroupement permet d'avoir des groupes équilibrés en nombre d'éléments et distribués en loi normale (test de shapiro). Nous observons (FIGURE 3) une augmentation du PPD en fonction de la sévérité de la dysarthrie. Pour les consonnes, $PPD_{control}$=0.31 ; PPD_{park0}=0.56 ; PPD_{park1}=0.64 ; PPD_{park2+}=0.94. Les différences sont significatives entre le groupe Contrôle et les groupes PARK, de même entre le groupe Park2+ et les autres groupes. En revanche, les différences entre le groupe PARK0 et PARK1 ne sont pas significatives.

Sur les voyelles, les résultats vont dans le même sens mais les différences entre les groupes sont moins marquées : $PPD_{control}$ =0.86 ; PPD_{park0}=0.97 ; PPD_{park1}=1.08; PPD_{park2+}=1.26. Les différences sont significatives entre le groupe Contrôle et les groupes PARK1 et PARK2+. En revanche, la différence entre le groupe Contrôle et le groupe PARK0 n'est pas significative. De même, nous n'observons aucune différence significative entre les groupes Park.

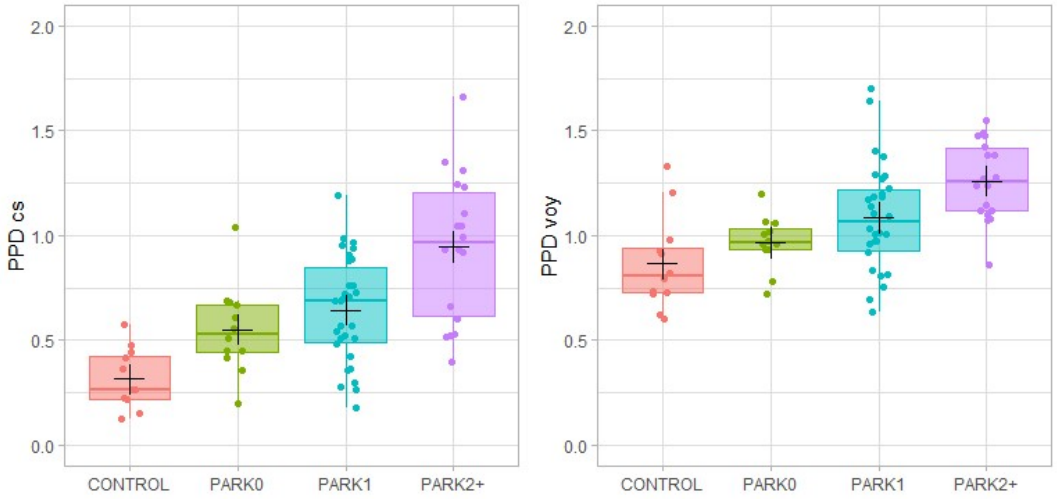


FIGURE 3: Score PPD en fonction de l'item parole de l'UPDRS pour les consonnes (à gauche) et pour les voyelles (à droite)

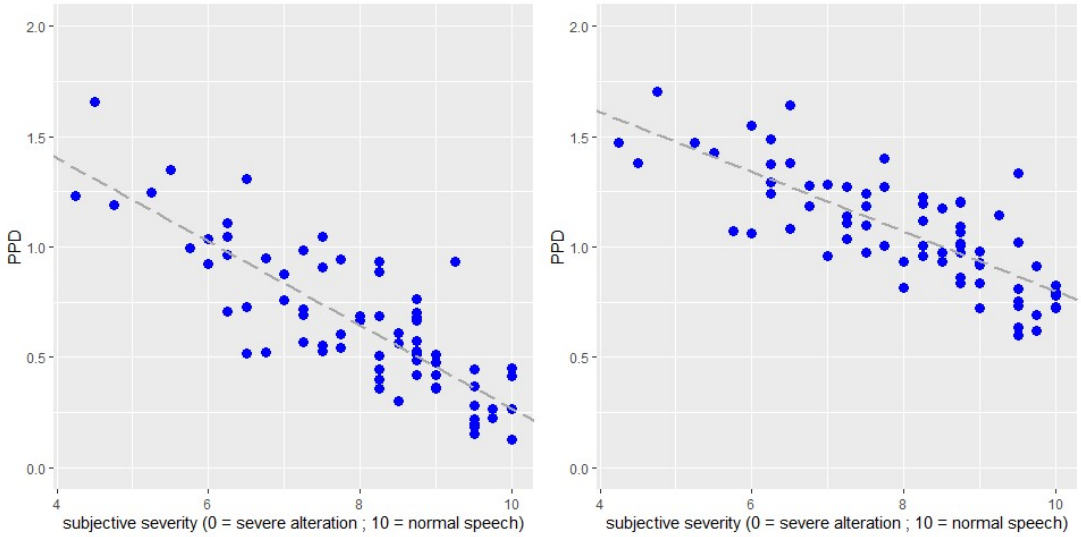


FIGURE 4: Score PPD en fonction de la sévérité de la dysarthrie pour les consonnes (à gauche) et pour les voyelles (à droite)

Si nous corrélons le score PPD avec les évaluations subjectives récoltées auprès d'un jury de 4 phoniatres/orthophonistes du CHU Toulouse (cf § 2.1 Locuteurs), nous obtenons un coefficient de corrélation de Pearson r = -0.84 pour les consonnes et r = -0.78 pour les voyelles (FIGURE 4).

4 Discussion et conclusion

Ces résultats valident notre méthode d'évaluation perceptive. Comme cette méthode est fondée sur du décodage acoustico-phonétique, opération fonctionnelle dans la perception de la parole, cela permet de réduire l'impact des biais perceptifs classiques, minimisant ainsi une forme de subjectivité indésirable dans un cadre d'évaluation clinique. Cette étude valide la proposition de Duez et al. (2020) sur un corpus important et dans une approche semi-automatique. Nous confirmons aussi la pertinence des consonnes par rapport aux voyelles pour la mesure de l'intelligibilité de la parole. En effet, que ce soit dans la capacité de distinguer la gravité du trouble moteur ou dans les corrélations avec des indices de sévérité, le recours aux consonnes permet une meilleure observation des différences.

D'un point de vue clinique, nous n'observons pas d'effet lié au traitement pharmaceutique à la dopa. Soit notre mesure n'est pas assez fine pour refléter l'effet du traitement, soit le traitement lui-même n'a pas d'effet sur la parole. A ce sujet, l'impact du traitement dopaminergique sur la parole reste controversé dans la littérature, notamment sur l'intelligibilité (De Letter et al., 2005).

L'obtention d'un PPD sur de la lecture et donc sur de la parole continue enlève les réticences que l'on a pu observer dans l'usage de séquences délexicalisées de type pseudo mots (Lalain et al., 2020). En effet, l'usage de pseudomots n'est pas naturel contrairement à une tache de lecture.

Au niveau de la clinique orthophonique, l'outil d'évaluation avec les scores PPD permettrait d'obtenir une évaluation précise de l'intelligibilité du patient parkinsonien. De plus, la mise en évidence des phonèmes altérés pourrait orienter les objectifs de la prise en soins. Le recours à des procédés semi-automatiques s'avère possible et ouvre donc des pistes à un usage réel.

Dans le cadre du projet ANR Rugbi, nous avons appliqué cette méthode sur un corpus plus vaste de 316 locuteurs (111 CTRL +205 patients ON/OFF). Les données sont en cours de traitement. Bien qu'imparfait, le repérage automatique des séquences cibles reste efficace et cela nous conduit à imaginer, à terme, de proposer une version utilisable par les orthophonistes en cabinet pour tous les patients ayant des troubles de la production de la parole. La prise en compte de la dimension prosodique est une perspective que nous envisageons à l'avenir ainsi que de la dimension compréhensibilité qui, elle, met les mécanismes de décodage acoustico-phonétique en contexte communicationnel.

Remerciements

Cette recherche a été financée en partie par le projet ANR-18-CE45-0008 RUGBI, de l'Agence nationale de la recherche, ainsi que par un BQR LPL. Les auteurs tiennent à remercier Alain Purson et Ludovic Jankowksi pour leur assistance lors de l'enregistrement des patients ainsi que le personnel du CEP pour leur assistance dans l'expérience perceptive.

Références

ANDRÉ C, GHIO A, CAVE C, TESTON B. (2003) PERCEVAL: a Computer-Driven Systemfor Experimentation on Auditory and Visual Perception. International Congress of Phonetic Sciences (ICPhS), Barcelona, Spain. pp.1421-1424.

AUZOU P, ROLLAND-MONNOURY V. (2006), Batterie d'Evaluation Clinique de la Dysarthrie. 1st ed. Isbergues: Ortho Edition; 2006.

DEFAIS C. (2021), L'évaluation de l'intelligibilité dans les troubles de la production de la parole : utilisation de séquences VCV en lecture chez des patients atteints de cancer des VADS, Certificat de Capacité d'Orthophonie, Aix-Marseille Univ. https://dumas.ccsd.cnrs.fr/dumas-03349024/document

DE LETTER, M., SANTENS, P., & BORSEL, J. V. (2005). The effects of levodopa on word intelligibility in Parkinson's disease. *Journal of Communication Disorders*, 38(3), 187-196.

DUEZ, D., GHIO, A., & VIALLET, F. (2020). Effect of linguistic context on the perception of consonants in Parkinsonian Read French speech. *Clinical Linguistics & Phonetics*, 2020, 35 (10), pp.926-944.

ENDERBY, P. (1983). Frenchay dysarthria assessment. Pro-Ed, Austin Tex.

GBD 2016 AND PARKINSON'S DISEASE COLLABORATORS, (2016), Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study, *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, 2018.

GHIO A, POUCHOULIN G, TESTON B, PINTO S, FREDOUILLE C, DE LOOZE C, ROBERT D, VIALLET F, GIOVANNI A (2012) How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? Speech Communication, 54 (5), pp.664-679.

GHIO, A., LALAIN, M., REBOURG, M., MARCZYK, A., FREDOUILLE, C., & WOISARD, V. (2021). Validation of an Intelligibility Test Based on Acoustic-Phonetic Decoding of Pseudo-Words : Overall Results from Patients with Cancer of the Oral Cavity and the Oropharynx. Folia Phoniatrica et Logopaedica: Official Organ of the International Association of Logopedics and Phoniatrics (IALP). https://doi.org/10.1159/000519427

KALIA L. V., LANG A. E. (2015), Parkinson's disease, *The Lancet*, vol. 386, no. 9996, pp. 896–912.

KENT RD, WEISMER G, KENT JF, ROSENBEK JC.(1989) Toward phonetic intelligibility testing in dysarthria. Journal of Speech and Hearing Disorders. 1989 Nov; 54(4):482-99.

LALAIN, M., GHIO, A., GIUSTI, L., ROBERT, D., FREDOUILLE, C., & WOISARD, V. (2020). Design and Development of a Speech Intelligibility Test Based on Pseudowords in French : Why and How? Journal of Speech, Language, and Hearing Research, 63 (7), pp.2070-2083

MILLER N., ANDREW S., NOBLE E., WALSHE M., (2011), Changing perceptions of self as a communicator in Parkinson's disease: a longitudinal follow-up study, *Disability and Rehabilitation*, vol. 33, no. 3, pp. 204–210, 2011.

PAHWA R., LYONS K. E., KULLER, W. C. (2007). *Handbook of Parkinson's Disease* (4th ed., p. 530). New York, NY:Informa Healthcare.

PINTO S., GHIO A., TESTON B., VIALLET F. (2010). La dysarthrie au cours de la maladie de Parkinson. Histoire naturelle de ses composantes : Dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, 166(10), 800-810. https://doi.org/10.1016/j.neurol.2010.07.005

REBOURG M. (2022), Évaluation de l'intelligibilité après un cancer ORL : Approche perceptive par décodage acoustico-phonétique et mesures acoustiques, Thèse de doctorat, Univ. Aix-Marseille

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Évaluation perceptive de l'anticipation de la prise de parole lors d'interactions dialogiques en français

Rémi Uro[1,2], Albert Rilliard[2], David Doukhan[1], Marie Tahon[3], Antoine Laurent[3]
(1) Institut National de l'Audiovisuel, Paris, France.
(2) Université Paris Saclay, CNRS, LISN, France.
(3) LIUM, Le Mans Université, France.
{ruro,ddoukhan}@ina.fr albert.rilliard@lisn.fr,
{marie.tahon,antoine.laurent}@univ-lemans.fr

RÉSUMÉ

Cette étude présente un test perceptif évaluant les indices permettant la planification de la prise de parole lors d'interactions orales spontanées. Des Unités Inter-Pauses (IPU) ont été extraites de dialogues du corpus REPERE et annotées en terminalité. Afin de déterminer quels paramètres affectent les jugements de la possibilité de prendre la parole, les stimulus ont été présentés sous forme audio ou textuelle. Les participant·es devaient indiquer la possibilité de prendre la parole « Maintenant », « Bientôt » ou « Pas encore », à la fin des IPU tronqués de 0 à 3 mots prosodiques. Les participant·es sont moins susceptibles de prendre la parole pour les frontières non terminales en modalité audio que textuelle. La modalité audio permet également d'anticiper une fin de tour de parole au moins trois mots avant sa fin, tandis que la modalité textuelle permet moins d'anticipation. Ces résultats soutiennent l'importance des indices contenus dans la parole pour la planification des interactions dialogiques.

ABSTRACT

A perceptual evaluation of the anticipation of turn-taking in French dialogic interactions

This study presents a perceptual test evaluating the cues allowing for turn-taking planning in French spontaneous interactions. Inter-Pausal Units (IPUs) were extracted from dialogues from the REPERE corpus and annotated with regard to terminality. In order to determine which parameters affect the possibility of turn-taking, stimuli were presented with audio-only or text-only modality. Participants had to indicate whether they could take the floor "Now", "Soon" or "Wait" after an IPU with 0 to 3 prosodic words removed. Participants were less likely to take the floor for non-terminal boundaries with the audio modality than with the text one. The audio modality also allows for the anticipation of the end of a turn up to three words before its end, while the text modality allows for less anticipation. These results support the importance of speech cues for the planning of dialogic interactions.

MOTS-CLÉS : Tour de parole, analyse de conversation, évaluation perceptive, TRP.

KEYWORDS: Turn-taking, conversation analysis, perceptual evaluation, interruption, TRP.

1 Introduction

Lors des interactions parlées, la gestion des tours de parole est critique pour intervenir au moment adéquat : sans couper la parole de son interlocuteur et sans laisser trop de silence (Bosch *et al.*, 2005;

Stivers *et al.*, 2009; Levinson & Torreira, 2015). Il est fondamental pour cela d'anticiper ces moments pertinents de prise de parole (communément appelés TRP - Transition Relevance Places (Sacks *et al.*, 1974)) pour planifier son énoncé (Levinson, 2016). L'étude de ces phénomènes répond à un grand nombre d'enjeux théoriques et applicatifs : analyse conversationnelle, description de comportements des locuteur·ices et des différences culturelles, conception de systèmes automatiques de gestion du dialogue pour les interactions Humain-Machine (Skantze, 2021).

Grosjean montre que la fin d'une phrase lue est prédictible grâce à des indices prosodiques (Grosjean, 1996). Les travaux de Magyari & de Ruiter (2012) sur des conversations téléphoniques en néerlandais démontrent la capacité des auditeur·ices à prédire si le tour de parole courant va se terminer ou continuer ; cette capacité augmente plus on s'approche de la fin du tour. Les pauses sont également un aspect important de l'analyse conversationnelle. Avec la syntaxe et la prosodie, les pauses fournissent des indices robustes pour permettre la détermination automatique de la fin d'un tour de parole (Christodoulides, 2018). Gotoh & Renals (2000) montre que l'analyse de la durée des pauses permet une meilleure définition des frontières de phrases (*"sentence boundaries"*) que des approches basées sur des modèles de langue, pour des contenus de médias audiovisuels anglophones. D'autres travaux, basé sur des conversations jouées ou des enregistrements téléphoniques, mettent en avant l'importance des indices visuels (Bi & Swerts, 2017) ou lexicaux (Hjalmarsson, 2011; Oliveira, 2008) dans la gestion des tours de parole. Les travaux de Gambi *et al.* (2015); De Ruiter *et al.* (2006) concluent à l'absence d'impact de l'intonation pour l'anticipation des fins de tour de parole. Ces différentes études travaillent sur des matériaux et des styles de parole divers (parole lue ou conversations téléphoniques, énoncés élicités), et dans différentes langues (anglais, néerlandais, français, etc.) dont certaines montrent des performances divergentes sur ces aspects (Grosjean, 1996). Il n'est donc pas clair quels sont les indices (prosodie, syntaxe, lexique, mouvements, etc.) les plus pertinents pour permettre des transitions fluides entre locuteur·ices en français.

Dans quelle mesure est-on capable d'anticiper le moment propice de prise de parole, et quels indices jouent dans cette décision ? Cet article présente une expérience visant à mieux comprendre entre les indices lexicaux et les indices transmis par le signal de parole, lesquels participent le plus à la capacité des auditeur·ices à anticiper les TRP, sur des énoncés de parole spontanée. Cette évaluation est fondée sur des segments de parole extraits semi-automatiquement de contenus de médias télévisuels présentant des interactions spontanées.

La Section 2 explique les processus de sélection et d'évaluation des données ainsi que le paradigme expérimental. Les résultats de perception sont en suite détaillés dans la Section 3 et discutés dans le contexte de la littérature en Section 4.

2 Méthode

2.1 Données

La tâche de perception envisagée consistait à présenter aux participant·es des unités de parole éventuellement tronquées de quelques mots à la fin, et de leur demander si, à la fin du stimulus, il leur semble possible de prendre la parole sans couper celle de leur interlocuteur·ice.

2.1.1 Sélection des unités

Les pauses étant des indices importants pour la segmentation de la parole, nous avons choisi d'utiliser des *Inter Pausal Units* (IPU, segment de parole entre deux pauses) –utilisées pour une variété de tâches d'analyse et de traitement de la parole (Levitan & Hirschberg, 2011; Prakash & Murthy, 2019; Bigi & Priego-Valverde, 2019)– comme unité de base pour cette étude.

Les IPU ont été extraites de REPERE (Giraudel *et al.*, 2012), corpus composé d'émissions de TV diffusées en France entre 2011 et 2012. Nous avons sélectionné uniquement les programmes des émissions *BFMStory*, *EntreLesLignes* et *CaVousRegarde*, qui présentent des interactions conversationnelles. Une détection automatique des pauses a été préférée à une annotation manuelle afin de limiter les biais humains lors de la sélection, et se rapprocher de conditions d'une tâche finale entièrement automatique. Pour cela, une segmentation en locuteur a été réalisée avec `LIUMSpkDiarization` (Meignier & Merlin, 2010) et les segments obtenus (en supprimant tous ceux contenant de la parole superposée) ont été transcrits avec le système du LIUM basé sur Kaldi (Povey *et al.*, 2011).

Les IPU ainsi obtenues sont les segments maximaux d'un·e même locuteur·ice entre deux pauses silencieuses, telles que prédites par `LIUMSpkDiarization`.

2.1.2 Annotation

Afin de proposer aux participant·es un ensemble varié d'IPU et d'éviter de leur faire évaluer le même extrait dans des conditions différentes, nous avons choisi de proposer aux participant·es deux IPU présentant les mêmes caractéristiques selon les facteurs contrôlés suivants : *Genre* du ou de la locuteur·ice (2 possibilités), *TRP* ou non à la fin de l'IPU (2 possibilités), *Modalité* de présentation du stimulus (2 possibilités, audio ou textuelle), et *Coupure* de 0 à 3 mots prosodiques (Nespor & Vogel, 2007; Wheeldon & Lahiri, 2002) à la fin de l'IPU (4 possibilités). Ainsi, un total de 64 (2 × Genre × TRP × Modalité × Coupure) IPU sont nécessaires.

Parmi les IPU obtenus automatiquement comme décrit ci-dessus, nous avons sélectionné un sous-ensemble d'IPU d'une durée de 6 à 12 secondes afin de limiter la complexité de l'étude. En raison du nombre restreint de femmes présentes dans le corpus REPERE, la durée maximale a été augmentée à 19 s pour les locutrices. Un certain nombre d'IPU a été retiré car présentant des questions ou des exclamations –exemples intéressants de TRP mais induisant un biais– ainsi que celles contenant des backchannels.

Les IPU sélectionnées ont été annotées par trois co-auteur·ices de cette étude, qui devaient pour chacune indiquer si elle faisait partie d'une conversation ou d'un monologue (e.g., présentation de nouvelles, discours, ...). Iels ont aussi annoté le type de frontière à la fin de chaque IPU pour différencier les fins terminales (présence de TRP) ou non-terminales (absence de TRP). Un total de 172 segments ont été annotés, résultant en un accord inter-annotateur de $0,70$ pour le dialogue et $0,73$ pour la terminalité (en utilisant le Kappa de Fleiss (Fleiss, 1971)) ce qui montre un accord substantiel pour ces deux tâches. Seuls les segments pour lesquels les trois annotateur·ices étaient d'accord ont été gardés. Les 64 IPU sélectionnées ont ensuite été traités manuellement afin de corriger la transcription automatique et déterminer les frontières des trois derniers mots prosodiques.

Des 32 IPU annotés comme terminales 22 apparaissaient en fin de tour dans l'émission originale et 10 apparaissaient à l'intérieur d'un tour de parole (i.e., la personne conserve la parole après). À l'inverse, sur les 32 IPU non terminales, 12 apparaissaient en fin de tour et 20 au sein d'un tour de

```
mais vous savez étant donné que les convocations sont attendues
étant donné que euh on a l' habitude de mettre la pression par
voie de presse à mon avis d' ici là on va | se | voir | souvent
```

FIGURE 1 – Exemple de transcription d'un énoncé utilisé dans l'expérience, les frontières de mots prosodiques sont représentées par le symbole « | »

parole. Du fait du faible nombre de femmes représentées dans le corpus REPERE, moins de locutrices différentes (20) que de locuteurs (27) sont présentes dans les stimulus.

La Figure 1 présente un exemple d'IPU transcrite avec les frontières de mots prosodiques de la fin représentées par le symbole « | ». Pour chaque IPU sélectionné, 8 versions différentes sont générées en découpant aux quatre positions différentes (0, 1, 2, ou 3 mots découpés) et en présentant dans les deux modalités (texte seul ou audio seul). Au final, un ensemble de 256 stimulus dont les caractéristiques sont présentées en Table 1, est utilisé durant le test.

TABLE 1 – Durée et nombre de mots des IPU sélectionnés

	Durée (s)	Nombre de mots
min. (s)	5.0	19
max. (s)	18.6	55
moy. (s)	8.9	34.7

Le nombre de syllabes de chaque énoncé (Table 2) a été calculé suivant la méthode décrite par Adda-Decker *et al.* (2005). Alors que Grosjean (1996) utilise des coupes de 3 syllabes dans une expérience similaire en français, nous observons une moyenne de 2 syllabes pour les coupes au niveau du mot prosodique effectuées sur les stimulus de notre étude.

2.2 Test de perception

2.2.1 Paradigme expérimental

Ce test utilise l'interface web PsyToolkit (Stoet, 2010, 2016), permettant la réalisation d'expériences et questionnaires en navigateur. Kochari (2019); Sasaki & Yamada (2019); Strickland & Stoops (2018), entre autres, montrent que des tests psychologiques classiques, effectués en ligne obtiennent des résultats comparables en condition de laboratoire, avec une même puissance de test statistique. Ainsi, l'utilisation d'une interface web a été préférée pour simplifier la tâche de recrutement de participant·es aux profils plus variés que ceux recrutés au laboratoire (Woods *et al.*, 2015).

TABLE 2 – Nombre de syllabes coupées

	Nb moyen de syllabe
IPU complet	53.4
1er mot prosodique	2.1
2ème mot prosodique	3.8
3ème mot prosodique	5.7

Après un court texte expliquant le cadre de l'étude et la durée estimée du test (20 min), il était demandé aux participant·es d'indiquer leur âge, genre et langue maternelle. L'expérience commençait par la modalité audio, supposée plus simple et motivante.

Pour chaque stimulus, les participant·es devaient soit écouter soit lire l'extrait. Les boutons de réponses apparaissaient à la fin de l'écoute pour la modalité audio et après la moitié de la durée de l'extrait audio pour la modalité textuelle. Les participant·es avaient ensuite 30 secondes pour indiquer si une prise de parole sans interrompre le tour courant était possible *Maintenant*, *Bientôt* ou *Pas encore*. Il leur était demandé de répondre le plus rapidement possible en suivant leur intuition. Le stimulus suivant démarrait après une seconde de pause, une fois la réponse soumise ou si les 30 secondes étaient écoulées.

Un lien vers l'interface de test en ligne a été envoyé à différentes listes de diffusion, incluant des communautés de recherche en informatique et en sociologie, et a été partagé sur des réseaux sociaux. Un total de 53 personnes francophones (29 s'identifiant comme femme, 21 comme homme et 3 comme autre), sans déficit visuel ou auditif non corrigé ont volontairement participé à l'expérience. Leurs âges varient de 20 à 63 ans, avec une moyenne de 35 ans. Cinq personnes ont indiqué avoir une langue maternelle autre que le français.

Soixante-quatre stimulus ont été présentés à chaque participant·e afin que chacun des 64 énoncés originaux soient évalués une et une seule fois, mais dans des conditions de présentations variées selon un design en carré latin mélangeant les facteurs contrôlés suivants : (i) 4 coupures de mot prosodiques (entre 0 aucun mot coupé et 3 derniers mots coupés), (ii) 2 genres, (iii) 2 modalités de présentation et (iv) IPU terminal/non-terminal. Les niveaux de ces quatre facteurs sont répartis selon quatre groupes de participant·es ; les participant·es sont attribués aléatoirement à l'un des quatre groupes.

2.2.2 Traitement statistique

Les variations de la proportion de chacune des trois réponses possibles (*Maintenant*, *Bientôt*, *Pas encore*) ont été modélisées en fonction des facteurs suivants : présence ou non de TRP, nombre de mots coupés (0 à 3), modalité de présentation (audio/texte) et genre du ou de la locuteur·ice (H/F) – regroupant ainsi ensemble les stimulus présentant les mêmes caractéristiques.

Les réponses sont analysées grâce à une régression polynomiale (Gries, 2021) à l'aide de la bibliothèque `R nnet`(Venables & Ripley, 2003), prenant la proportion de chaque catégorie de réponse (*Maintenant*, *Bientôt*, *Pas encore*) comme variable dépendante, et les variables *TRP*, *Coupure*, *Modalité* et *Genre* comme facteurs indépendants. Ces quatre facteurs indépendants et leurs interactions forment un modèle maximal qui est ensuite soumis à une procédure de simplification (Crawley, 2013) en supprimant itérativement les interactions d'ordre supérieur, tant que cela ne dégrade pas significativement le modèle. L'interaction quadruple et les interactions triples prenant en compte le genre, ainsi que les interactions doubles *(TRP* × *Genre)* et *(Modalité* × *Genre)* ont ainsi été supprimées. Le modèle minimal adéquat est ainsi basé sur les quatre facteurs principaux, quatre interactions doubles (*(TRP* × *Coupure)*; *(TRP* × *Modalité)*; *(Coupure* × *Modalité)*; *(Coupure* × *Genre)*) et l'interaction triple *(TRP* × *Coupure* × *Modalité)*.

TABLE 3 – Table présentant la sortie du modèle minimal adéquat (Type III tests) présenté dans le texte : test du rapport de vraisemblance (LR χ^2), degrés de liberté (df), degrés de significativité (0.001 : '***' ; 0.01 : '**' ; 0.05 : '*').

Facteur	**LR χ^2**	**df**	**p**
TRP	214.23	2	***
Coupure	366.14	6	***
Modalité	62.71	2	***
Genre	18.29	2	***
(TRP × Coupure)	69.03	6	***
(TRP × Modalité)	74.46	2	***
(Coupure × Modalité)	14.61	6	*
(Coupure × Genre)	12.63	6	*
(TRP × Coupure × Modalité)	14.52	6	*

3 Résultats

Moins de $0,5\%$ des réponses n'ont pas été traitées parce que les participant·es avaient atteint la limite de 30 s de temps de réponse. Sur les réponses restantes, le modèle de régression montre un rôle significatif de tous les différents facteurs contrôlés lors de cette expérience. La table ANOVA correspondante est présentée dans la Table 3.

La réponse « Pas encore » est la plus fréquente (46%), suivie par « Bientôt » (32%) et enfin « Maintenant » (21%). Cela est lié au fait que les stimulus pour lesquels un changement de tour peut effectivement se produire sont en minorité, du fait des coupures de mots.

L'interaction entre TRP, Modalité et Coupure est décrite à la Figure 2. Cette interaction triple est présentée dans quatre graphiques de façon à montrer l'influence relative de chaque catégorie de réponse, sur les différentes coupures (en abscisse).

La réponse « Pas encore » est en effet prédominante, sauf pour quelques combinaisons de facteurs. Sa probabilité décroît à la coupure 0, pour toutes les combinaisons de modalité et de présence de TRP. Elle atteint ses plus bas niveaux pour les stimulus présentant un TRP, étant déjà plus faible que pour les stimulus sans TRP quel que soit le nombre de mots coupés. La probabilité de la réponse « Pas encore » dépend également de la modalité, montrant des variations plus importantes avec la modalité audio qu'avec le texte : elle est la plus forte, quelle que soit la position dans la phrase, pour les stimulus audio sans TRP, au-dessus de 60%, tandis qu'elle est la plus faible pour les stimulus audio avec TRP. Les présentations audio d'énoncés terminant par un TRP sont les seuls cas où la réponse « Pas encore » n'est jamais la plus probable : les participant·es ont répondu à plus de 50% « Bientôt » pour les coupures > 0, et « Maintenant » à 77% pour la coupure 0.

La réponse « Bientôt » est plutôt stable pour les stimulus sans TRP avec une probabilité autour de 25%, alors que la capacité d'anticipation change en fonction des mots coupés pour les stimulus avec TRP. Pour la modalité textuelle, sa probabilité est la plus forte pour un et deux mots coupés (autour de 40%, comparable aux réponses « Pas encore »), tandis que pour la modalité audio, il s'agit de la réponse la plus probable, autour de 50%, jusqu'aux énoncés complets (coupure=0 ; pour lesquels la réponse « maintenant » est choisie).

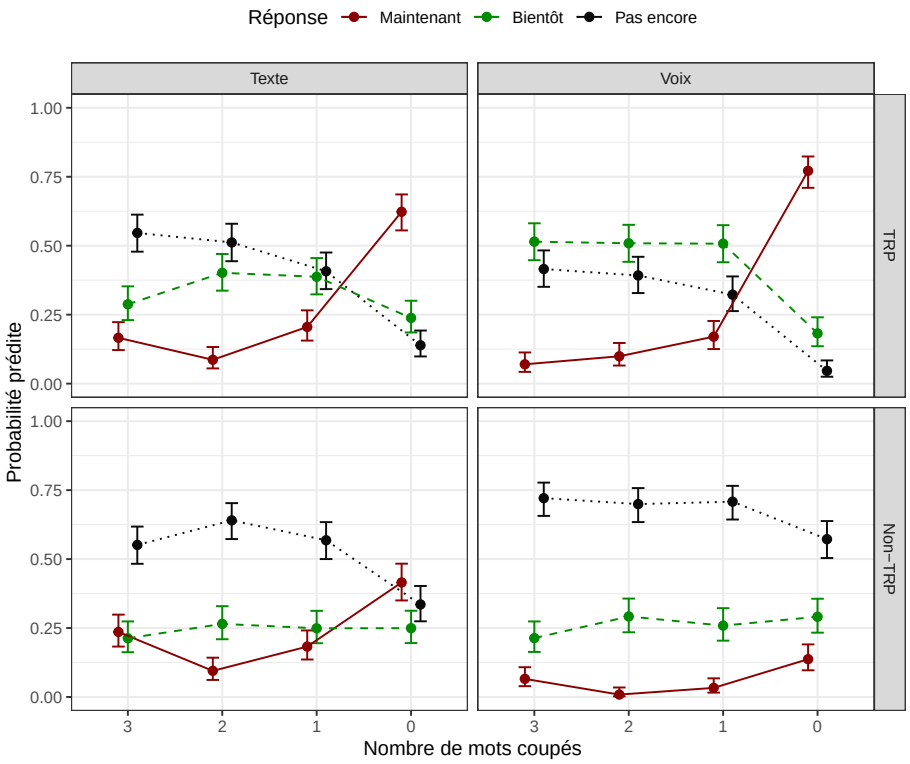


FIGURE 2 – Effet de la présence ou non de TRP (lignes), de la modalité de présentation (en colonne) et du nombre de mots coupés à la fin des IPU (abscisse) sur la probabilité des trois réponses possibles (« Maintenant » en rouge, « Bientôt » en vert, « Pas encore » en noir) estimée par le modèle.

La réponse « Maintenant » a un comportement opposé à la réponse « Pas encore », étant faible pour les coupures > 0 et augmentant pour les énoncés complets. Elle a aussi une probabilité plus élevée pour les énoncés complets avec TRP que sans, et surtout pour les énoncés en modalité audio : les stimulus complets audio sans TRP ont reçu 14% de réponse « Maintenant » contre 77% pour les ceux avec TRP. Pour la modalité textuelle, le taux de réponse passe de 42% pour les stimulus complets sans TRP à 62% pour ceux avec TRP.

Les inférences effectuées par les participants à propos de la terminalité des IPU sont plus tranchées pour les présentations orales que pour le texte. Avec la modalité textuelle, si la probabilité de la réponse « Maintenant » est plus faible pour les énoncés complets sans TRP qu'avec, il s'agit tout de même de la réponse la plus probable. À l'inverse, pour la modalité audio, les stimulus sans TRP ne sont presque jamais considérés comme pertinents pour une prise de parole, tandis que pour les stimulus avec TRP les participant·es anticipent la possibilité d'une prise de parole quel que soit le nombre de mots coupés (dans les limites de ce test).

L'interaction entre le nombre de mots coupés et le genre montre des différences à la coupure 0 (IPU complets), avec plus de réponses « Pas encore » et moins de « Maintenant » pour les énoncés produits par des hommes (18% « Pas encore » pour les femmes vs 28% pour les hommes, 53% « Maintenant » pour les femmes vs 45% pour les hommes). La probabilité de la réponse « Bientôt » augmente pour les femmes lorsque le nombre de mots coupés diminue. Cependant, ces observations sont les mêmes pour l'audio et le texte : il est possible que ce soit dû aux caractéristiques linguistiques et sémantique de ces énoncés plutôt qu'au genre, cette information n'étant pas disponible pour les stimulus textuels.

4 Discussion

L'effet principal sur la catégorie de réponse est lié à la triple interaction entre présence de TRP, nombre de mots coupés et modalité de présentation. S'il semble possible de décider qu'une IPU est terminée sur la base d'informations textuelles (hausse systématique des proportions « Maintenant » pour la coupure 0), distinguer entre IPU terminales ou non est bien mieux effectué si les participant·es ont accès aux informations de parole (proportion haute de « Bientôt » avant la fin des énoncés terminaux avant la coupure 0, puis réponse « Maintenant » très claire ; proportion élevée de « pas encore » pour les énoncés non terminaux, quelle que soit la coupure). Cela donne des arguments en faveur de l'importance des marques prosodiques pour la gestion du dialogue. Des indices de non-terminalité existent dans les deux modalités de présentation –il y a moins de réponses « Maintenant » pour les stimulus sans TRP dans les deux modalités– mais la présentation audio permet une meilleure distinction en fonction du nombre de mots coupés, sans confusion pour les énoncés complets : alors que les réponses « Maintenant » et « Pas encore » sont comparables pour les énoncés complets sans TRP présentés sous forme de texte, la réponse « Pas encore » n'est dominante que pour ceux présentés sous forme audio. Ainsi, les indices audio réduisent la probabilité de prendre la parole (réponses « Maintenant ») et augmentent la probabilité d'attendre (« Pas encore ») pour les séquences sans TRP.

À l'inverse, la prise de parole à la fin de séquences terminant par des TRP est plus probable pour les stimulus audio que textuels (62% de « Maintenant » pour le texte contre 77% pour l'audio). Les indices audio permettent aussi une meilleure anticipation : « Bientôt » n'est une réponse dominante que pour les stimulus audio, est l'est avant la fin de l'IPU. Les auditeur·ices peuvent prévoir au moins trois mots prosodiques (6 syllabes en moyenne) à l'avance s'il sera possible de prendre la parole. Ce résultat est cohérent avec les dynamiques de prise de parole présentée dans Levinson & Torreira (2015). Cette capacité d'anticipation n'est observée que pour la modalité audio.

L'augmentation de la réponse « Maintenant » est clairement liée aux énoncés avec 0 mots coupés, indiquant une forte capacité à déterminer qu'un énoncé est complet. Ceci est également observé pour la modalité textuelle, le gain permis par les indices audio est modeste (62% vs 77%). Les informations prosodiques auraient donc un rôle secondaire dans cette détermination. Elles sont par contre fondamentales pour l'anticipation –capacité nécessaire pour une prise de parole fluide–, résultat qui n'est pas reflété dans les réponses de ce test sur la seule base des contenus sémantiques. Nous montrons donc ici l'importance des indices audio pour le timing et l'efficacité de la gestion des tours de parole, sur des données dialogiques spontanées.

Si nous observons un effet du facteur *Genre* de la personne qui parle, le fait que cet effet soit visible aussi bien avec la modalité textuelle qu'audio pointe vers l'utilisation d'indices sémantiques ou syntaxiques plus que vers le genre perçu. Cette étude n'inclut pas un ensemble de locuteur·ices et de participant·es suffisamment large pour étudier l'impact des facteurs sociologiques tels que le rôle, le capital culturel ou la position sociale. Une étude multimodale incluant des indices visuels pourrait également être intéressante à l'avenir, le corpus d'origine étant issu de contenu télévisuel.

Remerciements

Ce travail a été partiellement financé par les projets ANR « Gender Equality Monitor » (ANR-19-CE38-0012) et ANR-DFG « La documentation automatique des langues à l'horizon 2025 » (CLD 2025, ANR-19-CE38-0015-04). Nous tenons à remercier les participant·es de cette étude.

Références

ADDA-DECKER M., BOULA DE MAREÜIL P., ADDA G. & LAMEL L. (2005). Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, **46**(2), 119–139. DOI : 10.1016/j.specom.2005.03.006.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BI R. & SWERTS M. (2017). A perceptual study of how rapidly and accurately audiovisual cues to utterance-final boundaries can be interpreted in chinese and english. *Speech Communication*, **95**, 68–77. DOI : 10.1016/j.specom.2017.07.002.

BIGI B. & PRIEGO-VALVERDE B. (2019). Search for Inter-Pausal Units : application to Cheese ! corpus. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 289–293, Poznań, Poland.

BOSCH L. T., OOSTDIJK N. & BOVES L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, **47**(1–2), 80–86. DOI : 10.1016/j.specom.2005.05.009.

CHRISTODOULIDES G. (2018). Acoustic correlates of prosodic boundaries in french a review of corpus data / correlatos acústicos de fronteiras prosódicas em francês : uma revisão de dados de corpora. *REVISTA DE ESTUDOS DA LINGUAGEM*, **26**(44), 1531–1549. DOI : 10.17851/2237-2083.26.4.1531-1549.

CRAWLEY M. J. (2013). *The R Book*. John Wiley & Sons, 2 édition.

DE RUITER J., MITTERER H. & ENFIELD N. (2006). Projecting the end of a speaker's turn : A cognitive cornerstone of conversation. *Language*, **82**, 515–535. DOI : 10.1353/lan.2006.0130.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382. DOI : 10.1037/h0031619.

GAMBI C., JACHMANN T. & STAUDTE M. (2015). The role of prosody and gaze in turn-end anticipation. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

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

GOTOH Y. & RENALS S. (2000). Sentence boundary detection in broadcast speech transcripts. In *in Proc. of ISCA Workshop : Automatic Speech Recognition : Challenges for the new Millennium ASR-2000*, p. 228–235.

GRIES S. T. (2021). *Statistics for linguistics with R*. Mouton Textbook. Berlin, Germany : De Gruyter Mouton, 3 édition.

GROSJEAN F. (1996). Using prosody to predict the end of sentences in english and french : Normal and brain-damaged subjects. *Language and Cognitive Processes*, **11**(1–2), 107–134. DOI : 10.1080/016909696387231.

HJALMARSSON A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, **53**(1), 23–35. DOI : 10.1016/j.specom.2010.08.003.

KOCHARI A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, **2**(1), 39. DOI : 10.5334/joc.85.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEVINSON S. & TORREIRA F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, **6**. DOI : 10.3389/fpsyg.2015.00731.

LEVINSON S. C. (2016). Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences*, **20**(1), 6–14. DOI : 10.1016/j.tics.2015.10.010.

LEVITAN R. & HIRSCHBERG J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH*. DOI : 10.7916/D8V12D8F.

MAGYARI L. & DE RUITER J. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, **3**.

MEIGNIER S. & MERLIN T. (2010). Lium spkdiarization : an open source toolkit for diarization. In *CMU SPUD Workshop*.

NESPOR M. & VOGEL I. (2007). *Prosodic Phonology*. DE GRUYTER. DOI : 10.1515/9783110977790.

OLIVEIRA M. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. *Speech Prosody*, p.4̃.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLIC˘EK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. p.4̃.

PRAKASH J. J. & MURTHY H. A. (2019). Analysis of inter-pausal units in indian languages and its application to text-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(10), 1616–1628. DOI : 10.1109/taslp.2019.2924534.

SACKS H., SCHEGLOFF E. A. & JEFFERSON G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, **50**(4), 696. DOI : 10.2307/412243.

SASAKI K. & YAMADA Y. (2019). Crowdsourcing visual perception experiments : a case of contrast threshold. *PeerJ*, **7**, e8339. DOI : 10.7717/peerj.8339.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SKANTZE G. (2021). Turn-taking in conversational systems and human-robot interaction : A review. *Computer Speech & Language*, **67**, 101178. DOI : 10.1016/j.csl.2020.101178.

STIVERS T., ENFIELD N. J., BROWN P., ENGLERT C., HAYASHI M., HEINEMANN T., HOYMANN G., ROSSANO F., DE RUITER J. P., YOON K.-E. & LEVINSON S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, **106**(26), 10587–10592. DOI : 10.1073/pnas.0903616106.

STOET G. (2010). PsyToolkit : A software package for programming psychological experiments using linux. *Behavior Research Methods*, **42**(4), 1096–1104. DOI : 10.3758/brm.42.4.1096.

STOET G. (2016). PsyToolkit. *Teaching of Psychology*, **44**(1), 24–31. DOI : 10.1177/0098628316677643.

STRICKLAND J. C. & STOOPS W. W. (2018). Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *Journal of the Experimental Analysis of Behavior*, **110**(1), 136–153. DOI : 10.1002/jeab.445.

VENABLES W. N. & RIPLEY B. D. (2003). *Modern applied statistics with S*. Statistics and Computing. New York, NY : Springer, 4 édition.

WHEELDON L. R. & LAHIRI A. (2002). The minimal unit of phonological encoding : prosodic or lexical word. *Cognition*, **85**(2), B31–B41. DOI : 10.1016/S0010-0277(02)00103-8.

WOODS A. T., VELASCO C., LEVITAN C. A., WAN X. & SPENCE C. (2015). Conducting perception research over the internet : a tutorial review. *PeerJ*, **3**, e1058. DOI : 10.7717/peerj.1058.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Frontières entre la perception de la voix normophonique et pathologique chez des auditeurs naïfs

Amelia Pettirossi[1] Nicolas Audibert[1] Lise Crevier-Buchman[1,2]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), Paris, France
(2) Hôpital Foch : Service de Laryngologie Phoniatre, Suresnes, France
amelia.pettirossi@sorbonne-nouvelle.fr, nicolas.audibert@sorbonne-nouvelle.fr, lise.buchman1@gmail.com

RÉSUMÉ

Nous étudions l'hétéroévaluation de la qualité de voix chez une population de femmes francophones comprenant des professionnelles de la voix et des patientes recrutées lors d'une consultation de phoniatrie. Des extraits de parole ont été évalués par un expert à l'aide du GRBAS et par deux groupes de 30 et 40 auditeurs naïfs avec des protocoles perceptifs destinés à évaluer la qualité de voix globale. Ces évaluations ont été mises en correspondance afin d'observer de potentielles corrélations entre la perception de la pathologie vocale chez les auditeurs naïfs et les paramètres du GRBAS expert. Les voix perçues comme plus pathologiques par les auditeurs naïfs sont bien associées à un grade global plus élevé, avec une influence prépondérante de la sévérité du serrage vocal. Les indices de raucité et de souffle sont moins liés à la perception par les naïfs des voix comme pathologiques.

ABSTRACT

Boundaries between the perception of normophonic and pathological voice by naive listeners
We study hetero evaluation of voice quality in a population of French female speakers including voice professionals and patients recruited during a phoniatric consultation. Speech samples were evaluated by an expert using the GRBAS and by two groups of 30 and 40 naïve listeners with perceptual protocols to assess overall voice quality. These assessments were matched to observe potential correlations between perception of vocal pathology in naive listeners and expert GRBAS parameters. Voices perceived as more pathological by naive listeners are well associated with a higher overall grade, with a predominant influence of vocal strain severity. The roughness and breathiness dimensions are less related to naive listeners' perception of voices as pathological.

MOTS-CLÉS : voix, dysphonie, hétéro-perception, auditeurs experts, auditeurs naïfs
KEYWORDS : voice, dysphonia, hetero-perception, expert listeners, naive listeners

1 Introduction

Les normes culturelles peuvent faire varier différents aspects de l'évaluation de la qualité de voix, même ceux supposés s'appuyer sur des références solides comme l'évaluation perceptive experte des dysphonies. Il a été observé que des professionnels italiens et français, évaluant des locuteurs dysphoniques de ces deux nationalités, sont en accord pour ce qui est du grade de sévérité général de la dysphonie, mais pas de la raucité, sous-cotée par les experts italophones en comparaison avec

les francophones (Ghio et al., 2011 ; 2015). Cet écart semble lié au fait que l'auditeur, bien qu'expert, est directement influencé par ses représentations sociales, c'est-à-dire à ce qui est localement et momentanément validé au sein d'une culture donnée (Rouquette et al., 1998). Ces représentations sociales viennent s'ajouter à une expérience personnelle et une éventuelle formation pour venir créer un référent interne propre à chaque auditeur.

Afin de quantifier les altérations de la qualité de voix, il existe des outils acoustiques composites tel que l'AVQI (Acoustic Voice Quality Index) mais également des outils perceptifs qui sont en général considérés comme la référence (*gold standard*) du diagnostic de la dysphonie. Bien que des revues de littérature mettent en évidence qu'il existe une grande variété de protocoles perceptifs pour l'hétéroévaluation de la qualité de voix pathologique, le GRBAS de Hirano (1981) reste l'un des protocoles les plus utilisés en recherche et en clinique (Kreiman et al., 1993 ; De Bodt et al., 1996 ; Suhail et al., 2016 ; Schuering et al., 2021). Le GRBAS permet de quantifier le grade général de dysphonie (G), la raucité (R), le souffle (B), l'asthénie (A) et le serrage vocal (S). Chaque critère pouvant être coté de 0 (aucun trouble) à 3 (trouble sévère).

Si nous savons que de manière générale les voix dysphoniques induisent des jugements plus négatifs que les voix saines sur différents aspects comme l'attractivité physique (Blood et al., 1979) ou la personnalité (Amir et al., 2013) il est aussi vrai que tout ce qui est catégorisé comme pathologique pour un expert n'est pas forcément associé à une représentation négative chez les auditeurs naïfs. Il a déjà été démontré que dans une population francophone, une légère raucité chez l'homme peut être considérée comme séduisante pour des auditrices alors que cette même raucité est considérée comme pathologique par un expert (Barkat-Defradas & al, 2012). Cette conclusion a également été vérifiée pour les femmes francophones, les locutrices ayant une raucité légère sont évaluées de manière plus positive par des auditeurs des deux sexes (Pettirossi et al., 2020). Ces observations sont facilement explicables par l'idée selon laquelle « *What sounds beautiful is good.* », largement mise en évidence dans les études sur le jugement vocal (Zuckerman & al, 1988).

Il semble nécessaire de comprendre où se place la limite entre une voix qui sera perçue comme pathologique par un auditeur naïf et qui engendrera une image détériorée du locuteur, d'une voix normophonique. Il pourrait être pertinent de prendre en compte qu'en l'absence de plainte physique ou fonctionnel, tous les paramètres perceptifs utilisés par les experts du domaine médical ou paramédical n'évoquent pas nécessairement la pathologie chez les naïfs.

Pour autant, comme susmentionné, l'évaluation des auditeurs naïfs peut être sévère et provoquer certains handicaps sociaux. Il a déjà été mis en évidence que des auditeurs ayant reçu des fiches explicatives visant à les sensibiliser aux troubles vocaux évaluent aussi sévèrement, malgré ce dispositif préalable, des locutrices atteintes de dysphonie et d'hyper-nasalité qu'un deuxième groupe d'auditeurs naïfs maintenu volontairement dans la méconnaissance de la pathologie vocale (Lallh et al., 2000). Les locutrices atteintes de troubles vocaux sont catégorisées comme peu amicales, ennuyeuses et peu attractives, ce qui peut évidemment être pénalisant dans divers contextes sociaux comme les entretiens d'embauche ou même les relations amicales. L'une des principales conclusions de ces travaux est donc que les cliniciens pourraient avoir besoin de faire de la prévention auprès de leurs patients quant aux attitudes sociales négatives des interlocuteurs.

Nous savons également que les auditeurs naïfs sont peu sensibles aux dysphonies légères (G1, G2), et n'obtiennent respectivement que 49% et 38% de bonnes catégorisations lors de l'évaluation de ces voix sur le paramètre G du GRBAS en comparaison avec une évaluation experte (Ghio et al.,

2011). A l'inverse, les dysphonies sévères (G3) obtiennent 68% de bonnes réponses et 86% pour les voix normophoniques (G0), ces résultats sont vérifiables même après un entrainement des naïfs.

D'autre part, certaines études traitent de la différence de perception de l'altération de la qualité de voix entre auditeurs experts (orthophonistes et oto-rhino-laryngologistes) et naïfs en termes d'accord inter-juges. Sans étonnement nous observons généralement de meilleurs degrés d'accord chez les experts (Helou et al., 2010 ; Misono et al., 2012) mais parfois un accord inter-juges aussi élevé pour les experts que les naïfs (Eadie et al., 2010). Ces approches mettent finalement en évidence qu'il y aurait un référent interne plus similaire chez les experts. Évidemment, il est légitime de penser que les naïfs, qui ont une expérience personnelle n'incluant pas de formation commune en santé, auront un référent interne plus variable que les experts.

Dans cet article, nous tenterons de répondre à la question suivante : où se trouve la frontière, aussi ténue soit-elle, entre ce qui est perçu comme une qualité de voix normophonique ou pathologique par un panel d'auditeurs naïfs ? Nous nous baserons sur la perception générale de la qualité de voix par des naïfs par rapport à la cotation des paramètres perceptifs du GRBAS réalisée par une médecin phoniatre à travers deux groupes de locutrices, l'un ayant été recruté dans le cadre d'une consultation de phoniatrie et l'autre étant composé de locutrices professionnelles de la voix volontaires.

2 Méthodologie

2.1 Locutrices, corpus et prises de données

Nous avons ici deux panels distincts. Nous nous réfèrerons à notre premier panel comme « groupe patientes », il s'agit ici de 10 femmes francophones (29-78 ans) recrutées dans le cadre d'une consultation phoniatrique avec laryngostroboscopie. Leurs métiers peuvent être variés, bien que nous retrouvions 9 professionnelles de la voix dont 5 professeures des écoles. Les critères d'inclusion de ce panel comprenaient le fait d'être diagnostiquées d'une dysphonie lors de la consultation et de ne pas avoir été opéré du larynx précédemment, ni avoir subi de traitement radio-chimiothérapique. Les atteintes du groupe patientes sont multiples : kyste, nodule, laryngite, granulome, œdème de Reinke, reflux, ou encore paralysie récurrentielle.

Nous appellerons le deuxième panel « groupe PE » car il est constitué de 61 femmes francophones, professeures des écoles (PE) en activité (23-61 ans). Il s'agit ici uniquement de volontaires tout-venant. Pour autant, nous nous attendons à recruter des locutrices avec des dysphonies dysfonctionnelles car d'après une étude menée pour l'INSERM la prévalence dans cette population, majoritairement féminine, est très élevée (Autesserre et al., 2006). En 2017, une enquête menée sur 709 femmes professeures des écoles travaillant en France a mis en évidence que 80% de la population auto-déclare des troubles vocaux alors que seulement 18% des PE ont déjà consultée un orthophoniste ou un ORL (Pettirossi, 2021). Les critères d'inclusion de ce panel comprenaient également de ne pas avoir été opéré du larynx précédemment, ni avoir subi de traitement radio-chimiothérapique.

Toutes les participantes ont été enregistrées à partir de la station Computerized Speech Lab 4500 de KayPENTAX avec un micro-casque AKG C 410 positionné à environ 5 cm des lèvres. Les

différentes évaluations perceptives de cette étude ont été réalisées à partir de phrases lues ou de la lecture du texte « La bise et le soleil ».

2.2 Hétéroévaluation de la qualité de voix

Dans le cadre de notre hétéroévaluation experte de la qualité de voix, une même phoniatre a réalisé une cotation GRBAS en direct pour le panel « patientes » et en différé avec une interface sur ordinateur réalisée avec Praat (Boersma et al., 2023) pour le panel « PE », dans les deux cas à partir d'un extrait de parole lue. Nous observons, des dysphonies beaucoup plus sévères dans notre groupe patientes et de nombreuses dysphonies légères dans notre groupe PE (Table 1).

	Groupe « patientes » (n = 10)				Groupe « PE » (n=61)			
Cotation experte	0	1	2	3	0	1	2	3
G (n)	0	2	3	5	37	22	2	0
R (n)	3	3	1	3	41	19	1	0
B (n)	2	5	3	0	52	9	0	0
A (n)	9	0	1	0	61	0	0	0
S (n)	4	2	3	1	57	4	0	0

TABLE 1 : Récapitulatif du nombre de locutrices dans chaque panel cotée 0, 1, 2 ou 3 sur les différents paramètres perceptifs du GRBAS par une experte phoniatre

Quant à eux, les auditeurs naïfs ont évalué la qualité de voix des locutrices sur des extraits de lecture de « La bise et le soleil » à l'aide d'interfaces générées avec le logiciel Praat (Boersma et al., 2023) avec des variations mineures dans les protocoles d'évaluation appliqués aux deux panels de locutrices. Nous comptons 30 auditeurs pour l'évaluation du panel patientes et 40 pour le panel PE parmi lesquels 15 auditeurs sont communs aux deux tâches, avec environ deux ans d'écart entre les deux évaluations.

Pour le panel patientes, les auditeurs ont dû répondre à la question « Pensez-vous que cette personne a des problèmes de voix ? » à l'aide de 4 boutons « Absolument pas », « Probablement pas », « Probablement », « Absolument ». À la fin de cette expérimentation les réponses ont été cotées 0 pour « Absolument pas », jusqu'à 3 pour « Absolument ».

Enfin, pour le panel PE, les évaluations ont été réalisées sur une échelle sémantique différentielle comprenant 5 échelons avec les mentions « Aucun trouble vocal » et « trouble vocal sévère » à chaque extrémité. Les réponses ont été cotées de 0 pour « Aucun trouble vocal » à 4 pour « trouble vocal sévère ».

3 Résultats

Nous réalisons ici des corrélations de Spearman pour comparer les évaluations de la qualité de voix globales réalisées par les auditeurs naïfs avec la cotation experte du GRBAS mais également entre le grade de dysphonie (G) et les autres dimensions (R, B, A et S) de l'évaluation experte (Table 2).

Corrélation avec les paramètres GRBAS	G expert : Coefficient ρ (Valeur-p)	Naïfs « patientes » : Coefficient ρ (Valeur-p)	Naïf « PE » : Coefficient ρ (Valeur-p)
G		0.822 (**0.003**)	0.630 (<**0.0001**)
R	0.786 (<**0.0001**)	0.227 (0.528)	0.579 (<**0.0001**)
B	0.653 (<**0.0001**)	0.508 (0.134)	0.391 (**0.002**)
A	0.180 (0.132)	-0.175 (0.630)	n/a
S	0.605 (<**0.0001**)	0.892 (**0.0003**)	0.304 (**0.02**)

TABLE 2 : Corrélation et valeur-p entre le Grade expert et les autres dimensions du GRBAS

Nous effectuons également des comparaisons entre ces corrélations expertes et naïves pour les dimensions perceptives R, B, A et S (Figure 1). Ces comparaisons sont effectuées à l'aide du package *cocor* (Diedenhofen et al., 2015) implémenté dans R (R Core Team, 2024), avec la méthode de Meng et al. (1992).

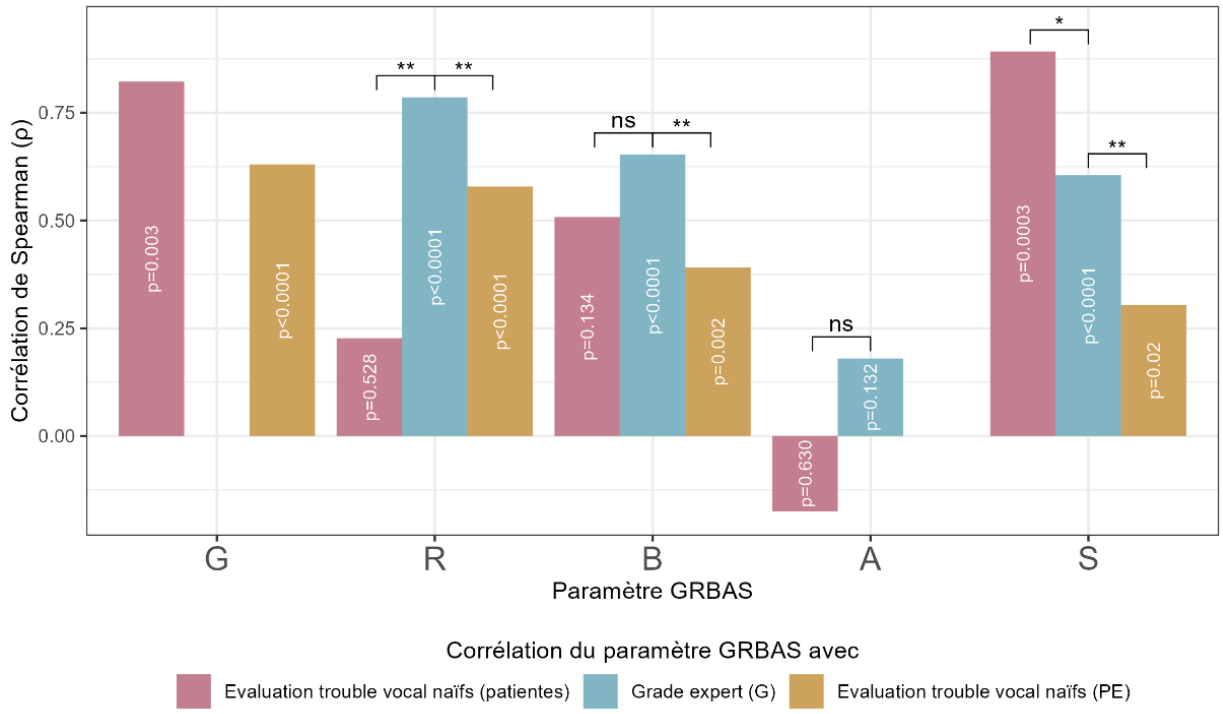


FIGURE 1 : Corrélation et valeur-p entre le Grade expert et les autres dimensions du GRBAS toutes locutrices confondues (bleu) et entre le GRBAS et le score moyen de l'évaluation globale de la qualité de voix naïve pour chaque groupe de locutrices (rose et ocre). Les accolades indiquent la significativité de la différence entre corrélations au sens de Meng et al. (1992).

Tout d'abord, nous obtenons une corrélation élevée entre le grade général (G) coté par l'expert et la raucité (R), des corrélations moyennes entre G et le souffle (B) et le serrage vocal (S) mais aussi une absence de corrélation entre G et l'asthénie (A). Si les auditeurs naïfs sont sensibles aux mêmes dimensions perceptives que les experts, ils devraient donc être plus influencés par la raucité, et dans une moindre mesure par le souffle, le serrage vocal et enfin l'asthénie.

Les corrélations entre le GRBAS expert et l'évaluation globale de la qualité de voix par les naïfs sont toutes plus modérées pour notre groupe PE. Ce résultat n'est pas étonnant car les dysphonies sont beaucoup plus légères que dans le groupe patientes. En revanche, il semble y avoir une sensibilité des auditeurs naïfs quant à « l'impression générale » d'altération des voix car le grade de sévérité coté par l'expert est fortement corrélé avec l'évaluation globale de la qualité de voix faite par les naïfs pour les patientes et modérément pour le groupe PE. Notons également des divergences presque systématiques entre les corrélations des dimensions R, B, A et S expert avec le G expert et l'évaluation globale de la qualité de voix par les auditeurs naïfs, la plupart des différences entre naïfs et expert étant significatives. Cela suggère que les dimensions perceptives prises en compte par les naïfs et les experts pour évaluer la pathologie vocale seraient foncièrement différentes.

Tout d'abord, pour ce qui est de la raucité, la corrélation avec l'évaluation globale de la qualité de voix réalisée par les naïfs est faible pour le groupe patiente et très modérée pour le groupe PE. Ainsi, la raucité semble plus impacter la perception de la qualité de voix globale dans notre population tout-venant comprenant des locutrices témoins comme des dysphonies légères plutôt que dans notre panel de patientes aux dysphonies plus sévères. Nous observons également une différence significative entre les corrélations experte et naïves pour le groupe patiente ($z=-3.071$, $p=0.002$), comme le groupe PE ($z=-2.809$, $p=0.005$). Ce résultat suggère que la raucité est plus prise en compte par les experts que par les naïfs pour catégoriser une voix comme pathologique.

La présence de souffle ne semble que très peu corrélée avec la perception de la pathologie vocale par les naïfs dans nos deux groupes de locutrices. Pour ce qui est de la comparaison entre les corrélations experte et l'évaluation naïve, nous observons une différence significative chez les PE ($z=-2.852$, $p=0.004$), mais pas chez les patientes ($z=-0.812$, $p=0.417$). Bien que ce résultat soit plus mitigé, l'expert semble plus influencé par le souffle que les naïfs pour évaluer la dysphonie.

Nous observons une absence totale de corrélation entre la perception d'une asthénie par l'expert et de la qualité de voix par les auditeurs naïfs. Rappelons tout de même que l'asthénie n'est présente chez aucune de nos locutrices du groupe PE et qu'une seule du groupe patiente en est atteinte. Cela limite grandement la portée de ce résultat, d'autant plus que l'asthénie est souvent le paramètre du GRBAS pour lequel l'accord inter-juges est le plus faible même chez des experts (Hidaka et al., 2022; Yamaguchi et al., 2003; Sellars et al., 2009). La comparaison entre les corrélations experte et naïves ne laisse apparaitre aucune différence significative chez les patientes ($z=-1.568$, $p=0.117$).

Quant au serrage vocal, il semble toujours assez fortement corrélé à la perception de la pathologie vocale chez les auditeurs naïfs. En effet, dans notre groupe patiente la corrélation de cette dimension avec l'évaluation de la qualité de voix globale des naïfs est élevée et dépasse même celle du grade de sévérité général défini par l'expert. De plus, la différence entre les corrélations experte et naïve est significative pour le groupe patientes ($z=2.240$, $p=0.025$) laissant entendre que le serrage vocal est plus considéré comme marqueur de voix pathologique par les naïfs qu'il ne l'est par l'expert. Étonnamment, une corrélation plus faible de l'évaluation naïve du trouble vocal avec le serrage vocal est observée pour le groupe PE, à un niveau significativement inférieur à la corrélation experte ($z=3.095$, $p=0.002$). Ce résultat peut néanmoins s'expliquer par le fait que nous n'avons que quatre locutrices avec un serrage vocal léger dans ce panel.

Enfin, nous notons un accord inter-juges modéré pour l'évaluation naïve de la qualité de voix globale des locutrices patientes (corrélation intraclasse ICC de 0.633) alors que ce même accord est faible pour le groupe PE (ICC de 0.201). Évidemment, ce résultat est attendu car des locutrices avec des

troubles dysphoniques plus sévères se trouvent dans le groupe patientes contrairement au groupe PE qui comprend des dysphoniques légères ainsi que des locutrices normophoniques.

4 Discussion et conclusion

Nos résultats confirment que la perception de la dysphonie et ce qui la caractérise n'est pas uniforme. En effet, la délicate limite entre la perception de la voix normophonique et de la voix pathologique a déjà été régulièrement mise en avant et commence évidemment par une simple question : qu'est-ce que la normophonie, qu'est-ce qu'une voix « normale » ? Nous ne sommes pas en mesure d'évaluer ce qui constitue une « déviation » sans une norme supposée, ce qui est loin d'être aisé puisqu'il n'y a pas de consensus sur la définition même de la qualité de voix (Kreiman et al., 2011). Une étude récente a mis en lumière que des auditeurs invités à classer 20 voix de femmes de la plus « normale » à la plus « non-normale » montrent un degré d'accord limité pour ce qui est de classer les voix « anormales » et un degré d'accord encore plus faible pour les voix « normales » (Kreiman et al., 2020). Dans cette étude, parmi les mesures acoustiques prises en compte, la fréquence fondamentale (f0) ainsi que le premier et second formant sont celles qui prédisent le mieux l'évaluation des auditeurs et non des paramètres plus spécifiques à la qualité de voix comme les mesures d'apériodicités ou de différence d'amplitude entre les premières harmoniques.

Il est important de prendre en compte qu'une voix peut être altérée par de nombreux biais comme un souffle, une raucité, une hyper ou encore hypo-fonctionnalité et que ces dimensions ne sont pas nécessairement indépendantes les unes des autres (Ghio et al., 2011). Encore une fois, la raucité semble être une dimension problématique puisqu'il a déjà été mis en lumière que son évaluation influence grandement la cotation du souffle, de manière générale, une forte interdépendance perceptive semble lier ces deux dimensions vocales (Kreiman et al., 1994). La littérature indique aussi que la perception de la pathologie n'est pas nécessairement constante, même lorsqu'elle est réalisée par des experts. En effet, l'évaluation de la raucité peut être catégorisée avec toutes les cotations possibles pour une même voix (Kreiman et al., 1993). La difficulté à isoler les dimensions perceptuelles est une des raisons pour lesquelles nous ne retrouvons pas un très bon accord entre les auditeurs sur la perception de la qualité de voix. A l'inverse, il existe un meilleur consensus sur la qualité de voix dans un corpus composé de voix de synthèse ne variant que par la f0 par rapport à des échantillons de voix naturelles (Kreiman et al., 2000). Il semble même que parmi diverses mesures acoustiques objectives, seule la f0 soit un indice robuste et presque unanimement utilisé par les auditeurs lors de l'évaluation de la qualité de voix d'un locuteur (Kreiman et al., 2020, 1992).

Cette grande variabilité dans l'évaluation perceptive de la qualité de voix semble être une conséquence de l'image que chacun se fait de ce qui est pathologique selon son expérience et sa culture. L'accord inter-juge modéré mis en évidence dans nos résultats même sur un panel de locutrices avec des dysphonies sévères en est une preuve. De plus, cette représentation n'est pas non plus uniforme parmi les experts. En effet, l'idée même d'un « espace perceptuel commun » a été remis en question par les divergences observées entre auditeurs et entre études (Kreiman et al., 1996).

Selon nos résultats, pour catégoriser une voix comme pathologique, les auditeurs naïfs sont particulièrement sensibles au-delà du grade général de dysphonie (G) à la perception du serrage vocal (S). Nous avons par ailleurs mis en évidence dans le cadre de travaux non-publiés sur le groupe patientes qu'il y a une forte corrélation positive entre la sévérité du serrage vocal d'une voix et le fait que les auditeurs naïfs la considèrent comme étant plus dérangeante à écouter. Les voix avec un

fort serrage vocal sont également connues pour entrainer un plus gros effort d'écoute chez des auditeurs naïfs (Farahani et al., 2020). Ce retour d'impression « dérangeant » ou « fatigant » pourrait provenir du fait que le serrage vocal donne l'impression de forcer sur la voix, une altération du comportement vocal qui serait perçue plus facilement et plus négativement par les naïfs.

En revanche, nous obtenons des corrélations très modérées à faibles entre la catégorisation d'une voix comme étant pathologique et la raucité (R) mais aussi le souffle (B). L'expert semble plus influencé par ces deux dimensions vocales que les naïfs pour évaluer une voix comme dysphonique. Pour le souffle, ces résultats pourraient nous mener à mettre en rapport l'évaluation par les naïfs et les représentations sociales liées à la voix soufflée. En effet, cet indice pourrait également faire partie de ceux qui sont perceptivement appréciés pour les femmes francophones comme c'est le cas pour les femmes anglophones aux États-Unis (Babel et al., 2014). Pour ce qui est de la raucité, ces résultats sont probablement en lien direct avec les représentations sociales et les référents internes des auditeurs naïfs qui semblent apprécier une raucité légère pour les femmes francophones. Nos résultats, ainsi que ceux de la littérature mise en avant dans le cadre théorique, tendent à montrer que la raucité n'est pas considérée comme pathologique par les naïfs francophones. Beaucoup moins de résultats existant sur le souffle, il pourrait être intéressant de réaliser une étude concernant plus précisément cette dimension vocale.

Enfin, au vu des résultats de la littérature et de ceux que nous apportons, il pourrait être intéressant de reconsidérer certains traits perceptifs comme relevant nécessairement de la pathologie, si toutefois ils ne font pas l'objet d'une plainte ou d'une atteinte organique. L'exemple de la raucité est particulièrement frappant puisqu'elle est visiblement dans la culture française, considérée comme vocalement attractive chez les hommes comme chez les femmes, du moins dans le cas d'une raucité légère et n'est que peu prise en compte par les auditeurs naïfs lors de l'évaluation d'une voix comme pathologique. Doit-on alors nécessairement considérer cette dimension comme un indice de pathologie ? L'idée selon laquelle une dimension perceptive pourrait être reconsidérée selon la norme sociale en cours au moment et à l'endroit de l'évaluation vocale pourrait être appuyé par certains arguments, comme le fait qu'en Angleterre, les femmes peuvent moduler leur voix chantée pour augmenter leur attractivité vocale en utilisant une qualité de voix plus rauque (Barkat-Defradas et al., 2013). Dans ce contexte, la raucité n'est pas la conséquence directe d'une altération de la qualité de voix mais d'une volonté de se rapprocher artificiellement d'une image vocale socialement validée.

Pour conclure, les voix perçues comme plus pathologiques par les auditeurs naïfs semblent être celles avec un grade de sévérité global plus élevé mais également celles avec un serrage vocal plus sévère. Cet indice est, selon nos résultats, plus important dans l'identification d'une pathologie vocale pour le jury naïf comparé au jury expert. Certaines autres dimensions comme la raucité ou le souffle semblent être plus à la frontière entre la perception de la voix pathologique et normophonique. Ces écarts entre experts et naïfs proviennent probablement d'un référent interne différent ainsi que de représentations sociales ancrées et hautement variable d'un individu à l'autre. Ce référent interne reste connu pour être plus commun entre experts qu'entre naïfs car bâti sur le nombre d'année d'exercice, ce qui permet aux experts de rattacher ce qu'ils perçoivent à ce qu'ils imaginent du mode de fonctionnement des plis vocaux.

Remerciements

Ce travail a été soutenu par le Laboratoire d'Excellence Empirical Foundations of Linguistics (LabEx EFL, ANR-10-LABX-0083). Il contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

AMIR, O., & LEVINE-YUNDOF, R. (2013). Listeners' Attitude Toward People With Dysphonia. *Journal of Voice*, 27(4), 524. https://doi.org/10.1016/j.jvoice.2013.01.015

AUTESSERRE D, CHARPY N, CREVIER-BUCHMAN L, et al. La Voix: Ses Troubles Chez Les Enseignants. Paris: Les éditions INSERM; 2006. https://hal-lara.archives-ouvertes.fr/hal-01570681v1

BABEL, M., MCGUIRE, G., & KING, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS one*, 9(2), e88616. https://doi.org/10.1371/journal.pone.0088616

BARKAT-DEFRADAS, M., BUSSEUIL, C., CHAUVY, O., HIRSCh, F., FAUTH, C., & RÉVIS, J. (2012). Dimension esthétique des voix normales et pathologiques : Approches perceptive et acoustique. TIPA. *Travaux interdisciplinaires sur la parole et le langage*, (28), 2-15. https://shs.hal.science/halshs-00778795

BARKAT-DEFRADAS, M., FAUTH, C., HIRSCH, F., AMY DE LA BRETÈQUE, B., SAUVAGE, J., & DODANe, C. (2013). Rauque « n » Roll : La raucité, entre symptôme pathologique & expression artistique. *Présenté à 5° Journées de Phonétique Clinique*, Liège, Belgium. https://hal.univ-lorraine.fr/hal-00918332

BLOOD, G. W., MAHAN, B. W., & HYMAN, M. (1979). Judging personality and appearance from voice disorders. *Journal of Communication Disorders*, 12(1), 63-67. https://doi.org/10.1016/0021-9924(79)90022-4

BOERSMA, P., & WEENINK, D. (2023). Praat: Doing phonetics by computer [Computer program]. Version 6.1.16, retrieved 15 December 2023 from http://www.praat.org/.

DE BODT, M. S., VAN DE HEYNING, P. H., WUYTS, F. L., & LAMBRECHTS, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto-Rhino-Laryngologica Belgica*, 50(4), 283-291.

DIEDENHOFEN, B. & MUSCH, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*, 10(4): e0121945. doi:10.1371/journal.pone.0121945

EADIE, T. L., KAPSNER, M., ROSENZWEIG, J., WAUGH, P., HILLEL, A., & MERATI, A. (2010). The Role of Experience on Judgments of Dysphonia. *Journal of Voice*, 24(5), 564–573. https://doi.org/10.1016/j.jvoice.2008.12.005

FARAHANI, M., PARSA, V., HERRMANN, B., KADEM, M., JOHNSRUDe, I., & DOYLE, P. C. (2020). An auditory-perceptual and pupillometric study of vocal strain and listening effort in adductor spasmodic dysphonia. *Applied Sciences*, 10(17), 5907. https://doi.org/10.3390/app10175907

GHIO, A., WEISZ, F., BARACCA, G., CANTARELLA, G., ROBERT, D., WOISARD, V., FUSSI, F., et al. (2011). Is the Perception of Voice Quality Language-Dependant ? A Comparison of French and Italian Listeners and Dysphonic Speakers. *Présenté à INTERSPEECH 2011*, Florence, Italy. https://hal.science/hal-01514687

GHIO, A., DUFOUR, S., ROUAZE, M., BOKANOWSKI, V., POUCHOULIN, G., RÉVIS, J., & GIOVANNI, A. (2011). Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle. *Revue de Laryngologie Otologie Rhinologie*, 132(1), 19-27. https://hal.science/hal-01491737v1

GHIO, A., CANTARELLA, G., WEISZ, F., ROBERT, D., WOISARD, V., FUSSI, F., GIOVANNI, A., et al. (2015). Is the perception of dysphonia severity language-dependent? A comparison of French and Italian voice assessments. *Logopedics Phoniatrics Vocology*, 40(1), 36-43. https://doi.org/10.3109/14015439.2013.837503

HELOU, L. B., SOLOMON, N. P., HENRY, L. R., COPPIT, G. L., HOWARD, R. S., & STOJADINOVIC, A. (2010). The Role of Listener Experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Ratings of Postthyroidectomy Voice. *American Journal of Speech-Language Pathology*, 19(3), 248–258. https://doi.org/10.1044/1058-0360(2010/09-0012)

HIDAKA, S., LEE, Y., NAKANISHI, M., WAKAMIYA, K., NAKAGAWA, T., & KABURAGI, T. (2022). Automatic GRBAS Scoring of Pathological Voices using Deep Learning and a Small Set of Labeled Voice Data. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2022.10.020

HIRANO, M. (1981). Clinical examination of voice. Wien; New York: Springer-Verlag. https://doi.org/10.1121/1.393788

KREIMAN, J., GERRATT, B. R., KEMPSTER, G. B., ERMAN, A., & BERKE, G. S. (1993). Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech, Language, and Hearing Research*, 36(1), 21–40. https://doi.org/10.1044/jshr.3601.21

KREIMAN, J., GERRATT, B. R., & BERKE, G. S. (1994). The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America*, 96(3), 1291-1302. https://doi.org/10.1121/1.410277

KREIMAN, J., & GERRATT, B. R. (1996). The perceptual structure of pathologic voice quality. *The Journal of the Acoustical Society of America*, 100(3), 1787-1795. https://doi.org/10.1121/1.416074

KREIMAN, J., & GERRATT, B. R. (2000). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*, 108(4), 1867-1876. https://doi.org/10.1121/1.1289362

KREIMAN, J., & SIDTIS, D. (2011). Foundations of voice studies: An interdisciplinary approach to voice production and perception. John Wiley & Sons.

KREIMAN, J., AUSZMANN, A., & GERRATT, B. R. (2020). What Does It Mean for a Voice to Sound "Normal"? Voice Attractiveness (p. 83-99). Springer.

LALLH, A. K., & ROCHET, A. P. (2000). The Effect of Information on Listeners' Attitudes Toward Speakers With Voice or Resonance Disorders. *Journal of Speech, Language, and Hearing Research*, 43(3), 782-795. https://doi.org/10.1044/jslhr.4303.782

MENG, X. L., ROSENTHAL, R., & RUBIN, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1), 172.

MISONO, S., MERATI, A. L., & EADIE, T. L. (2012). Developing Auditory-Perceptual Judgment Reliability in Otolaryngology Residents. *Journal of Voice*, 26(3), 358–364. https://doi.org/10.1016/j.jvoice.2011.07.006

PETTIROSSI, A., AUDIBERT, N., & CREVIER-BUCHMAN, L. (2020). Corrélats acoustiques et perceptifs de la personnalité perçue à travers la voix dans une population de dysphoniques légères. *In 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique*

pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (pp. 489-497). ATALA. https://hal.science/hal-02798576/

PETTIROSSI, A. (2021). La dysphonie chez les professeures des écoles: perception et représentations (Doctoral dissertation, Université Sorbonne Nouvelle). https://theses.hal.science/tel-03152574

R CORE TEAM. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Consulté à l'adresse https://www.R-project.org/.

ROUQUETTE, M.-L., & RATEAU, P. (1998). Introduction à l'étude des représentations sociales. Presses Universitaires de Grenoble.

SCHUERING, J. H. C., VAN HOF, K. S., HEIJNEN, B. J., VAN BENTHEM, P. P. G., SJÖGREN, E. V., & LANGEVELD, A. P. M. (2021). Proposal for a Core Outcome Set of Measurement Instruments to Assess Quality of Voice in Adductor Spasmodic Dysphonia Based on a Literature Review. *Journal of Voice*, 35(6), 933.e7-933.e21. https://doi.org/10.1016/j.jvoice.2020.02.010

SELLARS, C., STANTON, A. E., MCCONNACHIE, A., DUNNET, C. P., CHAPMAN, L. M., BUCKNALL, C. E., & MACKENZIE, K. (2009). Reliability of perceptions of voice quality: evidence from a problem asthma clinic population. *The Journal of Laryngology & Otology*, 123(7), 755-763. https://doi.org/10.1017/S0022215109004605

SUHAIL, I., KAZI, R., & JAGADE, M. (2016). Perceptual evaluation of tracheoesophageal speech: Is it a reliable tool? *Indian Journal of Cancer*, 53(1), 127-131. https://doi.org/10.4103/0019-509X.180814

YAMAGUCHI, H., SHRIVASTAV, R., ANDREWS, M. L., & NIIMI, S. (2003). A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatrica et Logopaedica*, 55(3), 147-157. https://doi.org/10.1159/000070726

ZUCKERMAN, M., & DRIVER, R. E. (1988). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, 13(2), 67-82. https://doi.org/10.1017/S0022215109004605

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Implémentation ouverte et étude de BEST-RQ pour le traitement de la parole

Ryan Whetten[1] Titouan Parcollet[2] Marco Dinarelli[3] Yannick Estève[1]
(1) Laboratoire Informatique d'Avignon, Avignon Université, France
(2) Samsung AI Center, Cambridge, United Kingdom
(3) Univervisté Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000, Grenoble, France

RÉSUMÉ

L'apprentissage auto-supervisé (SSL) a fait ses preuves pour le traitement automatique de la parole mais est généralement très consommateur de données, de mémoire et de ressources matérielles. L'approche BEST-RQ (BERT-based Speech pre-Training with Random-projection Quantizer) est une approche SSL performante en reconnaissance automatique de la parole (RAP), plus efficiente que wav2vec 2.0. L'article original de Google qui introduit BEST-RQ manque de détails, comme le nombre d'heures de GPU/TPU utilisées pour le pré-entraînement et il n'existe pas d'implémentation open-source facile à utiliser. De plus, BEST-RQ n'a pas été évalué sur d'autres tâches que la RAP et la traduction de la parole. Dans cet article, nous décrivons notre implémentation open-source de BEST-RQ et réalisons une première étude en le comparant à wav2vec 2.0 sur quatre tâches. Nous montrons que BERT-RQ peut atteindre des performances similaires à celles de wav2vec 2.0 tout en réduisant le temps d'apprentissage d'un facteur supérieur à deux. [1]

ABSTRACT

Open Implementation and Study of BEST-RQ for Speech Processing

Self-Supervised Learning (SSL) has proven to be useful in various speech tasks. However, these methods are generally very resource demanding. BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ), is an SSL method that has shown great performance on Automatic Speech Recognition (ASR) while being simpler than other SSL methods, such as wav2vec 2.0. Despite BEST-RQ's great performance, details are lacking in the original paper, such as the amount of GPU/TPU hours used in pre-training, and there is no official easy-to-use open-source implementation. Furthermore, BEST-RQ has not been evaluated on other downstream tasks aside from ASR and speech translation. In this work, we describe a re-implementation of BEST-RQ and perform a preliminary study with a comparison to wav2vec 2.0 on four downstream tasks. We discuss the details of our implementation. We show BEST-RQ can achieve similar downstream performance as wav2vec 2.0 while decreasing training time by over a factor of two.

MOTS-CLÉS : Apprentissage auto-supervisé, reconnaissance de la parole, reconnaissance du locuteur, repérage de mots-clés.

KEYWORDS: Self-supervised learning, speech recognition, speaker recognition, keyword spotting.

1. Cet article est une traduction de Whetten *et al.* (2024) publié au workshop ICASSP SASB 2024 sur l'apprentissage auto-supervisé.

1 Introduction

L'apprentissage auto-supervisé (SSL) est une technique d'apprentissage dans lequel les étiquettes à prédire sont extraites des étiquettes des données d'entrée elles-mêmes. Le SSL peut ainsi tirer parti de grandes quantités de données non étiquetées lors d'une phase de pré-apprentissage, et utiliser une quantité réduite de données déjà étiquetées pour obtenir des résultats très impressionnants sur une grande variété de tâches (Mohamed *et al.*, 2022). Dans le domaine du traitement automatique de la parole, le SSL a permis d'obtenir des résultats à l'état de l'art dans des tâches telles que la reconnaissance automatique de la parole (ASR), la reconnaissance automatique des émotions (ER), la vérification automatique du locuteur (ASV) et la compréhension du langage parlé (SLU) (Mohamed *et al.*, 2022; Yang *et al.*, 2021)

Cependant, le pré-entraînement SSL est très coûteux en termes de données, de mémoire et de calcul. Par exemple, les auteurs de wav2vec 2.0 ont déclaré avoir utilisé environ 2 400 heures de GPU V100 et une taille de lot (*batch*) de 1,6 heure uniquement pour le modèle de base (Baevski *et al.*, 2020). Pour des ensembles de données très volumineux comme pour *LeBenchmark*, les auteurs ont déclaré avoir utilisé 54 600 heures de GPU A100 pour leur modèle extra-large (Parcollet *et al.*, 2023a).

Malgré les efforts déployés pour améliorer l'efficacité d'autres modèles SSL largement utilisés pour la parole, tels que HuBERT et data2vec (Chen *et al.*, 2023; Baevski *et al.*, 2023), le processus reste gourmand en ressources.

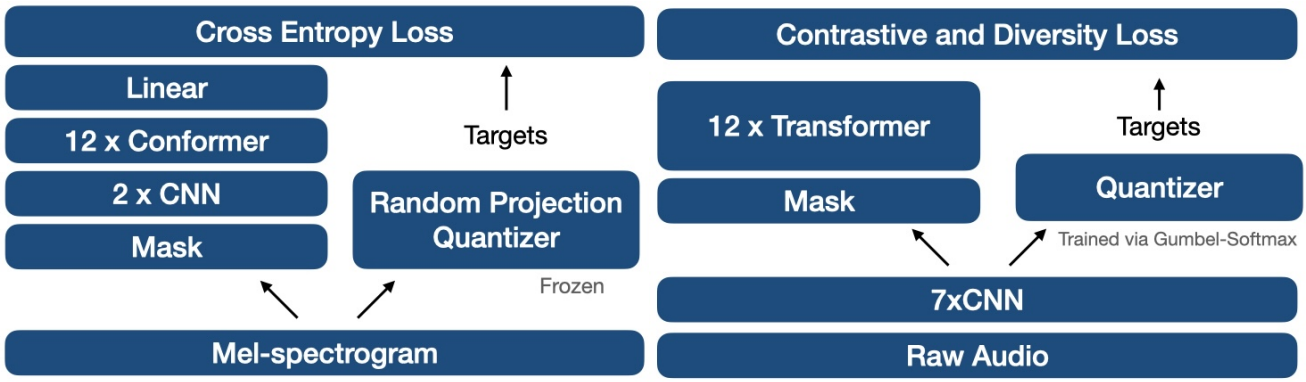
L'une des raisons de ce coût élevé est liée aux extracteurs de caractéristiques acoustiques qui sont généralement mis en œuvre sous la forme d'une série de couches de réseaux neuronaux convolutifs (CNN). Des études récentes ont montré qu'ils peuvent être remplacés par des solutions plus efficaces sans perte de performances (Parcollet *et al.*, 2023b).

Un modèle récent, BEST-RQ (*BERT-based Speech pre-Training with Random-projection Quantizer*) (Chiu *et al.*, 2022), réduit ce coût en réintroduisant l'emploi des banques de filtres Mel à la place de couches convolutives qui doivent être apprises. Grâce à cela et à d'autres simplifications (voir la section 2.1), BEST-RQ semble être l'une des méthodes SSL les plus efficaces proposées jusqu'à présent, tout en conservant des performances très compétitives pour la reconnaissance automatique de la parole.

Actuellement, il n'existe aucune implémentation officielle sous licence libre de BEST-RQ, ce qui limite l'accès à la communauté d'une méthode d'apprentissage SSL très efficace. De plus, les performances de BEST-RQ n'ont été étudiées que pour deux tâches, la reconnaissance automatique de la parole et la traduction vocale (Zhang *et al.*, 2023). Notre objectif est de combler ces lacunes.

Dans cet article, nous présentons notre implémentation *open-source* d'un discrétiseur à projection aléatoire utilisant SpeechBrain (Ravanelli *et al.*, 2021), et le résultat de nos premières expériences en comparant BestRQ à wav2vec 2.0. [2] Nous analysons le temps de calcul des pré-apprentissages ainsi que les performances sur les tâches suivantes : reconnaissance automatique de la parole, vérification automatique du locuteur, classification de l'intention et reconnaissance des émotions. Les résultats montrent qu'un discrétiseur à projection aléatoire peut atteindre des performances similaires à celles de wav2vec 2.0 sur ces différentes tâches, avec l'avantage supplémentaire de réduire de plus de la moitié le temps de pré-apprentissage auto-supervisé. Nous pensons que notre implémentation *open-source* pourra servir de point de départ à des recherches ultérieures, facilitant l'exploration de diverses

2. Le code de notre implémentation de BEST-RQ est disponible à https://github.com/speechbrain/speechbrain/pull/2309



(a) BEST-RQ (b) wav2vec 2.0

FIGURE 1 – Diagrammes de l'architecture BEST-RQ et de wav2vec 2.0. BEST-RQ opère sur des mel-spectrogrammes, utilise un discrétiseur statique et des couches de type *conformer*. De son côté, wav2vec 2.0 fonctionne sur de l'audio brut, apprend la discrétisation et utilise des couches de type *transformer*.

architectures pour l'apprentissage auto-supervisé (SSL) dans le cadre du traitement automatique de la parole.

2 Contexte

Bien qu'il existe d'autres modèles SSL efficaces pour le traitement de la parole, tels que HuBERT et data2vec (Chen *et al.*, 2023; Baevski *et al.*, 2023), wav2vec 2.0 semble plus répandu et utilisé par la communauté de la parole, avec plusieurs outils et modèles disponibles gratuitement, ce qui rend la comparaison plus aisée. Dans cette section, nous présentons une comparaison entre wav2vec 2.0 et BEST-RQ.

2.1 wav2vec 2.0 vs. BEST-RQ

BEST-RQ et wav2vec 2.0 présentent quelques différences majeures, l'une d'entre elles étant que **BEST-RQ n'opère pas directement sur la forme d'onde représentant le signal de parole**. Au lieu de cela, BEST-RQ traite plus classiquement des coefficients de banque de filtres log Mel à 80 dimensions avec deux couches CNN pour l'extracteur de représentations acoustiques. De son côté, wav2vec 2.0 opère sur la forme d'onde du signal audio et utilise sept couches CNN pour extraire des représentations acoustiques.

L'utilisation d'un extracteur de représentations acoustiques conçu à la main, comme les banques de filtres Mel, réduit considérablement la quantité de mémoire et de calculs nécessaires pour entraîner le modèle. Cependant, cela peut limiter les performances du modèle car les banques de filtres sont des vues compressées de la parole basées sur des connaissances psycho-acoustiques (c'est-à-dire basées sur l'expertise humaine) et, par conséquent, pourraient supprimer des indices acoustiques qu'un extracteur non construit à la main pourrait apprendre.

Une deuxième différence majeure est que **BEST-RQ utilise une couche linéaire initialisée de**

manière aléatoire et un livre de codes (*codebook*) pour la quantification et la discrétisation de l'audio. Ces deux composants sont gelés tout au long de l'apprentissage. Les banques de filtres Mel sont projetées via la couche linéaire, et l'indice de l'entrée du livre de codes la plus proche de la projection est utilisé comme cible. L'entrée du livre de codes la plus proche est trouvée en prenant l'*argmin* de la distance normalisée entre chaque entrée du livre de codes et la projection.

Ensuite, un masque est appliqué à une partie des banques de filtres Mel et l'objectif du modèle est de deviner les cibles correctes pour les sections masquées. Cette tâche est traitée comme une tâche de classification utilisant l'entropie croisée pour calculer la perte.

Inversement, dans wav2vec 2.0, il y a deux livres de codes qui sont appris en utilisant une fonction *softmax* de type Gumbel. La fonction de coût pour wav2vec 2.0 est plus complexe et consiste en une combinaison d'une fonction de type *constrative loss* (c'est-à-dire l'identification de l'étiquette correcte parmi un ensemble de fausses étiquettes sélectionnées au hasard) et d'une fonction de type *diversity loss* (pour empêcher le modèle de se réduire à l'utilisation d'une seule entrée dans le livre de codes).

Une dernière différence majeure est que **BEST-RQ utilise des couches de type *conformers* au lieu de couches de type *transformers***. Bien que l'utilisation de couches de type *conformers* dans une architecture wav2vec 2.0 a déjà été étudiée (Zhang *et al.*, 2020), nous ne considérons dans notre étude que l'architecture standard wav2vec 2.0, car notre objectif n'est pas de prouver qu'une méthodologie est meilleure qu'une autre, mais plutôt de montrer que notre implémentation fonctionne comme prévu et se compare favorablement à une autre solution bien connue. La figure 1 illustre schématiquement les principales différences entre ces deux architectures.

Dans (Chiu *et al.*, 2022), les auteurs ont montré qu'il est possible d'obtenir de bonnes performances en transcription automatique sans avoir à apprendre un extracteur de représentations acoustiques ou des livres de codes. En théorie, cela réduit considérablement la complexité du modèle, de la fonction de coût et de la passe arrière de mise à jour des poids. Le temps de pré-apprentissage n'ayant pas été indiqué dans l'article original de BEST-RQ, nous visons à montrer empiriquement comment cela affecte le pré-entraînement et proposons des évaluations sur transcription automatiques et trois autres tâches en aval. Pour des raisons de reproductibilité et pour permettre à d'autres chercheurs d'étudier les effets de l'utilisation d'un discrétiseur à projection aléatoire, notre implémentation est disponible sous licence libre.

3 Expériences

Pour nos expériences, nous pré-entraînons un modèle wav2vec 2.0 et notre modèle BEST-RQ, puis nous les utilisons pour un sous-ensemble des tâches du benchmarck MP3S (Zaiem *et al.*, 2023). Dans les sections suivantes, nous décrivons l'environnement de pré-apprentissage, l'architecture des modèles, et les tâches évaluées.

3.1 Paramètres de pré-apprentissage

Pour le pré-apprentissage, nous utilisons 960 heures de parole issues du corpus LibriSpeech (c'est-à-dire les sous-corpus *train-clean-100*, *train-clean-360* et *train-other-500*) (Panayotov *et al.*, 2015). L'apprentissage auto-supervisé de chacun des modèles est réalisé sur 42 époques ou environ 200k

itérations (*mises à jour des poids*) en utilisant huit cartes GPU Tesla V100 à 32Go de RAM. Nous sauvegardons un point de contrôle à l'époque 21 (environ 100k itérations), pour évaluer l'évolution de la performance au cours de l'apprentissage. Nous notons la différence entre ces versions du modèle en les nommant *100k* et *200k*.

Nous utilisons les mêmes paramètres de mise en lot (*batch*) dynamique pour les modèles wav2vec 2.0 et BEST-RQ, la taille maximale du lot étant de 100 secondes. Comme nous utilisons huit GPU, la taille totale des lots est de 800 secondes, soit environ 13,33 minutes. Nous avons choisi cette taille de lot pour pouvoir réaliser nos expériences sur de petits GPU et dans un temps limité en utilisant seulement quelques centaines d'heures de GPU.

3.2 Hyper-paramètres du modèle

Nous utilisons l'architecture du modèle wav2vec 2.0-base (Baevski *et al.*, 2020) comme référence, en utilisant l'implémentation disponible dans SpeechBrain. Nous avons seulement modifié les paramètres du lot dynamique pour avoir une taille de lot maximale de 100 secondes, comme indiqué précédemment.

Pour notre implémentation de BEST-RQ, nous suivons globalement la description des auteurs, tout en introduisant **quelques changements clés** découverts lors de nos expériences préliminaires, et qui se sont révélés très importants pour les performances du modèle en raison de la taille relativement petite de notre lot (13 minutes contre 18 heures dans l'article original des chercheurs de Google), la petite quantité de données de pré-entraînement (960 heures contre 12 millions dans (Zhang *et al.*, 2023)), et le nombre de GPUs à notre disposition. Nous **avont réduit le nombre de couches de type *conformers* de 24 à 12**. Ce choix a été fait pour correspondre au même nombre de couches que le modèle wav2vec 2.0-base, et a permis au modèle de s'adapter aux GPU utilisés. Nous avons **ajouté un *layer drop* avec une probabilité de 0,05**, nous avons **réduit le taux d'apprentissage initial à 0,0008** et **enfin nous avons augmenté le ratio de masquage à environ 60% de l'audio** (c'est-à-dire que 15% des trames du mel-spectrogramme sont sélectionnées aléatoirement pour être masquées avec les trois trames suivantes). Des résultats marquants de nos expériences préliminaires sont présentés dans la section 4.2.

3.3 Les tâches visées

Pour les tâches finales, nous suivons la méthodologie des benchmarks MP3S (Zaiem *et al.*, 2023) et SUPERB (Yang *et al.*, 2021), c'est-à-dire que le modèle SSL est gelé et que l'entrée du modèle neuronal spécifique à la tâche finale est une somme pondérée des sorties des couches cachées du modèle pré-entraîné.

Pour la tâche de **reconnaissance automatique de la parole**, nous utilisons respectivement le sous-corpus *train-clean-100 et dev-clean* pour l'apprentissage et la validation, puis nous évaluons sur les sous-ensembles *test-clean* et *test-other* de LibriSpeech (Panayotov *et al.*, 2015).

Pour cette tâche, le modèle neuronal spécifique à la tâche est composé de deux couches BiLSTM, suivie d'une couche linéaire, d'une couche softmax, puis d'une fonction de coût CTC (Zaiem *et al.*, 2023). Chaque couche BiLSTM a une taille de 1024 et se voit appliqué un dropout de 0,2 durant l'apprentissage. La métrique que nous utilisons pour évaluer les performances est le taux d'erreur sur

les mots (WER). Nous indiquons le WER avec et sans l'application du modèle de langage officiel [3] *4-gram*.

Pour la tâche de **vérification automatique du locuteur**, nous utilisons Voxceleb1 (Nagrani *et al.*, 2017). Ce jeu de données est composé d'énoncés de plus de 1 000 célébrités recueillis sur YouTube. Il s'agit d'une tâche de classification binaire où, étant donné deux fichiers audio, le modèle doit déterminer si le locuteur est le même ou non dans les deux enregistrements.

Nous utilisons l'architecture ECAPA-TDNN (Desplanques *et al.*, 2020) du benchmark MP3S. Pour mesurer les performances de l'ASV, nous utilisons l'*Equal Error Rate* (EER).

Pour **la classification des intentions**, nous utilisons SLURP (Bastianelli *et al.*, 2020), qui est connu pour être plus difficile que d'autres ensembles de données pour la même tâche, consistant en 177 locuteurs à partir de 72k fichiers audio totalisant 58 heures d'audio. La tâche consiste à classer un énoncé donné dans l'une des 18 catégories ou scénarios suivants : *email*, *calendar* ou *play* (comme dans *play next song*).

Pour l'architecture neuronale, nous utilisons à nouveau celle de MP3S. Cette architecture est composée une couche BiLSTM d'une taille de 1024, suivie d'une couche linéaire, d'une couche de *statistical pooling*, puis d'une couche linéaire pour la classification finale. La métrique utilisée pour cette tâche est la Précision.

Pour la tâche de **reconnaissance des émotions**, nous utilisons l'ensemble de données IEMOCAP (Busso *et al.*, 2008). Cet ensemble de données contient environ 12 heures de données provenant de 10 locuteurs jouant des scénarios avec quatre émotions différentes (neutre, heureux, triste et en colère). Les performances sont mesurées à l'aide d'une validation croisée sur 10 sous-ensembles (*10-fold cross validation*). Comme pour la tâche de vérification du locuteur, nous utilisons une architecture ECAPA-TDNN du benchmark MP3S.

Afin d'explorer les capacités de BEST-RQ au-delà de sa capacité à produire des représentations de parole, nous réalisons une expérience supplémentaire dans laquelle nous mettons à jour les poids du modèle pour la tâche de reconnaissance automatique de la parole sur les données LibriSpeech.

Le modèle est *fine-tuné* sur *train-clean-100* et nous l'évaluons ensuite sur *test-clean* et *test-other*, avec ou sans l'utilisation du modèle *4-gram*.

4 Résultats

Dans cette section, nous présentons des résultats sur le temps de pré-apprentissage, sur les tâches finales, ainsi que des résultats d'expériences préliminaires qu'il nous semble pertinent de partager.

4.1 Benchmark MP3S

Nos résultats sur les tâches présentées dans la section précédente sont rapportés dans le Tableau 1. Pour les tâches de vérification du locuteur, de classication d'intention et de reconnaissance des émotions, BEST-RQ est légèrement plus performant que wav2vec 2.0, alors que son pré-entraînement est environ 2,4 fois plus rapide que celui de wav2vec 2.0.

3. Disponible à l'adresse `openslr.org/11/`

Model / Task	# Par.	GPU hours	LibriSpeech train-100 ASR				VoxCeleb	SLURP	IEM.
Metric			WER ↓				EER ↓	Acc. ↑	Acc. ↑
			Clean	Clean LM	Other	Other LM	ASV	IC	ER
W2V2 100k	90.9M	130	16.26	10.63	40.17	30.83	4.56	72.8	60.9
W2V2 200k	90.9M	262	13.89	9.45	33.55	25.49	3.83	74.5	63.0
BRQ 100k	83.0M	54	16.79	10.79	38.09	28.31	3.84	74.3	61.3
BRQ 200k	83.0M	109	15.11	9.76	34.06	24.74	3.53	74.8	63.8

TABLE 1 – Résultats sur les tâches en aval à 100k et 200k pas. Les performances de BEST-RQ sont similaires à celles de wav2vec, mais avec moins de la moitié du temps d'apprentissage. Le temps d'apprentissage et le nombre de paramètres sont indiqués respectivement sous *GPU hours* et *# Par.*.

Model	Fine-Tune LibriSpeech train-100			
	Clean	Clean LM	Other	Other LM
W2V2 100k	16.42	10.29	36.62	27.25
W2V2 200k	13.47	8.73	29.64	21.92
BRQ 100k	14.59	9.00	31.49	23.04
BRQ 200k	12.21	7.78	26.81	19.67

TABLE 2 – Résultats de la mise au point sur l'ASR avec LibriSpeech *train-100*.

wav2vec 2.0 s'est avéré plus performant que BEST-RQ sans modèle de langage sur la tâche de reconnaissance automatique de la parole. Cependant, avec le modèle de langage, BEST-RQ obtient des résultats très proches de ceux de wav2vec 2.0 sur la tâche *test-clean* et même légèrement meilleurs sur la tâche *test-other*. Enfin, lorsque nous affinons BEST-RQ ou wav2vec 2.0 sur la tâche de reconnaissance automatique de la parole, BEST-RQ obtient de meilleurs résultats que wav2vec 2.0, comme l'illustrent les résultats présentés dans le tableau 2.

En examinant la différence entre les perfomances des modèles à 100k et 200k itérations dans ce tableau, nous remarquons wav2vec 2.0 s'améliore plus fortement durant l'apprentissage. Nous pensons que cela est dû au fait que la transformation du signal de parole en séquence d'*embeddings* dans wav2vec 2.0 doit être apprise – au contraire de BEST-RQ qui utilise des mel-spectrogrammes pré-calculés – ce qui implique une convergence plus lente.

4.2 Impact du taux de masquage

Avant de procéder aux expériences présentées dans la sous-section précédente, nous avons appris tous les modèles sur 18 époques, ou environ 87k itérations, sur 4 GPUs 2080Ti de 11Go, en nous focalisant uniquement sur les performances pour la reconnaissance de la parole sur le sous-corpus *dev-clean* de LibriSpeech. Nous avons fait varier le taux de masquage et la taille du livre de codes en utilisant des taux de masquage compris entre 1 % et 12 %.

Comme le montre le tableau 3, la réduction du livre de codes (CB) de 8192 à 1024 ne semble pas modifier les performances de manière cohérente ou significative. Pour un taux de masquage de 1 %, le livre de codes plus petit est légèrement plus performant, tandis que pour un taux de 10 %, il est légèrement moins performant. Au contraire, l'augmentation du nombre de trames masquées a un impact important, diminuant le WER sur les données *dev-clean* d'environ 15 % (d'un WER d'environ 35 % à un peu plus de 20 %) lorsque le taux de trames masquées passe de 1 à 12 %.

Mask %	Dev-Clean WER	CB
1%	34.08	1024
1%	36.45	8192
5%	25.24	8192
10%	21.10	1024
10%	20.68	8192
12%	20.11	8192

TABLE 3 – Impact de la modification du taux de masquage et de la taille du livre de codes. Le masque % concerne l'indice de la trame de départ choisi pour le masque. Étant donné que les trois trames suivantes sont également masquées, le taux de masquage réel est quatre fois plus important.

5 Discussion et conclusion

Dans ce travail, nous avons décrit notre implémentation *open-source* de BEST-RQ et nous l'avons comparé à wav2vec 2.0 en termes de temps de pré-entraînement et de performance sur diverses tâches en aval. BEST-RQ démontre des performances comparables à celles de wav2vec 2.0 tout en diminuant le temps de pré-apprentissage auto-supervisé de plus de la moitié. Le code de notre implémentation sera bientôt disponible dans la boîte à outils *SpeechBrain* (Ravanelli *et al.*, 2021).

Nous émettons l'hypothèse que BEST-RQ converge beaucoup plus rapidement que wav2vec 2.0 parce qu'il démarre avec des mel-spectrogrammes et qu'il a environ 8 millions de paramètres en moins, ce qui lui permet d'obtenir des performances comparables dans sa version à 200k itérations. Les différences de performance entre les modèles 100k et 200k suggèrent que le modèle wav2vec 2.0 pourrait surpasser BEST-RQ avec plus de temps ou d'autres paramètres d'entraînement (tels qu'une taille de lot plus importante).

Néanmoins, nous pensons que nos résultats sont révélateurs des capacités de ces méthodes avec seulement quelques centaines d'heures de GPU et nous continuerons dans cette voie pour nos travaux futurs.

Références

BAEVSKI A., BABU A., HSU W.-N. & AULI M. (2023). Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, p. 1416–1429 : PMLR.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BASTIANELLI E., VANZO A., SWIETOJANSKI P. & RIESER V. (2020). SLURP : A Spoken Language Understanding Resource Package. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7252–7262, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.588.

BUSSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S. & NARAYANAN S. S. (2008). IEMOCAP : Interactive emotional dyadic motion capture database. *Language resources and evaluation*, **42**, 335–359.

CHEN W., CHANG X., PENG Y., NI Z., MAITI S. & WATANABE S. (2023). Reducing Barriers to Self-Supervised Learning : HuBERT Pre-training with Academic Compute. *arXiv preprint arXiv :2306.06672*.

CHIU C.-C., QIN J., ZHANG Y., YU J. & WU Y. (2022). Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, p. 3915–3924 : PMLR.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). Ecapa-tdnn : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv :2005.07143*.

MOHAMED A., LEE H.-Y., BORGHOLT L., HAVTORN J. D., EDIN J., IGEL C., KIRCHHOFF K., LI S.-W., LIVESCU K., MAALØE L. *et al.* (2022). Self-supervised speech representation learning : A review. *IEEE Journal of Selected Topics in Signal Processing*.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : a large-scale speaker identification dataset. *arXiv preprint arXiv :1706.08612*.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, p. 5206–5210 : IEEE.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M. *et al.* (2023a). LeBenchmark 2.0 : a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech. *arXiv preprint arXiv :2309.05472*.

PARCOLLET T., ZHANG S., VAN DALEN R., RAMOS A. G. C. & BHATTACHARYA S. (2023b). On the (In) Efficiency of Acoustic Feature Extractors for Self-Supervised Speech Representation Learning. In *Interspeech 2023*.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J. *et al.* (2021). SpeechBrain : A general-purpose speech toolkit. *arXiv preprint arXiv :2106.04624*.

WHETTEN R., PARCOLLET T., DINARELLI M. & ESTÈVE Y. (2024). Open Implementation and Study of BEST-RQ for Speech Processing. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

YANG S.-W., CHI P.-H., CHUANG Y.-S., LAI C.-I. J., LAKHOTIA K., LIN Y. Y., LIU A. T., SHI J., CHANG X., LIN G.-T. *et al.* (2021). Superb : Speech processing universal performance benchmark. *arXiv preprint arXiv :2105.01051*.

ZAIEM S., KEMICHE Y., PARCOLLET T., ESSID S. & RAVANELLI M. (2023). Speech Self-Supervised Representation Benchmarking : Are We Doing it Right ? *arXiv preprint arXiv :2306.00452*.

ZHANG Y., HAN W., QIN J., WANG Y., BAPNA A., CHEN Z., CHEN N., LI B., AXELROD V., WANG G. *et al.* (2023). Google usm : Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv :2303.01037*.

ZHANG Y., QIN J., PARK D. S., HAN W., CHIU C.-C., PANG R., LE Q. V. & WU Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv :2010.10504*.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

L'impact du style de parole sur l'opposition de longueur des voyelles en arabe jordanien

Mohammad Abuoudeh[1] Jalal Al-Tamimi[2] Olivier Crouzet[3]
(1) Département de Langues et de Linguistique, Al-Hussein Bin Talal University, Ma'an, Jordanie
(2) Laboratoire de Linguistique Formelle, UMR 7110, Université Paris Cité, CNRS, 5 Rue Thomas Mann, 75013, Paris, France
(3) Laboratoire de Linguistique de Nantes (LLING), UMR6310 – Nantes Université / CNRS, France
mohammad.a.abuoudeh@ahu.edu.jo, jalal.al-tamimi@u-paris.fr, olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Cette étude examine l'impact des changements pouvant se produire dans deux styles de parole – lire vs. raconter une histoire – sur les informations spectrales et temporelles des voyelles longues et brèves en arabe jordanien. La lecture d'un texte est souvent considérée comme de la parole formelle ou soutenue, alors que la narration est plus spontanée. Le passage d'un style à l'autre peut engendrer des changements temporels et spectraux. C'est pourquoi un intérêt particulier a été porté au comportement des voyelles longues et brèves face à ces deux types de situations. Dix locuteurs de l'arabe jordanien ont lu puis raconté une histoire. Contrairement à ce qui était attendu, les caractéristiques spectrales et temporelles des voyelles n'ont pas été influencées par le changement de style. Cela suppose que dans cette expérience, le passage d'un style à l'autre a eu peu de conséquences sur la qualité et la quantité vocaliques. Cependant, les conditions comparées pourraient ne pas être suffisamment contrastées pour faire ressortir de telles différences. Les autres composantes du corpus en cours de constitution pourraient fournir des conditions plus à même de distinguer différents styles de parole.

ABSTRACT

Speaking style impact on vowel length opposition in Jordanian Arabic

This study examines the impact of changes in two speaking styles –story reading vs. storytelling– on the spectral and temporal properties of long and short vowels in Jordanian Arabic. Reading a text is usually associated with formal or clear speech, whereas storytelling is more spontaneous. The transition from one register to another may generate temporal and spectral modifications. This is why a particular interest has been paid to the behavior of long and short vowels in the context of these two speaking conditions. Ten speakers of Jordanian Arabic read then narrated the same short story. Contrary to what was expected, spectral and temporal vowel properties were not influenced by changes in speaking style. These results indicate that in this experiment, the transition from one register to the other had little impact on vowel quality and quantity. However, the conditions under scrutiny in this study may be too close to one another in order to enable such differences to emerge. Additional components of the currently collected corpus may be more appropriate to let differences between controlled and more spontaneous speech styles be revealed.

MOTS-CLÉS : style de parole, variations temporelles et spectrales, contraste de longueur vocalique, arabe jordanien.

KEYWORDS: speaking style, temporal and spectral variation, vowel length contrast, Jordanian

1 Introduction

Dans la parole continue, le style de parole change généralement de façon systématique selon la situation dans laquelle on se trouve. Par exemple dans une salle de classe, nous pouvons lire un texte (style « lecture »), parler à son professeur (style « formel ») et discuter avec son camarade de classe (style « informel »). D'un point de vue phonétique, le changement de style de parole pourrait provoquer des variations temporelles et spectrales des segments (Lindblom & Lindgren, 1985). Ces variations se produisent à cause du changement des stratégies dans la production de la parole. Il existe des situations dans lesquelles la parole doit être réalisée avec un grand degré de contraste perceptuel ; d'autres en exigent moins et permettent plus de variabilité. Par conséquent, les propriétés acoustiques du même son montrent de grandes variations reflétées à travers un continuum variant de l'hypo- à l'hyper-articulation (Lindblom, 1990; Farnetani & Recasens, 2010). En effet, les travaux de Lindblom et ses collègues (Lindblom & Lindgren, 1985; Lindblom, 1990; Lindblom *et al.*, 1992) – mais aussi ceux de Krull (1987) et Duez (1992) – mettent en avant que le changement de style de parole (parole formelle, informelle, lue, langage enfantin, etc.) peut faire subir de fortes transformations physiques aux mouvements articulatoires. Ces recherches soulignent également que le changement de style de parole vers la parole spontanée (« hypo-speech ») conduit à un degré de coarticulation maximal. De ce fait, les caractéristiques acoustiques des consonnes et des voyelles sont significativement modifiées. L'objectif de cette recherche est ainsi d'évaluer l'effet du changement de style de parole sur les informations temporelles et spectrales des voyelles phonologiquement longues et brèves en arabe.

Plusieurs chercheurs ont examiné l'influence des variations du style de parole sur la qualité et la quantité vocalique dans différentes langues (Leung *et al.*, 2016; DiCanio *et al.*, 2015; Blaauw, 1992; Bolotova, 2003; Meunier & Espesser, 2011). Les résultats de ces recherches ont révélé que dans la parole spontanée / informelle, la durée des segments ainsi que l'espace vocalique sont réduits en comparaison avec la parole lue / soutenue. Un effet d'*undershoot* est alors attendu lorsque l'on passe de la parole lue à la parole spontanée en arabe.

Néanmoins, peu d'études ont porté sur la relation entre les voyelles longues et brèves quand le style de parole varie. À titre d'exemple, DiCanio & Whalen (2015) ont montré qu'une influence asymétrique du style de parole sur les voyelles longues et brèves existe en arapaho [1]. La durée vocalique des voyelles longues est plus influencée par le changement du style de parole que celle des voyelles brèves. Toutefois, l'espace vocalique des voyelles longues est moins impacté que les voyelles brèves par ce facteur. Des résultats similaires ont été observés dans l'opposition tendue–relâchée en anglais (Leung *et al.*, 2016). La durée des voyelles tendues est plus impactée par la variation du style de parole que celle des voyelles relâchées. De plus, cette variation a moins de conséquences sur l'espace vocalique des voyelles relâchées que sur celui des voyelles tendues.

Des influences asymétriques ont également été remarquées entre les voyelles longues et brèves dans les études de variation de débit de parole dans plusieurs langues (en thaï Svastikula, 1986, en islandais Pind, 1995 et en japonais Hirata, 2004; Hirata & Tsukada, 2009). Selon ces recherches, la durée des voyelles longues est plus allongée que celle de leurs contreparties brèves lorsque le débit de parole ralentit. La durée des voyelles longues est aussi plus raccourcie que celle des voyelles brèves quand le débit de parole s'accélère. Toutefois, l'impact de variation du débit sur l'espace vocalique semble

1. Une langue algonquienne en danger d'extinction parlée dans l'État du Wyoming aux États-Unis.

dépendre de la langue. En thaï, les informations spectrales des voyelles longues et brèves restent relativement stables (Svastikula, 1986), contrairement à ce qui se passe en japonais, où les fréquences des voyelles brèves sont plus influencées par le changement de débit que leurs correspondantes brèves (Hirata & Tsukada, 2009). En résumé, les voyelles respectivement longues et brèves réagissent de manière différente lorsque le débit ou le style de parole est modifié.

2 Questions de recherche

Le but principal de cette recherche est d'évaluer à quel point le passage de la lecture d'une histoire à un style narratif pourrait influencer les informations temporelles et spectrales des voyelles longues et brèves en arabe jordanien. Selon les études mentionnées plus haut, on s'attend à ce que la lecture d'une histoire puisse déboucher sur des durées vocaliques plus longues et des espaces spectraux plus larges par rapport à la narration, puisque la tâche de lecture correspondrait à de la parole hyper-articulée alors que celle de narration se rapprocherait d'un style de parole plus hypo-articulé. De plus, cette influence pourrait être asymétrique entre les voyelles brèves et longues. Le choix de l'arabe jordanien a été motivé par le fait que cette langue est structurée autour d'un contraste entre voyelles phonologiquement longues et brèves, ce qui permettrait d'évaluer le comportement des phénomènes de durée face aux variations du style de parole. De plus, à notre connaissance, l'impact du style de parole n'a pas encore été examiné dans cette langue. Il est à noter que l'arabe jordanien contient trois voyelles brèves et leurs contreparties longues /i, iː, a, aː, u, uː/, ainsi que deux autres voyelles longues /eː, oː/. L'importance de la durée vocalique dans cette langue dépend du timbre ; autrement dit, /a, aː/ sont principalement différenciées par la durée, /u, uː/ sont quant à elles distinguées par la durée et l'information spectrale, tandis que /i, iː/ sont essentiellement opposées par l'information spectrale (Al-Tamimi, 2007; Abuoudeh, 2018). En outre, les voyelles longues occupent des positions spectrales plus périphériques que leurs contreparties brèves.

3 Méthodologie

3.1 Locuteurs

Pour répondre à la problématique de la présente étude, 10 locuteurs jordaniens (5 femmes et 5 hommes) ont participé bénévolement à une expérience de production. Les participants étaient tous étudiants en licence à l'Université Al-Hussein bin Talal, à Ma'an, dans le sud de la Jordanie, et étaient âgés entre 18 et 22 ans lors de l'enregistrement. Ils sont originaires d'Amman et de Zarqa, des villes situées dans la région Centre de la Jordanie. Les locuteurs ont déclaré qu'ils ne souffrent pas de troubles du langage.

3.2 Matériel

Le support linguistique de cette expérience est constitué de l'histoire du « Petit chaperon rouge », écrite en alphabet arabe dans une version de l'arabe jordanien rédigée par le premier auteur de l'article. Il est à remarquer que cette histoire étant populaire en Jordanie, l'ensemble des participants

enregistrés la connaissaient. Le choix d'une histoire connue et populaire a pour but de faciliter la tâche de narration [2].

3.3 Procédure

Tout d'abord, il a été demandé aux locuteurs de lire l'histoire du « Petit chaperon rouge » à partir d'un texte qui s'affichait sur un écran d'ordinateur. Par la suite, il leur a été demandé de raconter la même histoire, sans la lire. Avant l'enregistrement de la tâche de narration, les locuteurs pouvaient – s'ils le jugeaient nécessaire – relire à voix basse l'histoire afin de préparer leur narration. Avant le début de l'expérience, il a été indiqué aux participants de lire et de raconter l'histoire dans leur propre dialecte et non en arabe classique.

L'expérience s'est déroulée dans une salle calme des locaux de la Faculté des Lettres de l'Université Al-Hussein bin Talal. Le matériel utilisé pour les enregistrements est un micro Sennheiser e835 relié à un Tascam DR-100. Les fichiers sons ont été échantillonnés à 44100 Hz sur 32 bits en monophonique. Les enregistrements des deux tâches (lecture et narration) ont d'abord été transcrits et translittérés avec le nouveau système de translittération de l'arabe (convention ATR) puis segmentés par alignement forcé en utilisant le service `Arabic WebMAUS Basic` (Kisler *et al.*, 2017; Al-Tamimi *et al.*, 2022). Les résultats de l'alignement forcé ont été corrigés à la main dans un second temps par le biais du logiciel `Praat` (Boersma & Weenink, 2022). La durée des segments, la fréquence des formants (F1, F2, F3) et la f0 ont été automatiquement prélevées par un script `Praat`. L'algorithme d'extraction `Burg` (analyse LPC par auto-corrélation) a été employé avec une fenêtre d'analyse de 0.025 s et un pas de 0.01 s. Les seuils d'extractions des formants étaient adaptés au sexe du locuteur, autrement dit : 5000 Hz maximum pour les hommes et 5500 Hz maximum pour les femmes. Les données extraites par ce script ont ensuite été sauvegardées dans un fichier *.csv*. Pour cette étude, la durée et les fréquences des formants F1 et F2 des voyelles ont été analysées. Les fréquences des F1 et des F2 de tous les locuteurs ont été normalisées en utilisant la méthode de Lobanov afin de limiter la variation interlocuteurs (Lobanov, 1971) [3]. Les analyses de données ont été effectuées avec le logiciel `R` (R Core Team, 2023).

3.4 Analyses statistiques

Les relations entre chacune des variables dépendantes étudiées (Durée vocalique, F1 et F2) et les effets fixes (Voyelle et Tâche) ont été évaluées par des modèles linéaires mixtes avec la fonction `lmer` de la librairie `lme4` (Bates *et al.*, 2015). L'intercept pour les locuteurs a aussi été intégré aux modèles comme effet aléatoire. En outre, les pentes aléatoires par locuteur ont été incluses pour chaque effet fixe, correspondant à la variabilité inter-locuteur de l'effet de chaque facteur fixe sur les variables dépendantes de manière à éviter un taux élevé d'erreur de type I. Les *p-values* ont été obtenues par des approximations de type Satterthwaite par le biais de la fonction `anova` de la librairie `lmerTest` (Kuznetsova *et al.*, 2017). Ces analyses ont été suivies par des tests *post hoc* de Tukey en utilisant la fonction `glht` de la librairie `multcomp` (Hothorn *et al.*, 2008).

2. Les données de cette étude font partie d'une base de données plus large qui est actuellement en cours d'élaboration sur l'arabe jordanien (projet « Speech Database of Jordanian Arabic Dialects », abrégé en SDJAD), qui sera constituée de 100 participants provenant de différentes régions de Jordanie.

3. La normalisation a été effectuée par le biais de la fonction `normLobanov` de la librairie `phonR` (McCloy, 2016).

4 Résultats

Au total, les locuteurs ont produit 4972 voyelles en tâche de lecture et 3992 voyelles en tâche de narration comme détaillé dans la Table 1. Il était attendu que la tâche de narration contienne moins de réalisations que la tâche de lecture car les locuteurs pouvaient omettre ou résumer quelques évènements de l'histoire lors de leur narration. Par ailleurs, il est à remarquer que les voyelles brèves

	i	iː	a	aː	u	uː	eː	oː
Lecture	1120	393	1664	1185	81	188	278	63
Narration	942	360	1211	871	155	182	180	91

TABLE 1 – Nombre de réalisations pour chaque voyelle dans chaque style de parole.

– à l'exception de la voyelle /u/ – sont globalement plus fréquentes que les voyelles longues dans ces données, peu importe le style de parole, avec un total de 5173 voyelles brèves contre un total de 3791 voyelles longues.

4.1 La durée vocalique

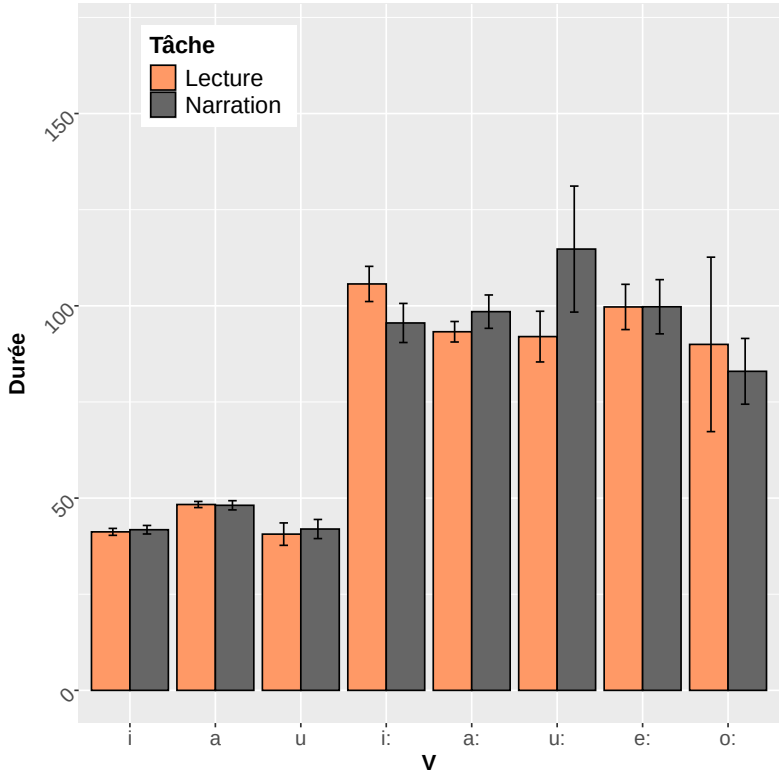


FIGURE 1 – Moyennes des durées vocaliques pour les deux conditions de style de parole (en ms, les barres d'erreur représentent l'Intervalle de Confiance à 95%) .

Les analyses descriptives indiquent que les deux styles de parole étudiés ont un impact faible sur les durées vocaliques (Figure 1). Les durées moyennes des voyelles brèves restent relativement stables dans les deux styles de parole. Quant aux voyelles longues, /iː, oː/ sont légèrement plus

longues en lecture qu'en narration. Les voyelles /aː, uː/, au contraire, sont plus allongées en narration qu'en lecture, notamment la durée de la voyelle /uː/. La durée de la voyelle /eː/ reste relativement inchangée dans les deux styles. Les observations des analyses descriptives sont confirmées par des analyses linéaires mixtes qui ne montrent aucune différence significative entre la tâche de lecture et celle de narration pour la durée ($F_{(1,7)} = 0.30, p = .587$). De plus, les analyses *post hoc* (Tukey) soulignent que la durée des voyelles n'est pas significativement différente en fonction du style de parole à l'exception des voyelles /iː, uː/. Ces résultats révèlent aussi que la relation temporelle entre les voyelles longues et brèves en arabe jordanien n'est pas influencée par le changement de style de parole en passant de la lecture à la narration.

Une série d'analyses inférentielles Bayésiennes (Package `brms` dans `R`; Bürkner, 2017) a été réalisée sur les données de durée vocalique à partir de 4 modèles de complexité décroissante. Les mesures ont été modélisées par une distribution *lognormale* afin de s'approcher au mieux des propriétés distributionnelles de durées vocaliques [4]. Les paramètres par défaut de la modélisation ont été utilisés (distributions de probabilités *a priori* non informatives, 1000 cycles de *warmup* suivis de 2000 itérations, sur 3 chaînes). Le modèle le plus complexe évalue l'effet des 3 prédicteurs que sont la *Catégorie de Voyelle*, le *Sexe* des participants et la *Tâche* réalisée et leurs interactions. Les 2 modèles suivants restreignent l'effet aux deux variables *Catégorie vocalique* et *Sexe* et leur interaction, puis *Catégorie vocalique* seule. Ces 3 modèles intègrent un intercept aléatoire par locuteur et lorsque c'est pertinent des pentes aléatoires par locuteur pour chaque variable intra-sujet (*Catégorie Vocalique* et *Tâche*). Le dernier modèle évalue l'effet de la *Catégorie vocalique* seule en n'intégrant qu'un intercept aléatoire et aucune pente aléatoire. Les temps de traitement de ces 4 modèles sur un serveur de calcul haute-performance ont été respectivement d'environ 3h10, 2h, 1h50 et 11 minutes.

Une comparaison des 4 modèles par méthode *leave-one out* fait ressortir que le modèle complet intégrant les 3 prédicteurs et leurs interactions (*Catégorie vocalique*, *Sexe* et *Tâche*) serait supérieur aux autres en termes de caractérisation des effets de durée mais que, à l'exception du modèle le plus simple (qui n'inclut que la *Catégorie Vocalique* et pas de pente aléatoire, il y a peu de différences –valeur absolue de différence d'ELPD < 4–, entre les 3 modèles les plus pertinents, qu'ils incluent les 3 prédicteurs ou que l'on retire le prédicteur *Tâche* aussi bien que le prédicteur *Sexe*. Ceci suggère que la contribution de la *Tâche* (et du *Sexe*) est relativement négligeable dans ces données. En complément, une exploration préliminaire des distributions *a posteriori* semble indiquer que, en dehors des variations liées à la *Catégorie Vocalique*, les effets sont très spécifiques et correspondent plutôt à des comportements particuliers d'interactions entre 2 ou 3 des prédicteurs plutôt qu'à des tendances globales liées à la *Tâche* ou au *Sexe* des participants.

L'exploration Bayésienne des effets et des distributions *a posteriori* devra être approfondie pour déterminer l'apport spécifique de cette approche aux données présentées. Ces analyses devront ultérieurement être élargies à l'étude des caractéristiques spectrales des voyelles.

4.2 L'espace vocalique

L'examen de l'espace vocalique met également en évidence que les deux styles de parole ont très peu d'influence sur les informations spectrales (Figure 2). En effet, les voyelles longues et brèves occupent des positions étroitement proches dans les deux styles de parole sur le plan F1-F2. Ces observations ont été confirmées par des analyses linéaires mixtes qui ne montrent aucune différence significative

4. En outre, cette modélisation fait gagner un temps de traitement considérable, souvent dans un rapport de 1 à 10, par rapport à la modélisation par une distribution Gaussienne.

entre la tâche de lecture et celle de narration pour les fréquences de F1 ($F_{(1,7)} = 0.48, p = .494$), et de F2 ($F_{(1,7)} = 0.0001, p = .99$). Les analyses *post hoc* (Tukey), quant à elles, confirment aussi que les fréquences de F1 et F2 observées pour toutes les voyelles ne changent pas significativement entre les deux styles de parole. En outre, ces résultats révèlent que la relation spectrale entre les voyelles longues et brèves en arabe jordanien n'est pas influencée par le changement de style de parole en passant de la lecture à la narration.

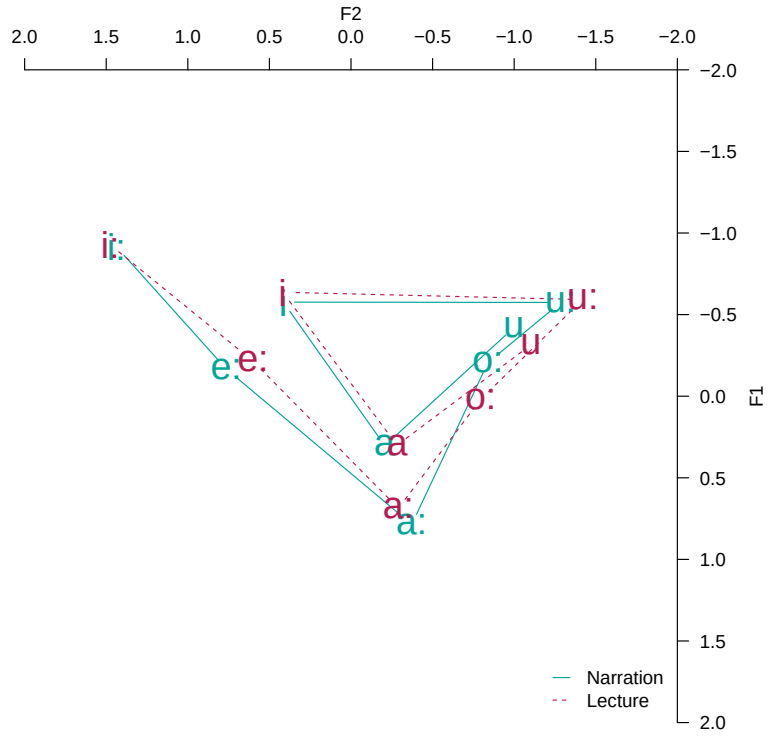


FIGURE 2 – Espace vocalique des voyelles longues et brèves normalisées avec la procédure de Lobanov en fonction du style de parole.

5 Discussion

Cette étude avait pour objectif d'évaluer l'effet du changement de style de parole sur l'opposition phonologique des voyelles longues et brèves en arabe jordanien. D'après les résultats présentés, ce changement influence très peu les informations spectrales et temporelles des voyelles longues et brèves. Dans les faits, la qualité vocalique n'a montré aucune différence significative entre les deux styles de parole pour toutes les voyelles. Quant à la quantité, seulement deux voyelles sur les huit ont révélé une différence significative en fonction du style (/uː/ et /iː/), dont une dans une direction inattendue. En effet, la voyelle /uː/ – et dans une moindre mesure la voyelle /aː/ avec un effet non significatif – présentent un allongement de leur durée en narration plutôt qu'en lecture. Cela pourrait être dû à des hésitations plus importantes dans la tâche de narration que dans celle de lecture.

Ces observations ne sont pas en accord avec les études antérieures sur d'autres langues. Pour rappel, ces études ont décrit que le passage de la parole formelle à la parole spontanée conduit à des variations spectrales et temporelles qui peuvent être asymétriques entre les voyelles longues et brèves. Les

résultats de cette recherche pourraient être expliqués par le fait que ces deux styles de parole ont potentiellement des effets limités au regard des différences temporelles dans une langue qui contient une opposition phonémique de longueur. Autrement dit, la distinction lecture *vs.* narration n'aboutit peut-être pas, dans le cas de cette étude, à produire une différence de style de parole à cause de la proximité des deux styles par comparaison aux styles étudiés dans les recherches mentionnées en introduction. Par exemple, DiCanio *et al.* (2015) – mais aussi DiCanio & Whalen (2015) – décrivent que leur condition d'élicitation est une prononciation répétée de mots isolés et que la parole « spontanée » est tirée de narration d'histoires personnelles. Ceci peut être nettement plus discriminant en termes de style de parole que la lecture *vs.* la narration du même conte telle que celle qu'on compare dans la présente étude.

De plus, l'importance de la séparation de durée entre les voyelles longues et brèves en arabe jordanien pourrait diminuer l'impact temporel et par conséquent, les variations associées à l'espace spectral dans ces deux de style de parole. Il est intéressant de signaler que les différences qualitatives entre les voyelles longues et brèves restent préservées. Par ailleurs, un autre facteur qui peut être avancé pour cette quasi-absence d'effet de style est que les locuteurs de l'arabe jordanien n'ont pas l'habitude de lire des histoires en arabe dialectal, puisque dans la majorité des cas, ils lisent des histoires en arabe classique. Ceci pourrait expliquer pourquoi les phénomènes phonétiques étudiés ici pourraient ne pas être clairement différenciés entre les tâches de lecture et de narration. Lors des enregistrements, une hésitation, voire un temps de réflexion, ont pu être observés chez certains locuteurs dans les deux styles de parole. L'étude des autres tâches du projet SDJAD (tel que les mots produits en isolation, la parole conversationnelle et la description d'une image), qui sont en cours de collecte, pourraient être utiles pour évaluer ces différentes observations.

Remerciements

Cette recherche fait partie du projet « Speech Database of Jordanian Arabic Dialects - SDJAD » financé par l'Université Al-Hussein Bin Talal avec la subvention numéro « 85/2022 ».

Références

ABUOUDEH M. (2018). *De l'impact des variations temporelles sur les transitions formantiques*. PhD thesis, Université de Nantes.

AL-TAMIMI J., SCHIEL F., KHATTAB G., SOKHEY N., AMAZOUZ D., DALLAK A. & MOUSSA H. (2022). A Romanization System and WebMAUS Aligner for Arabic Varieties. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), © European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*, p. 7269–7276.

AL-TAMIMI J.-E. (2007). *Indices dynamiques et perception des voyelles : Étude translinguistique en arabe dialectal et en français*. Thèse de doctorat, Université Louis Lumière - Lyon 2.

BATES D., MÄCHLER M., BOLKER B. & WALKER S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. DOI : https ://doi.org/10.18637/jss.v067.i01.

BLAAUW E. (1992). Phonetic differences between read and spontaneous speech. In *II International Conference on Spoken Language Processing ICSLP*.

BOERSMA P. & WEENINK D. (2022). Praat : doing phonetics by computer [computer program].

BOLOTOVA O. (2003). On some acoustic features of spontaneous speech and reading in russian (quantitative and qualitative comparison methods). In *15th International Congress of Phonetic Sciences (ICPhS-15)*.

BÜRKNER P.-C. (2017). brms : An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, **80**(1), 1–28. DOI : 10.18637/jss.v080.i01.

DICANIO C., NAM H., AMITH J. D., GARCÍA R. C. & WHALEN D. H. (2015). Vowel variability inelicited versus spontaneous speech : Evidence from mixtec. *Journal of Phonetics*, **48**, 45–59.

DICANIO C. & WHALEN D. (2015). The interaction of vowel length and speech style in an arapaho speech corpus. In *The 18th International Congress of the Phonetic Sciences*.

DUEZ D. (1992). Second formant locus-nucleus patterns : An investigation of spontaneous French speech. *Speech Communication*, **11**(4-5), 471–427.

FARNETANI E. & RECASENS D. (2010). Coarticulation and connected speech processes. In W. J. HARDCASTLE, J. LAVER & F. E. GIBBON, Éds., *The Handbook of Phonetic Sciences*, p. 316–352. Wiley-Blackwell, second édition.

HIRATA Y. (2004). Effects of speaking rate on the vowel length distinction in japanese. *Journal of Phonetics*, **32**, 565–589.

HIRATA Y. & TSUKADA K. (2009). Effects of speaking rate and vowel length on formant frequency displacement in japanese. *Phonetica*, **66**, 129–149.

HOTHORN T., BRETZ F. & WESTFALL P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, **50**(3), 346–363.

KISLER T., REICHEL U. & SCHIEL F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.*, **45**(C), 326–347. DOI : 10.1016/j.csl.2017.01.005.

KRULL D. (1987). Second formant locus patterns as a measure of consonant-vowel coarticulation. *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm-PERILUS*, **5**, 57–75.

KUZNETSOVA A., BROCKHOFF P. B. & CHRISTENSEN R. H. B. (2017). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **82**(13), 1–26. DOI : https ://doi.org/10.18637/jss.v082.i13.

LEUNG K. K. W., JONGMAN A., WANG Y. & SERENO J. A. (2016). Acoustic characteristics of clearly spoken English tense and lax vowelsa). *The Journal of the Acoustical Society of America*, **140**(1), 45–58. DOI : 10.1121/1.4954737.

LINDBLOM B. (1990). Explaining phonetic variation : A sketch of H&H theory. In W. HARDCASTLE & A. MARCHAL, Éds., *Speech production and speech modelling*, p. 403–439. Kluwer Academic Publishers.

LINDBLOM B., BROWNLEE S., DAVIS B. & MOON S.-J. (1992). Speech transforms. *Speech Communication*, **11**(4), 357–368.

LINDBLOM B. & LINDGREN R. (1985). Speaker-listener interaction and phonetic variation. *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm-PERILUS*, **4**, 77–85.

LOBANOV B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *The Journal of the Acoustical Society of America*, **49**(2B), 606–608. DOI : 10.1121/1.1912396.

MCCLOY D. R. (2016). Normalizing and plotting vowels with phonR 1.0.7. http://drammock.github.io/phonR/.

MEUNIER C. & ESPESSER R. (2011). Vowel reduction in conversational speech in french : The role of lexical factors. *Journal of Phonetics*, **39**(3), 271–278. DOI : https ://doi.org/10.1016/j.wocn.2010.11.008.

PIND J. (1995). Speaking rate, voice-onset time, and quantity : The search for higher-order invariants for two icelandic speech cues. *Perception & Psychophysics*, **57**(3), 291–304.

R CORE TEAM (2023). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

SVASTIKULA M. L. K. (1986). *A perceptual and acoustic study of the effects of speech rate on distinctive vowel length in Thai*. PhD thesis, The University of Connecticut.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

La reconnaissance automatique de phonèmes est-elle réellement adaptée pour l'analyse de la parole spontanée ?

Vincent P. Martin[1] Colleen Beaumard[2,3] Charles Brazier[2]
Jean-Luc Rouas[2] Yaru Wu[4]
(1) DDP Research Unit, Department of Precision Health, LIH, L-1445 Strassen, Luxembourg
(2) Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
(3) Univ. Bordeaux, CNRS, SANPSY, UMR 6033, F-33000 Bordeaux, France
(4) Université de Caen, CRISCO/UR4255, France
vincentp.martin@lih.lu, {colleen.beaumard, charles.brazier,jean-luc.rouas}@labri.fr, yaru.wu@unicaen.fr

RÉSUMÉ

La transcription phonémique automatique de la parole spontanée trouve des applications variées, notamment dans l'éducation et la surveillance de la santé. Ces transcriptions sont habituellement évaluées soit par la précision de l'identification des phonèmes, soit par leur segmentation temporelle. Jusqu'à présent, aucun système n'a été évalué simultanément sur ces deux tâches. Cet article présente l'évaluation d'un système de transcription phonémique du français spontané (corpus Rhapsodie) basé sur Kaldi. Ce système montre de bons résultats en identification des phonèmes et de leurs catégories, avec des taux d'erreur de 19,2% et 13,4% respectivement. Il est cependant moins performant en segmentation, manquant en moyenne 40% de la durée des phonèmes et 34% des catégories. Les performances s'améliorent avec le niveau de planification de la parole. Ces résultats soulignent le besoin de systèmes de transcription phonémique automatique fiables, nécessaires à des analyses plus approfondies de la parole spontanée.

ABSTRACT

Is automatic phoneme recognition suitable for spontaneous speech analysis ?

Automatic phonemic transcription of spontaneous speech has a wide range of applications, particularly in education and health monitoring. These transcriptions are usually evaluated either by the accuracy of phoneme identification or by their temporal segmentation. To date, no system has been evaluated simultaneously on both tasks. This article presents the evaluation of a Kaldi-based phonetic transcription system on spontaneous French in the Rhapsodie database. The system performed well in phoneme and category identification, with error rates of 19.2% and 13.4% respectively. However, its segmentation performances are low, missing on average 40% of phoneme duration and 34% of categories. Performance improves with the level of speech planning. These results underline the need for reliable automatic phonetic transcription systems, necessary for more in-depth analyses of spontaneous speech.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Reconnaissance de phonèmes, Segmentation de phonèmes.

KEYWORDS: Speech recognition, Phoneme recognition, Phoneme segmentation.

Actes des 35èmes Journées d'Études sur la Parole, pages 431–440.

1 Introduction

1.1 Contexte

Depuis les modèles fondateurs des années 1970 basés sur des modèles statistiques (Baker, 1975; Jelinek, 1976), le domaine de la transcription automatique de la parole a considérablement évolué, jusqu'au modèles d'apprentissage profond bout-en-bout (Alharbi *et al.*, 2021). Alors que les systèmes de transcription étaient initialement basés sur la modélisation des phonèmes, les modèles les plus récents et les plus performants fournissent directement une transcription des mots, à partir d'estimations des mots directement, de portions de mots, ou de caractères (Alharbi *et al.*, 2021). Cependant, un petit sous-ensemble de cas d'usages nécessitent encore une transcription phonémique, afin d'évaluer la précision de la prononciation lors de l'apprentissage d'une langue, de détecter les mots qui sont hors vocabulaire, (par exemple, pour la parole des enfants (Gelin *et al.*, 2021)) ou d'évaluer l'impact des pathologies sur l'articulation (Huckvale *et al.*, 2023; Beaumard *et al.*, 2023). Le domaine de la transcription automatique de la parole en phonèmes est partagé entre deux sous-applications différentes : d'un côté, l'estimation correcte de la séquence de phonèmes, mesurée par le Pourcentage d'Erreur de Phonèmes (PER) ; de l'autre, la segmentation correcte du fichier audio en phonèmes, délimitant leurs emplacements (généralement mesurée en termes de sensibilité, de spécificité et de score F1).

Concernant l'estimation automatique des phonèmes, le modèle le plus récent pour le français à notre connaissance repose sur le modèle Wav2Vec2 de Meta ré-entraîné plus finement sur Common Voice v13 et publié par Huggingface [1]. Ce système atteint des PER de 5,5% et 4,4% sur Common Voice v13 et Librispeech respectivement, qui sont tous deux des corpus de parole lue. Nous n'avons trouvé aucune évaluation récente concernant les performances de transcription phonémique pour la parole spontanée en français. En ce qui concerne la segmentation des signaux de parole en phonèmes, les dernières approches comprennent l'apprentissage auto-supervisé (Strgar & Harwath, 2023) et les modèles autorégressifs(Kim & Choi, 2023), atteignant des scores F1 autour de 90% sur les corpus TIMIT et Buckeye. Cependant, à notre connaissance, ces systèmes n'ont été évalués que sur l'une de deux tâches de détection ou de segmentation des phonèmes. Aucun système n'a donc été évalué conjointement sur les deux tâches.

1.2 Objectif

Notre objectif est d'évaluer un système standard de transcription de la parole en phonèmes pour différents styles de parole spontanée en français, tant en termes de reconnaissance des phonèmes (taux d'erreur phonémique) que de précision temporelle (rappel, précision et score F1). Cette double évaluation sera utile pour tout type d'analyse des phonèmes extraits automatiquement, que ce soit par exemple pour l'évaluation de la prononciation pour des apprenants de langues ou l'analyse de la parole pathologique.

Ce document est organisé comme suit. Nous introduisons le corpus Rhapsodie, notre modèle et les métriques de performance dans la Section 2. Nous rapportons et discutons les résultats du système conçu dans la Section 3 et concluons dans la Section 4.

1. https ://huggingface.co/Cnam-LMSSC/wav2vec2-french-phonemizer

2 Méthode

2.1 Système de reconnaissance automatique de la parole

Dans cette étude, nous utilisons un système de transcription automatique de la parole entraîné avec la boîte à outils Kaldi (Povey *et al.*, 2011). Il s'agit d'un modèle TDNN-HMM entraîné avec la fonction LF-MMI. Le réseau neuronal est un réseau à délai temporel échantillonné avec 7 couches TDNN, chacune ayant 1024 unités. La valeur de pas temporel est réglée sur 1 pour les trois premières couches, 0 pour la quatrième, et 3 pour les suivantes. Le modèle acoustique est basé sur un vecteur MFCC de haute résolution à 40 dimensions concaténé avec un i-vecteur de 100 dimensions (Gupta *et al.*, 2014). Les données d'apprentissage sont un sous-ensemble des corpus ESTER 1 et 2 (Galliano *et al.*, 2009) (discours radiophoniques). Ce système atteint un taux d'erreur de mots de 13.7% sur le sous-corpus de test d'ESTER (Boyer, 2021), ce qui est proche des performances systèmes état-de-l'art sur le même corpus (taux d'erreurs en mots légèrement inférieur à 12% (Heba, 2021)). Les phonèmes et leur alignement temporel sont obtenus avec la commande `lattice-align-phones`, qui permet d'obtenir une annotation et une segmentation en 35 phonèmes standards.

Source	Description	#enr.	#loc.	Durée	Style
CFPP2000	*Corpus de Français Parlé Parisien*, interviews à propos des quartiers de Paris (Branca-Rosoff & Lefeuvre, 2016)	3	2 H / 5 F	15min.	Semi-spt (3)
Avanzi	Collecté par M. Avanzi pour l'étude intonosyntaxique des phénomènes macrosyntaxiques (Avanzi, 2013)	18	7 H / 15 F	14 min	Spontané (17)
Lacheret	Collecté pour la modélisation continue et fonctionnelle du français (Lacheret-Dujour, 2003)	2	3 H / 1 F	9 min.	Planifié (1), Spontané (1)
Mertens	Collecté pour la modélisation intonosyntaxique du français (Mertens, 1987)	2	4 H / 0 F	10 min	Planifié (1), Semi-spt (1)
C-Prom	Collecté pour étudier les proéminences syllabiques en français (Avanzi & Simon, 2010)	1	1 H / 0 F	3 min.	Planifié (1)
ESLO	*L'Enquête Sociolinguistique à Orléans*, recueillie à Orléans, France en 1968-74 avec un objectif sociolinguistique (Eshkol-taravella *et al.*, 2011)	1	2 H / 0 F	7 min.	Planifié (1)
PFC	*Phonologie du français contemporain*, conversations dirigées entre un sujet et un intervieweur et conversations informelles entre deux personnes appartenant à un réseau social dense, (Durand *et al.*, 2009)	3	2 H / 4 F	14 min.	Spontané (3)
Film	Monologues dans lesquels 7 intervenants différents sont invités, dans un cadre informel, à décrire une courte scène d'un film de Charlie Chaplin collectée pour le projet Rhapsodie	7	4 H / 3 F	9 min.	Spontané (7)
Professionnel	Monologues et dialogues dans un contexte professionnel collectés pour le projet Rhapsodie	3	2 H / 2 F	8 min.	Spontané (3)
Télédiffusion	14 monologues diffusés, dialogues et conversations téléchargés d'Internet pour le projet Rhapsodie	14	22 H / 6 F	67 min.	Planifié (7), Spontané (6)
Tous		54	49 H / 36 F	2h 41m	

TABLE 1 – Description du corpus Rhapsodie : nombre d'échantillons, nombre de locuteurs, durée du corpus. *Semi-spt* : Semi-spontané. Les trois styles de parole spontanée sont ceux fournis dans les métadonnées du corpus.

2.2 Corpus Rhapsodie

Nos analyses ont été réalisées sur le corpus Rhapsodie, un corpus multigenre de français parlé (Lacheret-Dujour *et al.*, 2019). Le corpus contient au total trois heures de parole (∼33000 mots), composées de 54échantillons courts (5 minutes en moyenne). Il inclut des interviews en face à face, des émissions de radio et de télévision, pour un total de 89 locuteurs. Les transcriptions phonétiques sont obtenues en utilisant un outil de conversion automatique graphème-vers-phonème (g2p) (Easyalign (Goldman, 2011) dans Praat (Boersma, 2001)), suivi d'une vérification manuelle (Lacheret *et al.*, 2014). Les pauses ont été détectées automatiquement. Deux enregistrements, D0001 et D1003 (respectivement dans les sous-corpus CFPP2000 et Rhapsodie Professionnel) ont été exclus en raison de leur mauvaise qualité acoustique. Un autre fichier, M2006 (sous-corpus Télédiffusion), a été exclu en raison d'erreurs dans les frontières temporelles de l'annotation phonétique de la vérité terrain. Tous les résultats et statistiques ultérieurs n'incluent pas ces fichiers. Les différentes sources de données utilisées dans le corpus Rhapsodie sont décrites dans le Tableau 1.

Le corpus Rhapsodie contient plusieurs variables pour représenter les caractéristiques discursives de chaque échantillon. Nous nous concentrons dans cette étude sur l'analyse des résultats du système de transcription phonémique automatique en fonction du degré de planification de la parole, tel qu'annoté dans les métadonnées du corpus : parole planifiée, semi-spontanée ou spontanée.

2.3 Vérité terrain : phonèmes et catégories

Les 54 fichiers du corpus représentent un total de 96756 phonèmes, qui ont une durée moyenne de 81.2ms. Pour faciliter l'interprétation des résultats, les 35 phonèmes ont été regroupés en 10 catégories standards : 5 catégories pour les consonnes (occlusives, fricatives, nasales, liquides, glissantes) et 5 catégories pour les voyelles (antérieures arrondies, antérieures non arrondies, centrales, postérieures arrondies, nasales).

2.4 Métriques de performance

Le Pourcentage d'Erreur Phonétique (PER) est la métrique utilisée dans le domaine de la reconnaissance automatique de la parole pour mesurer la précision de la transcription phonémique. Le PER est calculé par la somme du nombre de substitutions (S), d'insertions (I) et de suppressions (D) de phonèmes dans l'hypothèse fournie par le système par rapport au nombre total de phonèmes dans la transcription de référence (N) : $PER = 100 \times (S + I + D)/N$

Une faible valeur de PER indique une bonne précision dans la reconnaissance automatique des phonèmes. Le PER ne prend cependant pas en compte les erreurs dues aux mauvais placement des frontières puisqu'il ne prend en compte que la séquence de symboles phonétiques.

Pour compléter le PER, nous souhaitons également mesurer les performances du système en termes de durée. À cette fin, nous utilisons l'outil *trackeval* (marge d'erreur = 0) qui a été utilisé lors de la campagne d'évaluation ESTER pour estimer les performances de la détection d'événements audio (Galliano *et al.*, 2009). Ici, les événements à identifier correctement sont les phonèmes. Ce faisant, nous avons mesuré trois indicateurs :

— *le score de Rappel*, qui est le rapport de la durée de détection correcte d'un phonème sur la durée totale de cet événement dans le fichier de référence : $R = \hat{d}_{corr}(phon)/d_{ref}(phon)$

— *le score de Précision*, qui est le rapport entre la durée de détection correcte d'un phonème sur la durée totale de détection de ce phonème (y compris les insertions) : $P =$

$\hat{d}_{corr}(phon)/\hat{d}_{corr+ins}(phon)$

— et *le F-score*, une métrique combinant la Précision et le Rappel en une seule métrique : $F = 2 \times (P \times R)/(P + R)$

Une score de *Rappel* élevé indique que le phonème considéré est bien détecté, tandis qu'un score de *Précision* élevé montre que le système détecte le phonème principalement lorsqu'il est réellement présent (peu d'insertions). Idéalement, un bon système de segmentation des phonèmes doit donc avoir à la fois des scores de *Précision* et de *Rappel* élevé, et donc une *Mesure-F* élevée.

Les métriques de PER et de segmentation sont également recalculées sur les catégories phonétiques.

3 Résultats et discussion

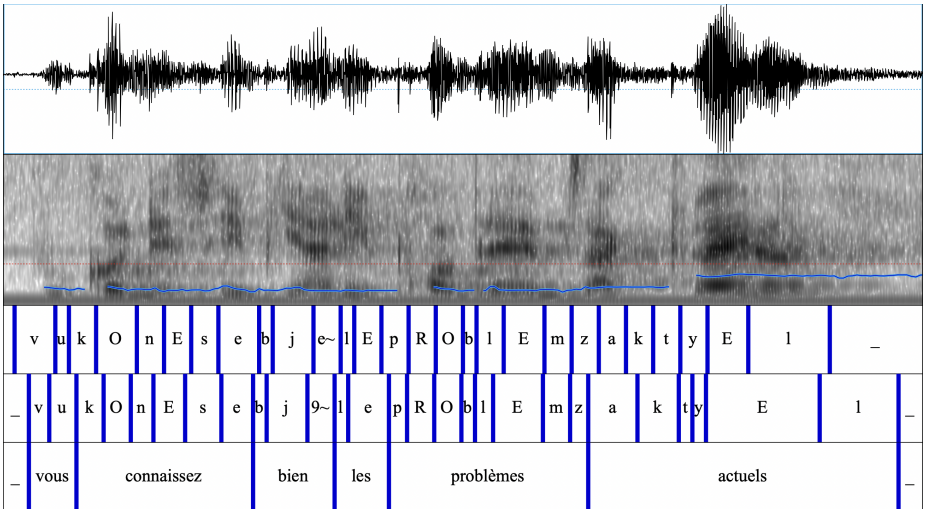


FIGURE 1 – Exemple de sortie du système de reconnaissance phonétique. En haut : résultats de la détection automatique ; au milieu : annotation phonétique de référence ; en bas : transcription en mots.

La Figure 1 montre un exemple de sortie de notre système automatique de reconnaissance phonétique. Cet exemple est un extrait du fichier D1001 du corpus Rhapsodie, provenant de la base de données ESLO (Eshkol-taravella *et al.*, 2011). Cet enregistrement date de 1968 et est un monologue d'un locuteur masculin.

Dans cet exemple, la plupart des phonèmes sont correctement détectés à l'exception de deux substitutions, dans deux couples de phonèmes proches l'un de l'autre : /9~/ a été substitué par /e~/ ; et /e/ par /E/. Étant donné les bonnes performances d'identification des phonèmes, si le système de reconnaissance phonétique se comporte de la même manière pour tous les fichiers, nous espérons obtenir de faibles valeurs de PER, démontrant l'efficacité du système automatique dans l'identification des phonèmes. Ceci est discuté dans la section 3.1.

Cependant, alors que la séquence de phonèmes est correctement identifiée, nous observons des inexactitudes sur les emplacements des frontières des phonèmes, particulièrement à la fin de l'extrait (pour le mot « actuels »). Ces inexactitudes peuvent poser problème dans l'utilisation de cette segmentation automatique pour d'autres analyses, telles que l'analyse prosodique ou l'analyse de la qualité de la voix sur des phonèmes spécifiques. La qualité de la segmentation est discutée dans la section 3.3.

3.1 Performances de reconnaissance des phonèmes

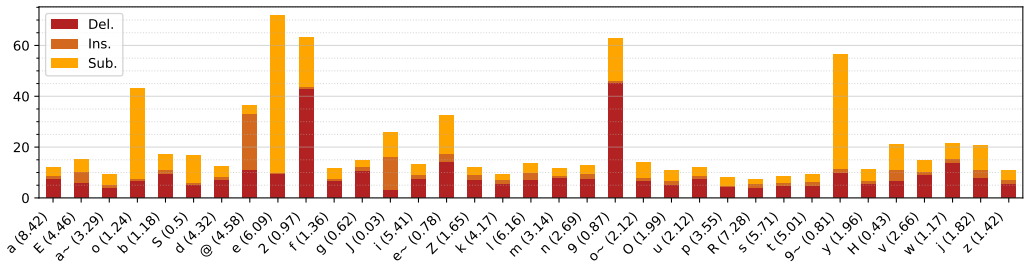


FIGURE 2 – Performances du système automatique de reconnaissance phonétique pour chaque phonème. Les valeurs entre parenthèses indiquent le ratio du nombre d'occurrences du phonème par rapport au nombre total de phonèmes

Les performances de notre système sur chaque phonème sont rapportées dans la Figure 2. Les erreurs les plus communes sont les substitutions, sauf pour /2/ et /9/ qui sont principalement supprimés. En particulier, la voyelle /e/ a un taux de substitution élevé. En effet, dans la parole continue, /e/ est interchangeable avec /E/ par les locuteurs natifs du français. Puisque le système ne permet pas de choix libre entre les deux phonèmes dans le dictionnaire, nous supposons qu'il tend à étiqueter /e/ lorsqu'il rencontre un son similaire à /e, E/. D'autre part, le taux d'insertion élevé pour le schwa /@/ est fortement lié au fait que le système n'est pas informé du caractère facultatif de la voyelle en français.

Lorsque l'on considère les catégories phonétiques (Figure 3), nous observons une réduction drastique du nombre de substitutions, montrant que la plupart de ces erreurs sont faites sur des phonèmes appartenant à la même catégorie. De plus, une observation intéressants peut être faite sur les taux d'erreur observés pour les consonnes : plus la consonne est sonore, plus il est difficile pour le système de l'identifier.

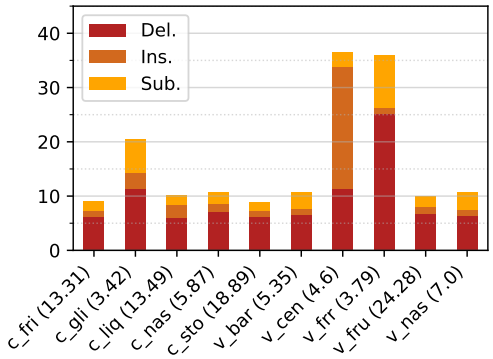


FIGURE 3 – Performances du système sur les catégories de phonèmes. Les valeurs entre parenthèses indiquent le ratio du nombre d'occurrences du phonème appartenant à la catégorie par rapport au nombre total de phonèmes

3.2 Performances de reconnaissance des phonèmes en fonction du style

Les performances de reconnaissance des phonèmes en termes de phonèmes et de catégories de phonèmes en fonction de chaque sous-corpus sont rapportées dans les Tableaux 2 et 3. Pour les deux

unités de mesure, le taux d'erreur diminue avec le degré de préparation de la parole spontanée.

Cependant, puisque la plupart des fichiers du sous-corpus planifié proviennent de la source "Télédiffusion" de la base de données Rhapsodie, nous nous attendions à ce que le système de transcription automatique fonctionne mieux sur ces données que sur les autres sous-corpus, car il est entraîné sur des échantillons du même type (bien que d'une période différente). Ce n'est pas le cas, en partie à cause d'un échantillon de nature plus spontanée que la parole planifiée (D2002, discussion sur un livre) résultant en de mauvaises performances (PER=24,9%). Néanmoins, le système obtient d'assez bonnes performances sur les autres sous-corpus semi-spontanés. Alors que le taux d'erreur augmente avec le degré de spontanéité, les taux d'insertion restent relativement constants, tandis que les taux de substitution augmentent légèrement. Le type d'erreur qui augmente le plus est le taux de délétion (de 3% sur la parole planifiée à 8,5% sur la parole spontanée), atteignant un maximum de 24,9% sur le fichier D2004 de la source Lacheret (locuteur avec un fort accent régional).

Concernant le fait que même pour la parole spontanée, plus de 80% des phonèmes sont détectés correctement (87% pour les catégories phonétiques) et que les erreurs sont principalement dues à des suppressions, nous pourrions penser que la détection automatique des phonèmes peut être adaptée pour l'analyse phonétique de la parole spontanée.

style	# fichiers	# phonèmes	Corr	Sub	Sup	Ins	Err
Tout	54	96756	83,9	9,4	6,7	3,4	19,5
planifié	11	30549	89,2	7,6	3,2	3,0	13,8
semi-spt	4	15302	82,8	9,3	7,9	3,3	20,5
spontané	39	50905	81,0	10,5	8,5	3,6	22,6

TABLE 2 – Erreurs pour la détection des tokens phonétiques pour différents styles de parole

style	# fichiers	# phonèmes	Corr	Sub	Sup	Ins	Err
Tout	54	96756	89,7	3,5	6,7	3,4	13,6
planifié	11	30549	94,9	1,9	3,2	3,0	8,1
semi-spt	4	15302	88,7	3,5	7,8	3.3	14,6
spontané	39	50905	87,0	4,5	8,6	3,6	16,7

TABLE 3 – Erreurs pour la détection des catégories phonétiques pour différents styles de parole

3.3 Performances de segmentation de phonèmes

Le Tableau 4 rapporte les performances de la segmentation des phonèmes selon les métriques détaillées dans la section 2.4.

Sur un total de 7912 secondes à détecter, seules 4746 s. (R=60,0%) sont correctement détectées au niveau du phonème, avec une précision de P=68,2%, conduisant à un score F de 0.62. Cette valeur atteint 5232 s. lorsque l'on considère les catégories phonétiques (66,1%), avec une précision supérieure P = 71,0%, conduisant à un score F correspondant de 0.68.

Concernant l'effet du degré de planification, de manière similaire à ce que nous avons observé dans la section 3.1, ajouter plus de spontanéité dégrade les résultats, à la fois pour les phonèmes (d'un score F1 de 0,67 pour le discours planifié à 0,58 pour le discours spontané) et pour les catégories phonétiques (de F=0,73 à F=0,68).

De plus, notre système fonctionne de manière inégale selon les classes phonétiques : alors qu'il segmente avec de bonnes performances les voyelles nasales (durée cible=823 s., R=71%, P=81%, F=0,76) et les consonnes fricatives (durée cible=1117 s., R=71%, P=78%, F=0,74), il a du mal à estimer les frontières des voyelles centrales (durée cible=330 s., R=60%, P=47%, F=0,53), des voyelles antérieures arrondies (durée cible=595 s., R=30%, P=72%, F=0,42) et des consonnes glissantes (durée cible=170 s., R=57%, P=48%, F=0,52). Les autres catégories phonétiques sont segmentées avec des scores F entre 0,60 et 0,68.

		Phonèmes			Catégories de phonèmes		
	cible	%R	%P	F	%R	%P	F
planifié	2596s	66,5	68,2	0,67	72,3	74,1	0,73
semi-spt	1198s	59,0	64,6	0,62	65,2	71,3	0,68
spontané	60,8	55,5	61,5	0,58	62,5	68,8	0,65
Tous	7912s	60,0	64,4	0,62	66,1	71,0	0,68

TABLE 4 – Évaluation de la segmentation pour la détection des tokens phonétiques pour différents styles de parole

4 Conclusion

Cet article évalue les performances d'un système de reconnaissance automatique des phonèmes pour le français spontané, non seulement en termes de détection de phonèmes mais aussi sur l'identification correcte de leurs frontières. Basé sur le corpus Rhapsodie, qui contient de la parole spontanée de plusieurs sources avec trois degrés de spontanéité, nous avons calculé des métriques d'identification et de segmentation tant au niveau phonémique qu'en fonction de dix catégories phonétiques standards (5 types de voyelles et 5 types de consonnes). Nous avons montré qu'un système de transcription automatique pouvait en même temps obtenir des performances satisfaisantes en identification des phonèmes (PER global de 19,5%, taux d'erreur de 13,6% sur les catégories) et des performances de segmentation insatisfaisantes (F-score de 0,62 et 0,68 pour les phonèmes et les catégories respectivement). Cependant, dans les deux évaluations, de nombreuses disparités ont été observées en fonction du type de parole et des phonèmes considérés.

La prise en compte des catégories de phonèmes améliore les performances dans les deux évaluations, suggérant que les substitutions sont effectuées dans le même groupe phonétique. De plus, toutes les métriques de performance augmentent avec le degré de planification de la parole spontanée considérée. Comme la plupart des systèmes de reconnaissance phonétique sont uniquement évalués sur la parole lue, ces résultats incitent à être très prudents lors de l'utilisation de tels systèmes pour l'analyse linguistique ou prosodique de la parole spontanée.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche (ANR) dans le cadre de l'axe Autonom-Health du PEPR Santé Numérique, convention de subvention n°ANR-22-PESN-0009. VPM a reçu le soutien financier du programme de recherche et d'innovation européen Horizon Europe à travers le projet Marie Skłodowska-Curie MATER (No. 101106577). CB a reçu le soutien financier de la MITI du CNRS (projet PRIME 80 DSM-HEALTH).

Références

ALHARBI S., ALRAZGAN M., ALRASHED A., ALNOMASI T., ALMOJEL R., ALHARBI R., ALHARBI S., ALTURKI S., ALSHEHRI F. & ALMOJIL M. (2021). Automatic speech recognition : Systematic literature review. *IEEE Access*, **9**, 131858–131876. DOI : 10.1109/ACCESS.2021.3112535.

AVANZI M. (2013). *L'interface prosodie/syntaxe en français*.

AVANZI M. & SIMON A. C. (2010). C-PROM : An Annotated Corpus for French Prominence Study. *Speech Prosody*.

BAKER J. K. (1975). *Stochastic modeling as a means of automatic speech recognition.* Carnegie Mellon University.

BEAUMARD C., MARTIN V. P., WU Y., ROUAS J.-L. & PHILIP P. (2023). Automatic detection of schwa in French hypersomniac patients. In *Journée Santé et Intelligence Artificielle (Evènement affilié à PFIA 2023)*.

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

BOYER F. (2021). *Reconnaissance de Parole Pour Le Français et Intégration Dans Un Système de Compréhension Du Langage Parlé*. Thèse de doctorat, Université de Bordeaux.

BRANCA-ROSOFF S. & LEFEUVRE F. (2016). Le CFPP2000 : constitution, outils et analyses. Le cas des interrogatives indirectes. *Corpus*, (15). DOI : 10.4000/corpus.3043.

DURAND J., LAKS B. & LYCHE C. (2009). Phonologie, variation et accents du français. chapitre Le projet PFC : une source de données primaires structurées, p. 19–61. Hermès.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral disponible : le Corpus d'Orléans 1968-2012 [A Large available oral corpus : Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, **52**(3), 17–46.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech 2009*, p. 2583–2586. DOI : 10.21437/Interspeech.2009-680.

GELIN L., PELLEGRINI T., PINQUIER J. & DANIEL M. (2021). Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition. In *Proc. Interspeech 2021*, p. 3860–3864. DOI : 10.21437/Interspeech.2021-2202.

GOLDMAN J.-P. (2011). Easyalign : an automatic phonetic alignment tool under praat. In *Twelfth Annual Conference of the International Speech Communication Association*.

GUPTA V., KENNY P., OUELLET P. & STAFYLAKIS T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP*, p. 6334–6338. DOI : 10.1109/ICASSP.2014.6854823.

HEBA A. (2021). *Reconnaissance Automatique de La Parole à Large Vocabulaire : Des Approches Hybrides Aux Approches End-to-End*. Theses, Université toulouse 3 Paul Sabatier.

HUCKVALE M., LIU Z. & BUCIULEAC C. (2023). Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech. *Biomedical Signal Processing and Control*, **86**, 105201. DOI : https ://doi.org/10.1016/j.bspc.2023.105201.

JELINEK F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, **64**(4), 532–556. DOI : 10.1109/PROC.1976.10159.

KIM H. & CHOI H.-S. (2023). Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5. DOI : 10.1109/ICASSP49357.2023.10096748.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : a prosodic-syntactic treebank for spoken French. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 295–301, Reykjavik, Iceland : European Language Resources Association (ELRA).

LACHERET-DUJOUR A. (2003). *La prosodie des circonstants en français parlé*. Collection linguistique (Paris). Paris : Peeters.

LACHERET-DUJOUR A., KAHANE S. & PIETRANDREA P. (2019). *Rhapsodie : A Prosodic and Syntactic Treebank for Spoken French*. John Benjamins.

MERTENS P. (1987). *L'intonation Du Français : De La Description Linguistique à La Reconnaissance Automatique*. Thèse de doctorat.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society.

STRGAR L. & HARWATH D. (2023). Phoneme segmentation using self-supervised speech models. In *IEEE Spoken Language Technology Workshop (SLT)*, p. 1067–1073. DOI : 10.1109/SLT54892.2023.10022827.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

La sonorité n'est pas l'intensité : le cas des diphtongues dans une langue tonale

Yunzhuo XIANG Jiayin GAO Cédric GENDROT
Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)
yunzhuo.xiang@sorbonne-nouvelle.fr

RÉSUMÉ

Cette étude explore le lien entre la sonorité et l'intensité dans la production des diphtongues ouvrantes et fermantes en mandarin de Pékin. Étant donné qu'une voyelle ouverte est considérée comme plus sonore qu'une voyelle fermée, nous nous attendons à constater une augmentation d'intensité dans une diphtongue ouvrante et une diminution d'intensité dans une diphtongue fermante. Or, nos résultats, basés sur les modèles GAMM (modèles additifs généralisés à effets mixtes) révèlent un pattern différent de nos attentes : la dynamique d'intensité au sein de la diphtongue n'est pas liée à l'aperture vocalique. En revanche, conformément aux études précédentes, nous trouvons une corrélation positive entre la F0 et l'intensité. Nous nous interrogeons ainsi sur la validité de définir la sonorité à base de l'intensité seule. Enfin, nous discutons du rôle de la F0 dans la définition de la sonorité et l'apport de notre étude pour modéliser la sonorité dans une langue tonale.

ABSTRACT

Sonority is not intensity : the case of diphthongs in a tonal language.

This study explores the link between sonority and intensity in the production of rising and falling diphthongs in Beijing Mandarin. Considering that a low vowel is more sonorant than a high vowel, we expected an increase in intensity in a rising diphthong and a decrease in intensity in a falling diphthong. However, our results based on GAMM (Generalized Additive Mixed Models) reveal an unexpected pattern. That is, the change of intensity within a vowel is not related to vowel height. On the other hand, we find a positive correlation between F0 and intensity, in line with previous findings. We thus raise our concern about the legitimacy of defining sonority based on intensity alone. We conclude our paper with a discussion on the role of F0 in the definition of sonority and its implication for the modeling of sonority in a tonal language.

MOTS-CLÉS : intensité, sonorité, fréquence fondamentale, diphtongue, ton, Mandarin.

KEYWORDS: intensity, sonority, fundamental frequency, diphthong, tone, Mandarin.

1 Introduction

L'échelle de sonorité est largement utilisée en phonologie pour expliquer divers phénomènes synchroniques (tels que la formation syllabique et la resyllabation) et diachroniques (tels que la fortition et la lénition). Cependant, les corrélats phonétiques de sonorité sont débattus. Au niveau articulatoire, la sonorité est reliée au degré d'ouverture du conduit vocal. Au niveau perceptif, la sonorité est associée à la perception de la force sonore (voir Parker, 2002, pp. 43–48, pour une revue de littérature).

Quant aux corrélats acoustiques, de nombreuses tentatives ont été faites pour définir la sonorité à base de paramètres acoustiques tels que périodicité, intensité, durée, fréquence fondamentale (F0), structure formantique et enveloppe spectrale. La complexité est telle que certains chercheurs proposent d'abandonner complètement l'échelle de sonorité puisqu'il est difficile de la vérifier empiriquement (Ohala & Kawasaki-Fukumori, 1997). Parmi les paramètres acoustiques, le lien entre la sonorité et l'intensité est le plus recherché (par ex. Ladefoged & Johnson, 2015). D'autre part, l'intensité est corrélée avec la F0 et elle varie en fonction de multiples facteurs tels que la structure prosodique (Gordon & Roettger, 2017). Le présent article approfondit la question sur la relation entre l'intensité et la sonorité sous l'influence tonale à travers une étude sur les diphtongues en mandarin de Pékin.

1.1 Intensité et sonorité : classification des voyelles

Il est accepté que les voyelles sont plus sonores que toutes les autres classes des sons. Alors que la version minimale de l'échelle de sonorité ne donne pas de détail sur les différences de sonorité entre les voyelles (par ex. Clements, 1990), toutes les versions plus nuancées traitent les voyelles ouvertes comme plus sonores que les voyelles fermées (par ex. Selkirk, 1984). Ici, le lien avec l'intensité semble évident : toutes choses égales par ailleurs, une voyelle ouverte est intrinsèquement plus intense qu'une voyelle fermée (Fairbanks *et al.*, 1950; Rossi, 1971).

Les diphtongues, quant à elles, sont quasi-inexistantes dans les propositions de l'échelle de sonorité, sans doute en raison de la dynamique au sein de la voyelle. En revanche, deux catégories de diphtongues sont définies en termes de *dynamique* de sonorité : « montantes » et « descendantes ». Les diphtongues montantes telles que [ua] et [ia] sont caractérisées par une augmentation de sonorité, tandis que les diphtongues descendantes telles que [au] et [ai] par une diminution de sonorité (Jones, 1954). (En terme articulatoire, elles sont appelées « ouvrantes » et « fermantes », respectivement.) Selon Miret (1998), la cohésion d'une diphtongue est favorisée par une différence marquée de sonorité entre les deux parties vocaliques, donnant lieu à un pic de sonorité au sein de la syllabe, conformément au principe de sonorité séquentielle.

Toutes ces propositions reposent sur la prémisse selon laquelle il y a une forte corrélation entre la sonorité et l'intensité intrinsèque liée à la voyelle. Cependant, à notre connaissance, peu d'études mettent en évidence le lien entre la dynamique de sonorité et celle d'intensité dans une diphtongue. Ainsi, le but principal de notre étude consiste à vérifier si le lien entre sonorité et intensité est étayé par nos données sur les diphtongues en mandarin.

1.2 Intensité et F0 : le cas des langues à tons

Il est également reconnu que l'intensité varie en fonction de la F0 pour des raisons physiologiques comme la pression sous-glottique et l'adduction des plis vocaux, contrôlés par les muscles respiratoires et la configuration laryngée (par ex. les muscles crico-thyroïdiens et thyro-aryténoïdiens) (Van den Berg, 1957; Giovanni *et al.*, 2003; Honda, 2004). Cela constitue une des raisons pour lesquelles on trouve une corrélation positive entre la F0 et le niveau d'intensité dans les langues tonales comme le taiwanais et le mandarin, même si cette corrélation n'est pas linéaire (Zee, 1978; Whalen & Xu, 1992; Zhang, 2017). La question se pose donc de savoir comment l'intensité d'une voyelle est modulée par le ton lexical dans une langue tonale. Le deuxième but de notre étude est donc d'explorer cette modulation dans les diphtongues d'une langue tonale, le mandarin de Pékin.

1.3 La présente étude : diphtongues en mandarin de Pékin

Notre étude examine la dynamique d'intensité dans quatre diphtongues du mandarin de Pékin : deux diphtongues descendantes ([ai], [au]) et deux diphtongues montantes ([ia], [ua]). La comparaison entre ces diphtongues ainsi qu'entre elles et les monophtongues [i, u, a] nous permettrait de vérifier les hypothèses suivantes :

H1. Toutes choses étant égales par ailleurs, les voyelles ouvertes sont plus intenses que les voyelles fermées.
H2. Toutes choses étant égales par ailleurs, les diphtongues descendantes et montantes diffèrent entre elles en termes de changement dynamique d'intensité : les diphtongues descendantes diminuent en intensité et les diphtongues montantes augmentent en intensité.
H3. L'intensité vocalique est modulée par le ton lexical : pour la même voyelle, l'intensité augmente quand la F0 augmente.

Si H1 et H2 sont confirmées, nous anticipons un schéma de dynamique d'intensité comme illustré dans le cas de [i, a, ia, ai] dans la figure 1. Toutes choses égales par ailleurs (c.-à-d. les caractéristiques telles que la F0, la durée et l'accentuation étant comparables), [i] a une intensité plus faible que [a]. [ia] et [ai] se trouvent entre le niveau d'intensité de [i] et celui de [a], avec [ia] qui augmente en intensité et [ai] qui diminue en intensité. Quant à l'hypothèse H3, nous anticipons une modulation d'intensité sous l'interaction entre la F0 et le type de voyelle. Ensemble, ces questions contribuent à mieux comprendre la relation entre la sonorité, l'intensité, et la F0 à l'intérieur d'une diphtongue, et à explorer la base acoustique pour la bonne formation syllabique dans une langue tonale.

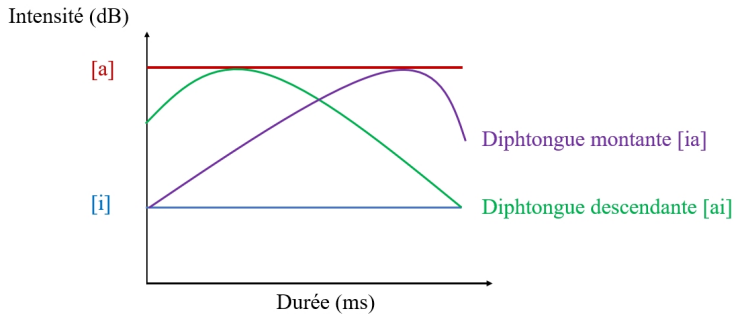


FIGURE 1 – Schéma qui prédit les intensités des monophtongues et des diphtongues montantes et descendantes, quelle que soit leur unité tonale.

2 Méthode

2.1 Participants et Matériels

Dix locutrices originaires de Pékin résidant à Paris ont été recrutées pour cette étude. Leur âge varie entre 20 et 27 ans, avec une moyenne de 23,7 ans. Aucune d'entre elles ne présente de trouble pathologique lié au langage. Les enregistrements ont été effectués entre novembre et décembre 2023 dans un studio d'enregistrement ou au domicile de l'expérimentateur, à l'aide d'un microphone serre-tête AKG-C520L positionné à environ 2,5 cm devant le côté inférieur droit de la bouche et une carte son KOMPLETE AUDIO 2. Les matériaux d'enregistrement sont des mots monosyllabiques

du mandarin pékinois, comprenant trois monophtongues [i], [u], [a], ainsi que deux diphtongues descendantes [ai] et [au] et deux diphtongues montantes [ia] et [ua]. La voyelle cible est précédée d'une occlusive, une affriquée ou une latérale selon les contraintes phonotactiques.

Pour les tons lexicaux, nous avons sélectionné pour notre analyse le ton 1 (plat et haut, transcrit comme "55" [1]), le ton 2 (montant, transcrit comme "35") et le ton 4 (descendant, transcrit comme "51"). Dans nos enregistrements, le ton 3 (descendant-montant, transcrit comme "214") est souvent prononcé sans sa partie montante (donc seulement "21") avec la voyelle devenue craquée et/ou dévoisée, ce qui affecte considérablement l'analyse de la F0 et de l'intensité. (La partie dévoisée est très diminuée en intensité, ce qui rend la comparaison d'intensité insensée quand le dévoisement n'est pas systématique ni entre les items ni à l'intérieur d'un item.) Par conséquent, le ton 3 n'a pas été inclus dans notre analyse. Les syllabes analysées sont répertoriées dans le tableau 1. Chaque syllabe a été produite d'abord en isolation, ensuite dans la position de focus d'une phrase cadre pour assurer une intonation comparable « X, zhè zèr zán běijīnghuà niàn X ba » (X représentant la syllabe cible) 'X, ce caractère est prononcé X en pékinois'. Pour cette étude, seule la syllabe prononcée dans la phrase cadre sera analysée.

Voyelle \ Consonne	p	t	k	k^h	l	tɕ	tʂ	Sans-attaque
i	pī pí pì	tī tí tì			lí lì			
u	pù	tū tú tù			lū lú lù			
a	pā pá pà	tā tá tà			lā là			
ai	pāi pái pài	tāi tài			lái lài		tʂāi tʂái tʂài	āi ái ài
au	pāu pàu	tāu tàu	kāu kàu		lāu láu làu			āu áu àu
ia						tɕiā tɕiá tɕià		iā iá ià
ua			kuā kuà	k^huā k^huà				uā uá uà

TABLE 1 – Syllabes analysées. Les syllabes sont transcrites avec le *pīnyīn*, le système officiel de romanisation des caractères chinois. L'accent macron représente le ton 1 (55), l'accent aigu représente le ton 2 (35) et l'accent grave représente le ton 4 (51).

2.2 Analyse de F0 et d'intensité

Nous avons extrait les valeurs de F0 et d'intensité sur 21 points temporels [2] équidistants du début jusqu'à la fin de chaque voyelle cible, à l'aide d'un script Praat (Boersma & Weenink, 2023). Afin de tester notre hypothèse, des modèles de GAMMs (Wood, 2017; Sóskuthy, 2017; van Rij *et al.*, 2022; Wieling, 2018) ont été ajustés aux données d'intensité en fonction du type de la voyelle. Le modèle de GAMMs nous permet de modéliser le changement non-linéaire des données et de comparer statistiquement la différence entre deux courbes. Nous avons établi 2 (monophtongue, diphtongue) $\times$ 3 (ton 1, 2 et 3) = 6 modèles pour comparer l'intensité des voyelles en fonctions du ton. Ils sont expliqués dans les codes suivants. Pour ces modèles, nous avons divisé les données en TYPE DE VOYELLE $\times$ TON.

Chaque modèle inclut des *smooths* de facteurs d'effet fixé de VOYELLE, CONSONNE et LOCUTEUR, ainsi que des *smooth* d'effet aléatoire non-linéaire de LOCUTEUR ajusté par CONSONNE et VOYELLE. Pour les *smooths*, nous avons choisi K = 21 parce qu'on a 21 points temporels.

Modèles = bam(Intensité ~Voyelle + s(Points temporels, bs='cr', k=21) + s(Points temporels, by=Consonne(by=Voyelle, Locuteur), bs='cr', k=21)

1. sur l'échelle de Chao (1930)
2. Toutes les voyelles sont supérieures à la durée minimale (105ms) requise de 21 points.

+ s(Points temporels, locuteur, m=1, xt=list(bs='tp')),bs='fs', k=21)+ s(Points temporels, locuteur, by=Consonne(by=Voyelle), bs='fs',k=21), data=nos données, method='fREML', AR.start=données$pointInitial, rho=rho calculé

Ensuite, nous voudrions explorer la relation entre l'intensité et les tons. Un modèle de la F0 ~ton est établi pour montrer le changement mélodique des tons. Comparé aux modèles intensité ~voyelle, le modèle de la F0 ~ton inclut un *random intercept* par LOCUTEUR et deux *random slope* par TON pour CONSONNE et VOYELLE :...s(locuteur,bs='re', k=21)+ s(tone,consonant,bs='re', k=21) + s(tone,vowel,bs='re', k=21)...

Enfin, compte tenu des erreurs possibles de l'intensité absolue, nous avons mesuré la valeur de l'intensité relative en prenant le point initial de la voyelle comme la référence et en soustrayant la valeur de référence de la valeur de chaque point qui suit ce point initial, et puis nous avons modélisé l'intensité normalisée en fonctions des tons. Par rapport aux modèles de l'intensité ~voyelle. Cette modèle inclut un facteur d'effet non-linéaire de VOYELLE adjusté par TON, CONSONNE et LOCUTEUR : ...s(Points temporels, voyelle, by=consonne(by=ton, locuteur), m=1, xt=list(bs='tp')), bs='fs', k=21) +...

3 Résultat

3.1 Intensité et voyelles

La figure 2 montre les modèles Intensité ~Voyelle pour chaque type de ton. Les monophtongues sont à gauche et les diphtongues sont à droite. La durée normalisée est en x et l'intensité en y. Les ombres de différentes couleurs représentent les intervalles de confiance (95%) de chaque courbe. Les résumés des modèles montrent que les déviances expliquées de tous les modèles sont supérieurs à 90%.

Rappelons-nous que l'hypothèse H1 prédit une intensité plus forte pour [a] que pour [i, u], et que l'hypothèse H2 prédit une opposition de direction dans la dynamique d'intensité entre les diphtongues descendantes [ai, au] et les diphtongues montantes [ia, ua]. Or, nos résultats ne permettent pas de montrer de différence d'intensité entre les voyelles. Puisque les intervalles de confiance des courbes se recouvrent entre eux, nous ne pouvons pas dire avec certitude qu'une classe de voyelles est plus intense qu'une autre. Ces résultats montrent un caractère non systématique lié à la qualité vocalique. Par exemple, la courbe d'intensité de [a] du ton 2 est plus élevée que [i] et [u] au début, mais elle est plus inférieure au milieu. La seule différence significative est trouvée sur la partie 50% – 80% des diphtongues du ton 4, où les diphtongues descendantes significativement plus intenses que les diphtoneus montante.

3.2 Intensité et tons

Nous allons maintenant regarder la relation entre l'intensité et le ton, pour vérifier l'hypothèse H3. L'image en haut à gauche de la figure 3 montre notre modèle F0 ~ton, qui explique 89.9% de déviances. Les tons 1 (55) et 4 (51) conforment à la transcription tonale sauf que le ton 4 commence plus haut que le ton 1. Le ton 2, au lieu de monter dès le départ, présente une courbe plus basse et un retard de montée par rapport à sa transcription. L'image en haut à droite de la figure 3 montre notre modèle de l'intensité normalisée ~ton (les déviances expliquées = 80.7%). Par rapport au modèle de la F0, la direction de changement d'intensité suit en général le changement de F0. Un test de corrélation de Pearson montre une corrélation positive et significative entre F0 et intensité (Pearson's $r = 0.407$, $t = 54.536$, $df = 14941$, $p < 0.001$).

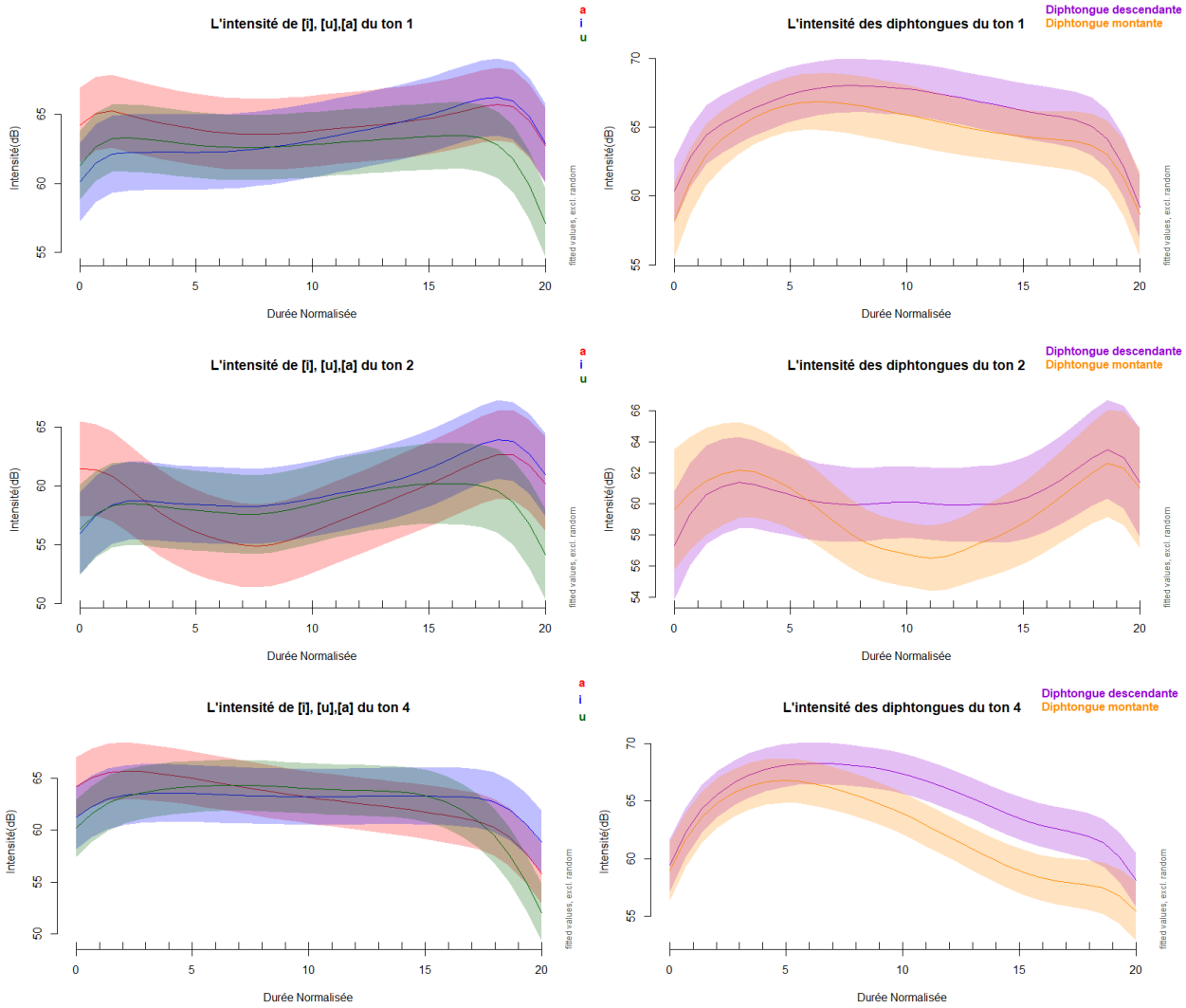


FIGURE 2 – Les courbes d'intensité prédites par les modèles de GAMMs pour les monophtongues [i], [u], [a] (à gauche), les diphtongues montantes[ia], [ua] et diphtongues descendantes [ai], [au] (à droite) pour chaque ton. L'intensité est représentée en y.

Les différences d'intensité entre les tons sont illustrées avec les packages *tidymv* (Coretta, 2023) et *ggplot2* (Wickham, 2016), présentées dans les trois images en bas de la figure 3 sous forme d'une courbe de deux couleurs. Cette courbe représente le chevauchement d'intervalle de confiance (95%) entre les deux courbes d'intensité de l'image en haut à droite, où la ligne $y = 0$ ne représente aucune différence de valeurs entre les deux courbes. La partie rouge représente une différence significative.

Les images montrent que les différences entre ces courbes d'intensité des différents tons sont statistiquement significatives. La courbe d'intensité relative du ton 1 est plus élevée que celle du ton 2 pour la majeure partie du milieu, de même que plus élevée que celle du ton 4 dans la seconde moitié de la voyelle. La courbe d'intensité relative du ton 4 est plus élevée que celle du ton 2 dans la première moitié, mais l'intensité du ton 4 diminue avec la baisse de F0, donc le ton 2 est plus élevé que le ton 4 dans la partie finale.

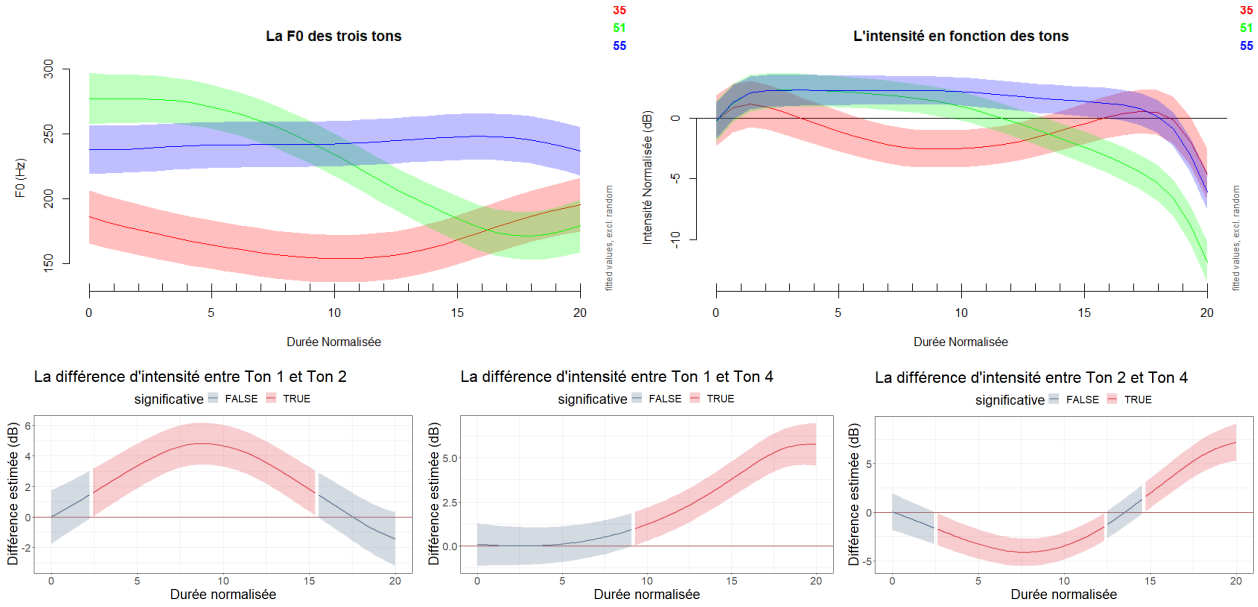


FIGURE 3 – Les courbes de F0 (en haut à gauche) et les courbes d'intensité (en haut à droite) prédites par les modèles GAMM en fonction des tons. Les différences entre les courbes d'intensité entre chaque paire de tons (en bas).

4 Discussion

En résumé, les hypothèses H1 et H2 ne sont pas validées ni par nos données des monophtongues ni par celles des diphtongues. Il n'y a pas de lien évident entre l'intensité et l'aperture vocalique pour les monophtongues ou à l'intérieur d'une diphtongue. En revanche, l'hypothèse H3 est validée par nos résultats : l'intensité augmente quand la F0 augmente.

Le résultat nous mène ainsi à poser des questions sur la validité de définir la sonorité basée sur l'intensité. Dans la littérature, nous trouvons des tentatives pour quantifier la sonorité à partir du seul paramètre d'intensité. Inspiré des données de Fry (1979), Koffi (2020) propose une formule logarithmique pour dériver la sonorité à partir de l'intensité mesurée sur un segment [3]. Parker (2002) conçoit une version des indices de sonorité en intégrant plus de 100 propositions antérieures, et trouve que la corrélation entre l'indice de sonorité et l'intensité est presque parfaite pour les segments en anglais et en espagnol. Selon l'auteur, cela justifierait la mesure de l'intensité seule pour déduire la sonorité. De même, bien que le rôle de la F0 soit reconnu, selon Ladefoged & Johnson (2015, p. 255), la quantification de la sonorité est basée sur l'intensité, à condition que la F0, ainsi que la durée et l'accentuation, soient comparables.

Nos données vont clairement à l'encontre de ces propositions. Si la sonorité dépendait de l'intensité, même en prenant en compte l'effet de la F0, nous devrions constater une différence d'intensité entre les types de voyelles pour la même catégorie tonale. Or, pour toutes les catégories tonales, y compris le Ton 1 où la F0 est égale, cette différence d'intensité liée à la voyelle n'est pas constatée (Figure 2). En revanche, nos résultats sur une corrélation positive entre l'intensité et la F0 sont cohérents avec les travaux précédents (voir §1.2). Une possibilité est que cette corrélation prévaut sur l'intensité intrinsèque des voyelles dans une langue tonale, peut-être parce que le rôle de l'intensité

3. index de sonorité $= 2 \times 10 \times \log 10 \times \frac{\text{Intensité maximale du segment}}{60 \text{ dB}}$

dans l'identification tonale est plus important que ce à quoi l'on pourrait s'attendre. Nous notons ici qu'afin d'isoler l'effet tonal de l'effet lié à l'intensité intrinsèque des voyelles, nous aurions besoin de plus de travaux sur l'intensité des voyelles, et notamment sur la dynamique d'intensité dans les diphtongues en comparant des langues non tonales à des langues tonales.

Ainsi, notre étude montre que l'intensité ne pourrait pas être utilisée pour dériver l'échelle de sonorité pour les segments vocaliques en mandarin de Pékin, à moins d'abandonner la notion de sonorité du moins pour les segments vocaliques. Alors, *quid* du rôle de la F0 dans la sonorité ? Bien que la F0 ait été proposée comme un des corrélats acoustiques de la sonorité (par ex. Nathan, 1989), ce n'est que récemment qu'une proposition très complète a été faite sur la modélisation de la sonorité à partir de l'intelligibilité de la hauteur de la voix ('*pitch intelligibility*'), estimée sur l'énergie des composantes périodiques ('*period energy*')[4] (Albert & Nicenboim, 2022; Albert, 2023). Autrement dit, la dynamique de la F0 y joue un rôle important, mais l'énergie périodique représente des données plus riches que la F0, en intégrant les dimensions spectrales et temporelles. Selon ces auteurs, la relation entre intensité et sonorité n'est que corrélationnelle, tandis que la relation entre l'intelligibilité de la hauteur et la sonorité est causale. C'est aussi pour cette raison qu'ils considèrent que cette relation causale est aussi importante pour expliquer la sonorité (au niveau segmental et syllabique) et la proéminence (au niveau prosodique) que ce soit dans une langue à tons ou à intonation.

Si cette proposition est sur la bonne voie, nous devrions avoir un autre regard sur la notion de sonorité. La sonorité (interprétée comme la perception de la force sonore) dépendrait de l'intégration de multiples paramètres, dont la F0 qui joue un rôle primordial dans l'intelligibilité de la hauteur. Si nous continuons ce raisonnement, dans une langue tonale en particulier, la bonne formation syllabique (c.-à-d. en suivant le principe de sonorité séquentielle) dépendrait non seulement des propriétés segmentales, mais aussi de leur intégration avec la dynamique de la F0 liée aux tons lexicaux. Nos études futures cherchent à creuser cette spéculation au moyen de plus de données empiriques, ainsi que de l'exploration du lien entre la perception de la force sonore et la modélisation de la notion de l'intelligibilité de la hauteur proposée par Albert & Nicenboim (2022).

5 Limites de l'étude

Comme un des reviewers l'indique, l'intensité liée à l'ouverture vocalique doit correspondre à l'intensité perçue (Rossi, 1971). Notre étude exploratrice est uniquement basée sur l'intensité en dB SPL sans pondération, comme dans les études citées dans §1.2 sur les langues tonales ainsi que certaines tentatives de l'index de sonorité (Koffi, 2020). Une mesure des intensités pondérées selon les bandes fréquentielles (par ex. pondération A) pourrait sans doute donner des résultats plus indicatifs.

Remerciements

Cette étude a bénéficié du soutien de l'Agence Nationale de la Recherche, dans le cadre d'une bourse de M2 du projet « Investissements d'Avenir » (référence : ANR-10-LABX-0083-LabExEFL), attribuée au premier auteur. Elle contribue à IdEx U. Paris (ANR-18-IDEX-0001). Nous remercions Jalal Al-Tamimi pour ses conseils sur les modèles GAMMs de cette étude.

4. periodic energy $= 10 \times \log 10 \times \frac{\text{periodic power}}{\text{periodic floor}}$

Références

ALBERT A. (2023). *A model of sonority based on pitch intelligibility*. language science press.

ALBERT A. & NICENBOIM B. (2022). Modeling sonority in terms of pitch intelligibility with the nucleus attraction principle. *Cognitive Science*, **46**, e13161.

BOERSMA P. & WEENINK D., Éds. (2023). *Praat : doing phonetics by computer [Computer program]*.

CHAO Y. (1930). A system of tone letters. *Le Maître Phonétique*, **3**, 1–30.

CLEMENTS G. N. (1990). The role of the sonority cycle in core syllabification. **1**, 283–333.

CORETTA S. (2023). tidymv : Tidy model visualisation for generalised additive models. R package.

FAIRBANKS G., HOUSE A. S. & STEVENS E. L. (1950). An experimental study of vowel intensities. *The Journal of the Acoustical Society of America*, **22**(4), 457–459.

FRY D. B., Éd. (1979). *The physics of speech*. Cambridge University Press.

GIOVANNI A., OUAKNINE M. & GARREL R. (2003). Physiologie de la phonation. *Encyl Méd Chir Paris : Elsevier SAS*, p. 20–632.

GORDON M. & ROETTGER T. (2017). Acoustic correlates of word stress : A cross-linguistic survey. *Linguistics Vanguard*, **3**(1), 20170007.

HONDA K. (2004). Physiological factors causing tonal characteristics of speech : from global to local prosody. In *Proc. Speech Prosody 2004*, p. 739–744.

JONES D. (1954). Falling and rising diphthongs in Southern English. *Le Maître Phonétique*, **32**, 1–12.

KOFFI E. (2020). A comprehensive review of intensity and its linguistic applications. *Linguistic Portfolios*, **9**(1).

LADEFOGED P. & JOHNSON K. (2015). *A course in phonetics [5th edition]*. Cengage learning.

MIRET F. S. (1998). Some reflections on the notion of diphthong. *Papers and studies in contrastive linguistics*, **34**, 27–51.

NATHAN G. S. (1989). Preliminaries to a theory of phonological substance : The substance of sonority. In R. L. CORRIGAN, F. R. ECKMAN & M. NOONAN, Éds., *Linguistic categorization*, p. 55–67.

OHALA J. & KAWASAKI-FUKUMORI H. (1997). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. *Language and its ecology : Essays in memory of Einar Haugen*, **100**, 343.

PARKER S. G. (2002). *Quantifying the sonority hierarchy*. Thèse de doctorat, University of Massachusetts Amherst.

ROSSI M. (1971). L'intensité spécifique des voyelles. *Phonetica*, **24**(3), 129–161.

SELKIRK E. (1984). On the major class features and syllable theory. p. 107–136. MIT press.

SÓSKUTHY M. (2017). Generalised additive mixed models for dynamic analysis in linguistics : A practical introduction. *arXiv preprint arXiv :1703.05339*.

VAN DEN BERG J. (1957). Subglottic pressures and vibrations of the vocal folds : Remarks on a high-speed film of piquet, decroix and libersa. *Folia Phoniatrica et Logopaedica*, **9**(2), 65–71. DOI : 10.1159/000262761.

VAN RIJ J., WIELING M., BAAYEN R. H. & VAN RIJN H. (2022). itsadug : Interpreting time series and autocorrelated data using gamms. R package version 2.4.1.

WHALEN D. H. & XU Y. (1992). Information for mandarin tones in the amplitude contour and in brief segments. *Phonetica*, **49**(1), 25–47.

WICKHAM H., Éd. (2016). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York.

WIELING M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling : A tutorial focusing on articulatory differences between l1 and l2 speakers of english. *Journal of Phonetics*, **70**, 86–116.

WOOD S. N., Éd. (2017). *Generalized additive models : an introduction with R*. CRC press.

ZEE E. (1978). Duration and intensity as correlates of f0. *Journal of Phonetics*, **6**(3), 213–220.

ZHANG X. (2017). *Les tons lexicaux du chinois mandarin en voix modale et en voix chuchotée*. Thèse de doctorat, Université de Strasbourg.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Le /r/ du mandarin est-il une fricative plutôt qu'une liquide ?

Yezhou Jiang[1], Rachid Ridouane[1], Pierre Hallé[1]
[1]Laboratoire de Phonétique et Phonologie (CNRS & Université Sorbonne Nouvelle)
4 Rue des Irlandais, Paris, France
yezhou.jiang@sorbonne-nouvlle.fr, rachid.ridouane@sorbonne-nouvelle.fr, pierre.halle@sorbonne-nouvelle.fr

RÉSUMÉ

Cette étude contribue au débat sur la nature phonologique de la consonne du mandarin notée <r> en pinyin : liquide ou obstruante ? /r/ ou /ʐ/ ? Nous savons que les clusters C1C2 sont d'autant plus sujets à la réparation perceptive C1C2 > C1əC2 que le profil de sonorité de C1C2 est marqué : pour C1=occlusive, nous devrions observer davantage de réparations lorsque C2 = /s/ que lorsque C2 = /l/. Qu'en sera-t-il avec C2=(/r/ présumé) ? Nous utilisons la difficulté de discrimination de C1C2-C1əC2 comme index de réparation de C1C2 (auditeurs mandarins ; stimuli produits par une bilingue mandarin-russe). Conformément aux prédictions, la discrimination est moins bonne pour /s/ que pour /l/. Mais de manière cruciale, la discrimination est aussi mauvaise pour le /r/ présumé que pour /s/ (pərou–prou ≈ pəsou–psou << pəlou–plou). Ces données suggèrent que la consonne notée <r> du mandarin est plutôt une obstruante qu'une liquide : /ʐ/ plutôt que /r/.

ABSTRACT

Is Mandarin /r/ a fricative rather than a liquid?

This study contributes to the debate on the phonological nature of the Mandarin consonant written <r> in pinyin: is it a liquid or an obstruent? /r/ or /ʐ/?

We know that C1C2 clusters are all the more prone to the C1C2 > C1əC2 perceptual repair that the C1C2 sonority profile is marked: For C1=stop, we anticipate more repairs for C2=/s/ than for C2=/l/. What about C2=(presumed /r/)? To assess C1C2 repairs, we took as a proxy the difficulty at discriminating between C1C2 and C1əC2 (Mandarin-speaking listeners; stimuli produced by a Mandarin-Russian bilingual). As expected, discrimination is poorer for /s/ than for /l/. Crucially, discrimination is as poor for the presumed /r/ as it is for /s/ (pərou–prou ≈ pəsou–psou << pəlou–plou). These data suggest that the Mandarin consonant written <r> should be phonologically classified as an obstruent rather than a liquid: that is, /ʐ/ rather than /r/.

MOTS-CLÉS : liquides, rhotiques, sonorité, perception, groupes consonantiques
KEYWORDS : liquids, rhotics, sonority, perception, consonant clusters

1 Introduction

Les consonnes liquides “rhotiques” (“R-sounds”) sont difficiles à définir. Habituellement, on les note collectivement /r/, même lorsque leur réalisation n’est pas le [r] de l’API ; on les transcrit avec la lettre “r” (ou un équivalent) dans les systèmes d'écriture alphabétiques. Il existe une grande variabilité dans la réalisation phonétique de ces consonnes, tant articulatoire qu’acoustique, selon les langues et les dialectes (Lindau, 1980 ; Spreafico & Vietti, 2013), les contextes et les locuteurs. Il y a cependant un consensus sur la place des liquides (dont les rhotiques) dans les hiérarchies de sonorité proposées (e.g., Clements, 1990): les liquides sont placées entre les glides et les nasales. Parmi les liquides, les rhotiques sont parfois considérées comme plus sonores que les latérales (Wiese, 2001, 2011). Elles sont en tout cas plus sonores que les obstruantes : un point crucial pour notre étude.

En mandarin, il existe un son transcrit en pinyin par la lettre <r>, qui apparaît principalement en position initiale mais aussi en position finale de syllabe, soit comme une coda dans quelques mots monosyllabiques (par exemple, 二, 耳, etc., tous transcrits <er>), soit, plus fréquemment, comme un suffixe rétroflexe -r, reflétant le processus morphophonologique appelé "erhua" (儿化). En position initiale de mot, la nature phonologique du son transcrit <r> est controversée. Pour cette position, les analyses phonologiques récentes suggèrent que le <r> du pinyin note une rhotique /r/, proche de l'approximante [ɹ] ou [ɻ] de l'anglais américain, qui forme un contraste avec la latérale /l/ : le contraste liquide /r/-/l/. Plusieurs études ont montré que ce /r/ supposé (pour simplifier, nous le notons désormais simplement ‘/r/’) est réalisé avec un F3 particulièrement bas, ce qui est vu comme une signature acoustique des approximantes rétroflexes, et souvent avec peu de bruit de friction, ce qui indiquerait une approximante plutôt qu’une fricative (Lee, 2005 ; Smith, 2010 ; Chen & Mok, 2019 ; Xing, 2021). Concernant l'articulation en position prévocalique, le /r/ mandarin est fréquemment articulé comme une consonne rétroflexe, similaire à celle du /r/ anglais américain. Pendant la production, la pointe ou la lame de la langue est soulevée vers le palais post-alvéolaire ou le palais dur, tandis que la partie antérieure de la langue reste relativement plate ou neutre. (Gick et al., 2006 ; Chen & Mok, 2021 ; Xing, 2021).

Du point de vue diachronique, les travaux en phonologie historique du chinois ont montré que le /r/ du mandarin n'est pas le réflexe d’une rhotique. Il provient d’une nasale alvéo-palatale voisée (*ȵ) du chinois archaïque (CA), qui a évolué en chinois moyen en une fricative alvéolo-palatale /ʑ/, pour finalement devenir la fricative rétroflexe /ʐ/ du mandarin (Baxter & Sagart, 2014). Le /ʂ/ du mandarin provient entre autres de la nasale sourde qui correspond à la sonore *ȵ en CA (Baxter & Sagart, 2014). Ces changements diachroniques sont en accord avec le point de vue traditionnel selon lequel le /r/ mandarin est une fricative /ʐ/ dont la contrepartie sourde est /ʂ/ (Karlgren, 1915 ; Chao, 1968 ; Ye, 1981 ; Duanmu, 2007 ; Lin, 2007). Dans les manuels actuels d'enseignement du chinois langue étrangère, le pinyin <r> est regroupé avec les fricatives et affriquées rétroflexes /ʂ/, /tʂ/ et /tʂh/ (<sh, zh, ch>), prononcées avec la pointe de la langue relevée, approchant ou touchant la partie avant du palais dur. Ce regroupement est phonologiquement justifié par une distribution similaire des rimes possibles après <sh, r, zh, ch>, distribution très différente de celle trouvée après /l/. Cet argument distributionnel suggère un appariement des sons transcrits <sh> et <r> (contraste /ʂ/-/ʐ/) plutôt que de <l> et <r> (contraste /l/-/r/). En particulier, les fricatives et affriquées rétroflexes, et non /l/, peuvent précéder la voyelle apicale rétroflexe /ʅ/. Inversement, /l/, et non le groupe rétroflexe, peut précéder les voyelles ou les semi-voyelles hautes et antérieures (/i, j, ɥ/). Tout cela suggérerait que le mandarin possède un contraste /ʂ/-/ʐ/ plutôt qu'un contraste /r/-/l/.

Il existe peu de recherches explorant la nature du /r/ mandarin du point de vue de la sonorité perçue. La sonorité perçue du /r/ mandarin est-elle plus proche de celle d'une liquide ou de celle d'une obstruante ? Nous tentons de répondre à cette question, en suivant la hiérarchie de sonorité la plus simple et consensuelle : voyelle > glide > liquide > nasale > obstruante (Clements, 1990).

Selon le Principe de Sonorité Séquentielle (PSS), les syllabes où l'attaque a un profil de sonorité montante sont mieux formées (moins marquées, plus acceptables) que celles où l'attaque a un profil "plateau" ou, pire encore, tombant (Clements, 1990 ; Berent et al., 2007). Berent et ses collègues (Berent et al., 2008, 2012) ont montré que les attaques plus marquées sont perçues moins fidélement et sont plus souvent "réparées" perceptivement, en général par insertion de voyelle épenthétique. Par exemple, Zhao et Berent (2016) ont étudié la perception par des locuteurs mandarins de clusters CC en position initiale, différant par le profil de sonorité. Le mandarin bannit tout cluster. Les sujets mandarins avaient tendance à percevoir les clusters CC comme des séquences CəC avec schwa épenthétique (par exemple, blif perçu bəlif). Cela se traduisait par une mauvaise discrimination des contrastes CəC-CC (par exemple, blif-bəlif). La discrimination était d'autant plus mauvaise que les profils de sonorité des CC étaient mal-formés. Les locuteurs du mandarin sont donc sensibles au PSS : Ils discriminent mieux bl-bəl que bd-bəd et mieux bd-bəd que lb-ləb, suivant la marque croissante bl < bd < lb. Nous exploitons ici cette sensibilité pour déterminer où se place, pour un locuteur mandarin, le /r/ mandarin dans l'échelle de sonorité : du côté des liquides ou des obstruantes ?

Les effets PSS en perception sont ainsi susceptibles d'éclairer la nature phonologique du /r/ mandarin (fricative ou liquide ?). Cette étude compare la sonorité perçue du /r/ mandarin à celle de la fricative /s/ et la latérale /l/, obstruante et liquide sans équivoque. Nous testons des auditeurs chinois sur leur discrimination de contrastes C1C2-C1əC2 en début de mot, qui varient selon le profil de sonorité de C1C2. C1 est toujours une occlusive et C2 est soit /l/, soit /s/, soit le /r/ mandarin. Pour que cette étude soit possible, les stimuli ont été produits par une locutrice bilingue sino-russe : en tant que locutrice du russe, elle pouvait parfaitement produire les clusters CC requis (le russe autorise en attaque des CC avec profil de sonorité montant, plateau, et même tombant) ; en tant que locurice du mandarin, elle pouvait parfaitement produire tous les C1s et C2s propres au mandarin, en particulier le /r/ mandarin comme C2 après une occlusive C1. Pour chaque type de contraste C1C2-C1əC2 (C2=/l, s, r/), le taux de non-discrimination et le temps de réponse indexent la difficulté de discrimination de ces contrastes (ou le taux de réparation perceptive C1C2 > C1əC2) qui doit dépendre du profil de sonorité de C1C2.

Tout d'abord, nous prédisons un taux de réparation plus élevé pour /s/ que pour /l/, étant donné les nombreuses résultats allant dans ce sens dans la littérature (Berent et al., 2008, 2012), et en particulier pour des locuteurs du mandarin (Zhao & Berent, 2016). Comme C1 est une occlusive, les clusters C1+/s/ (profil de sonorité plateau) devraient induire davantage de réparation que les clusters C1+/l/ (profil de sonorité montant donc mieux formé). La question cruciale porte sur les clusters C1+/r/ : se comportent-ils plutôt comme C1+/s/ ou plutôt comme C1+/l/ ? Le second cas suggérerait que le /r/ mandarin fonctionne comme un liquide, tandis que le premier cas indiquerait qu'il fonctionne comme une obstruante. Les cas intermédiaires seraient difficiles à interpréter.

2 Méthodes

2.1 Stimuli

Une locutrice bilingue mandarin-russe a enregistré cinq répétitions de 18 paires de non-mots du type {C1əC2R, C1C2R}, où C1 est l'occlusive /p/ ou /t/ ; C2 est /l/, /s/ ou le /r/ mandarin ; R est l'une des rimes suivantes, possibles en mandarin : /aŋ/, /oŋ/ ou /ou/ (Tableau 1).

belou-blou	belang-blang	belong-blong	delou-dlou	delang-dlang	delong-dlong
berou-brou	berang-brang	berong-brong	derou-drou	derang-drang	derong-drong
besou-bsou	besang-bsang	besong-bsong	desou-dsou	desang-dsang	desong-dsong

TABLEAU 1: Contrastes (notés en pinyin) utilisés pour l'expérience de discrimination.

Après analyse acoustique et prosodique sous Praat (Boersma & Weenink, 2024), trois stimuli ont été sélectionnés parmi les 5 répétitions de chaque item en sorte que (1) les CC ne contiennent pas de voyelle épenthétique et que (2) les contrastes CC-CəC soient minimalement marqués par des différences de contour f0 et/ou de durée. La fréquence fondamentale moyenne des stimuli sélectionnés varie entre de 240 à 270 Hz. Le pic d'intensité a été égalisé à 75 dB SPL. La durée de /ə/ dans CəCR a été égalisée à environ 40 ms en utilisant l'implémentation de PSOLA dans Praat.

Au cours de la sélection, nous avons observé des variations phonétiques intéressantes dans les /r/ mandarin de la locutrice bilingue. Pour les items du type CərR et CrR, environ 78% de ses /r/ avaient une structure formantique claire et peu de bruit de friction. Seuls ~22 % de ses /r/ présentaient un bruit de friction notable.

2.2 Participants

Vingt-quatre étudiants (15 femmes, 9 hommes) de 18 à 20 ans, recrutés à l'université Jimei en Chine, ont participé à l'expérience. Tous ont indiqué que leur langue maternelle était le mandarin et que leurs deux parents parlaient également le mandarin. Aucun d'entre eux n'avait vécu à l'étranger, mais tous avaient suivi des cours d'anglais à l'école pendant au moins neuf ans ; leur niveau moyen d'anglais auto-évalué sur une échelle de 1 à 5 était de 3,6 pour la compréhension et de 3,3 pour la production orale. Aucun ne souffrait de troubles auditifs ou linguistiques.

2.3 Procédure

Nous avons utilisé le paradigme AXB de discrimination, avec un intervalle inter-stimuli de 1 sec. et un time-out de 3 secondes pour les réponses. Chacun des 18 contrastes C1C2R-C1əC2R (2 C1 x 3 C2 x 3 rimes) était présenté 12 fois : pour chacune des quatre combinaisons AAB, BBA, ABB et BAA, trois triplets ont été construits par rotation des trois tokens des items A et B de sorte que chaque token apparaisse de façon équiprobable dans chacune des trois position et jamais deux fois dans le même essai. D'où un total de 18 x 12 = 216 essais test. La phase de test était précédée par une phase d'entraînement de huit essais faciles (e.g., belang-belang-belou) pour familiariser les participants avec la tâche.

L'expérience a été conduite en ligne avec PsyToolkit (Stoet, 2017). Il était demandé aux participants de passer l'expérience dans une pièce calme, avec aussi peu de distractions visuelles ou auditives que possible, de mettre leur téléphone en mode silencieux, et de porter des écouteurs avec le niveau sonore réglé à un niveau confortable. L'expérience proprement dite était accompagnée d'un questionnaire pour collecter les métadonnées sur les participants (en particulier sur leur background linguistique). Le questionnaire ainsi que toutes les instructions aux participants apparaissant à l'écran étaient rédigées en chinois.

À chaque essai, les participants recevaient trois stimuli (AXB) et devaient indiquer si le deuxième stimulus X leur semblait plus semblable au premier (A) ou au troisième (B) stimulus en appuyant sur l'une de deux touches (S ou L). Il leur était demandé de répondre aussi rapidement et correctement que possible. Les temps de réponse (RT) étaient mesurés à partir du début du troisième stimulus B. Au cours de l'entraînement, les participants recevaient un feed-back sur leur réponse : RT (en ms) affiché en vert ou rouge (correct ou incorrect) ou "trop lent" en cas de dépassement des 3 secondes. Au cours de la phase de test, ils ne recevaient plus que le feed-back de time-out. Chaque participant recevait les essais dans un ordre aléatoire différent. La durée de l'expérience était d'environ 30 minutes.

3 Résultats

La figure 2.a montre les pourcentages de discrimination correcte en fonction de la consonne critique C2 (/l/, /s/, ou /r/ : i.e., les trois contrastes testés Cəl-Cl, Cəs-Cs et Cər-Cr). La figure 2.b montre le détail des taux d'erreur de discrimination : selon C2, C1 (/p/ ou /t/) et R (rime -ang, -ong ou -ou). Considérant C2 comme le facteur principal, la performance de discrimination est la meilleure pour les contrastes Cəl-Cl (taux d'erreur moyen le plus bas : 14.6%), suivie des contrastes Cəs-Cs (taux d'erreur de 32.2%), puis Cər-Cr (taux d'erreur de 34.8%). Du point de vue de la rime, la performance est la meilleure pour les contrastes C1əC2ong-C1C2ong en '-ong' (taux d'erreur 15,7%), suivis par les contrastes en '-ang' et '-ou' (taux d'erreur 31,4% et 34.5%, respectivement). Nous avons analysé les données de taux d'erreur et de temps de réponse à l'aide de modèles linéaires mixtes (régression logistique pour les données de taux d'erreur) sous R (R Core Team, 2016), avec pour effets fixes principaux C1, C2, Rime et leurs interactions, et pour effets aléatoires les intercepts par sujets. Nous avons également inclus les effets fixes structurels Pattern (primauté vs. récence : AAB ou BBA vs. ABB ou BAA) et Target (X avec vs. sans schwa). La significativité de chaque effets a été évaluée en comparant les meilleurs modèles avec et sans cet effet (statistique Chi2). Nous avons effectué des comparaisons par paires entre les trois niveaux de C2 (/l, s, r/) en testant les modèles sur des sous-ensembles ne contenant que les deux niveaux à comparer.

Les effets fixes structurels n'étaient significatifs ni pour les taux d'erreur ni pour les RTs. Pour les taux d'erreur, C2 était significatif (Figure 2.a), avec moins d'erreurs pour /l/ que pour /s/ ($\chi^2(1) = 6.62$, $p < .01$) ou /r/ ($\chi^2(1) = 8.26$, $p < .001$), donc une meilleure discrimination pour Cəl-Cl que pour Cəs-Cs ou Cər-Cr. La différence entre /r/ et /s/ n'était pas significative ($\chi^2(1) = 0.10$, $p = .76$), indiquant une performance de discrimination similaire pour Cəs-Cs et Cər-Cr (taux d'erreurs 32.2% et 34.8%, respectivement). Rime était significatif (Figure. 2.b), avec moins d'erreurs pour '-ong' que pour '-ang' ($\chi^2(1) = 5.23$, $p = .02$) et '-ou' ($\chi^2(1) = 7.04$, $p < .01$), donc une meilleure discrimination pour CəCong-CCong. C1 n'était pas significatif ($\chi^2(1) = 0.66$, $p = .41$), malgré un taux d'erreur numériquement plus élevé pour /p/ (30.2%) que pour /t/ (24.2%).

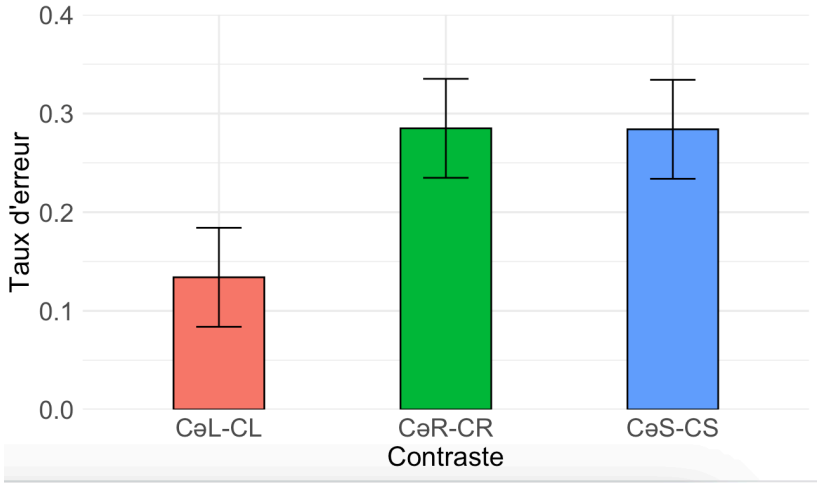


FIGURE 2.a. Taux d'erreur de discrimination selon C2 = /l/, /s/ et /r/

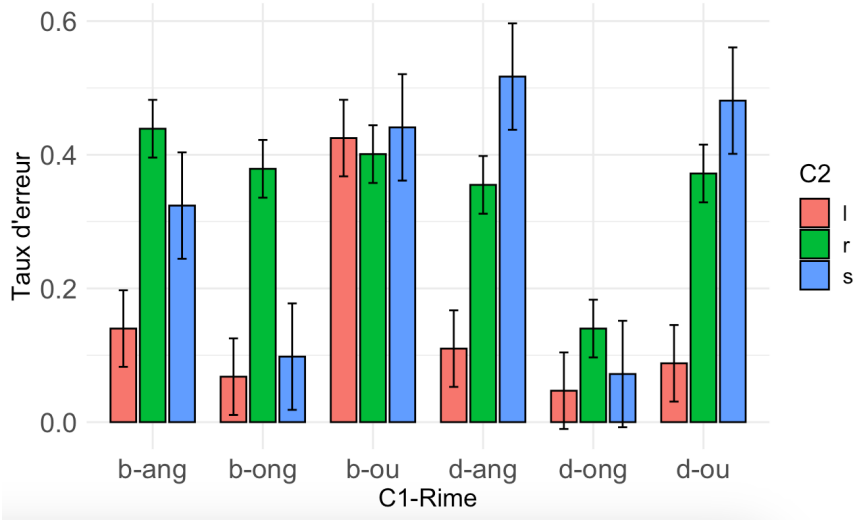


FIGURE 2.b. Taux d'erreur selon C1, C2, et la rime R

Pour les RTs (réponses correctes), les effets de C2 étaient parallèles à ceux trouvés pour les taux d'erreur. Les temps de réponse étaient plus courts pour /l/ que pour /s/ ou /r/ (938 ms < 1025 ou 1064 ms : $\chi^2(1) = 3.85$, $p < .05$ ou $\chi^2(1) = 7.93$, $p < .01$, respectivement). Les RTs ne différaient pas entre /r/ et /s/, $\chi^2(1) = 0.73$, $p = .39$. La Figure 3 montre les distributions des temps de réponses correctes pour /l/, /s/ et /r/. Encore une fois, nous constatons que la discrimination est plus facile pour C2=/l/ que pour /s/ ou /r/, et que /s/ et /r/ ne diffèrent pas de manière significative. En d'autres termes, les trois contrastes peuvent être regroupés en deux catégories : /l/ d'un côté et /s, r/ de l'autre. La discrimination est plus facile pour /l/ par rapport à /s, r/, suggérant moins de réparations par schwa épenthètique avec /l/ qu'avec /s, r/ comme C2 dans les clusters occlusive+C2.

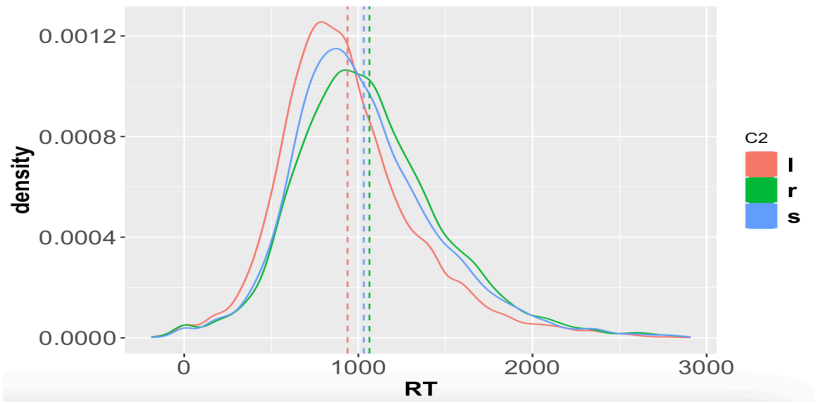


FIGURE 3. Distributions des temps de réponses correctes pour C2=/l/, /s/ et /r/

Le Tableau 2 montre un résumé d'une analyse ANOVA effectuée en parallèle sur les taux d'erreur, avec les facteurs intra-sujets C1, C2, Rime, Pattern et Target et les interactions C1 x C2 et C2 x Rime. Cette analyse confirme celle par modèles mixtes.

Facteur	d.f.	F	p
C1	1, 24	1.05	0.32
C2	2, 23	6.15	0.007
Rime	2, 73	8.98	0.0003
Pattern	3, 22	1.19	0.34
Target	1, 24	0.15	0.7
C1 x C2	5, 20	3.93	0.01
C2 x Rime	4, 67	2.00	0.1

TABLEAU 2. Sommaire de l'analyse ANOVA des taux d'erreur

4 Discussion

Nos résultats, qui comparent les performances de discrimination des auditeurs du chinois mandarin sur les contrastes Cəl-Cl, Cəs-Cs et Cər-Cr, montrent des performances plus faibles pour Cəs-Cs et Cər-Cr que pour Cəl-Cl, en termes de précision et temps de réponse. L'effet de réparation phonotactique par insertion de schwa est donc similaire pour les clusters occlusive+/s/ et occlusive+/r/, et plus fort que pour les clusters occlusive+/l/. Notre prédiction initiale que les clusters occlusive+/s/ de profil de sonorité plateau induisent davantage de réparation que les clusters occlusive +/l/ de profil montant est confirmée, en accord avec les résultats de la litérature (Berent et al., 2007 ; Zhao & Berent, 2016). Quant au /r/, les clusters occlusive+/r/ se comportent comme les occlusive+/s/ de profil de sonorité plateau. Nous devons en conclure que, pour les auditeurs du mandarin, ce /r/ présumé a une sonorité perçue bien plus proche de celle de /s/ que de /l/. Autrement dit, le /r/ mandarin se comporte perceptivement comme une obstruante plutôt qu'une liquide : le son transcrit par le pinyin <r> n'est pas une rhotique /r/ mais plutôt une fricative /ʐ/.

Nous avons noté que la plupart des '/r/' de nos stimuli (~78%) avaient peu de friction. La sonorité perçue d'une consonne n'est donc sans doute pas directement liée à sa réalisation phonétique.

Quelques remarques supplémentaires sont nécessaires. Tout d'abord, trois contrastes spécifiques: besong-bsong, desong-dsong et derong-drong ont été très bien perçus par les participants, avec des taux d'erreur de 9.8%, 7.2% et 14.1%, respectivement, bien inférieurs aux taux moyens de 35.7% et 28.9% pour les contrastes de type dəs-ds et dər-dr (Figure 2.b). L'examen détaillé des stimuli impliqués dans ces contrastes suggère des explications phonétiques, même si elles restent spéculatives. (1) Le /r/ dans drong-derong est réalisé proche du tap russe [ɾ] dans drong, mais proche du /r/ mandarin typique dans derong. On peut supposer que la locutrice bilingue est passé par inadvertance au russe pour les items "drong". Les données pour le contraste drong-derong auraient donc dû être écartées. (2) Pour les items "dsong", la séquence /t/+/s/ est plus étroitement coarticulée que les deux autres rimes, ce qui rend la séquence /ts/ plus facile à confondre avec l'affriquée /t͡s/ du mandarin (pinyin <z>) dans laquelle il est peu probable de percevoir un schwa épenthétique. (3) De même, pour les items "bsong", la séquence /p/+/s/ est plus étroitement coarticulée que pour les deux autres rimes, avec une confusion possible entre /ps/ et /p^{h}/ (pinyin <p>). L'aspiration perçue aiderait les auditeurs mandarins à distinguer besong de bsong. Enfin, le taux d'erreur pour les contrastes dəl-dl est assez faible (8.1%). Une explication plausible est la réparation perceptive /tl/>/kl/ que l'on trouve chez les auditeurs mandarins (Chen et al., 2022) tout comme chez les français ou américains (Hallé & Best, 2007). Le contraste dəl-dl équivaudrait ainsi au contraste /təl/-/kl/, logiquement plus facile à percevoir que /pəl/-/pl/. Un dernier aspect atténuant le pattern d'ensemble de la Figure 2.a est la difficulté de discrimination du contraste belou-blou (cf. Figure 2.b), pour laquelle nous n'avons pas d'explication.

Pour résumer ces remarques, différents aspects ont pu affecter nos données pour la rime '-ong' et pour l'initiale C1 /t/ suivie de /l/ ou /s/, rendant moins net le pattern global de sonorité perçue /l/ > {/s/ ≈ '/r/'}. Ces perturbations expliquent en partie l'effet du facteur Rime et de l'interaction C1 × C2. Malgré ces perturbations, le pattern global qui ressort est bien la hiérarchie de sonorité /l/ > {/s/ ≈ '/r/'}. Nous en tirons donc pour l'instant la conclusion que le /r/ mandarin est, en perception, une fricative plutôt qu'une liquide.

5 Conclusion

Nos résultats répliquent tout d'abord ceux de Zhao et Berent (2016) pour ce qui est du rôle du profil de sonorité dans l'acceptabilité perceptive des clusters dont le profil montant versus plateau est non équivoque (occlusive+/l/ vs. occlusive+/s/). Même si tous ces clusters sont *également* illégaux en chinois, ceux à profil montant sont plus acceptables que ceux à profil plateau : ils sont perçus plus fidèlement, induisent moins de réparation perceptive par épenthèse de schwa, et donc une meilleure perception des contrastes CC-CəC. Nous trouvons en effet que, par exemple, belang-blang est mieux discriminé que besang-bsang, tout comme Zhao et Berent rapportent que belif-blif est mieux discriminé que bedif-bdif. Ce qui est nouveau dans nos données relève de la consonne du mandarin transcrite <r> en pinyin. Cette consonne, un /r/ présumé, se comporte, du point de vue de la perception de la sonorité, comme un /s/ plutôt qu'un /l/. Autrement dit, ce /r/ présumé du mandarin se comporte en perception comme une obstruante plutôt qu'une liquide. Dans le débat encore non résolu sur la nature de ce '/r/' mandarin, nos résultats vont donc à l'appui du point de vue traditionnel : le pinyin <r> représente une fricative plutôt qu'une liquide.

Références

BAXTER W., SAGART L. (2014). Old Chinese: A New Reconstruction. Oxford University Press. New York. DOI: 10.1093/acprof:oso/9780199945375.001.0001.

BERENT I., STERIADE D., LENNERTZ T. & VAKNIN V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104 (3), p. 591-630. DOI: 10.1016/j.cognition.2007.01.006.

BERENT I., LENNERTZ T. & ROSSELLI M. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, 105(14), p. 5321–5325. DOI: 10.1073/pnas.0801469105.

BERENT I., LENNERTZ T., JUN J., MORENO M.A. & SMOLENSKY P. (2012). Universal phonological restrictions and language-specific repairs: Evidence from Spanish. *The Mental Lexicon*, 13, p. 275-305. DOI: 10.1075/ml.7.3.02Ber.

BOERSMA, P. & WEENINK, D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.05, retrieved 27 January 2024 from http://www.praat.org/.

CHAO Y. (1968). A Grammar of Spoken Chinese. University of California Press. Berkeley.

CHEN S., MOK P. (2021). Articulatory and acoustic features of Mandarin /ɹ/: a preliminary study. *12th international symposium on Chinese spoken language processing*, p. 1-5. DOI: 10.1109/ISCSLP49672.2021.9362070.

CHEN X., RIDOUANE R., HALLÉ P. (2022). Perception des clusters selon leur profil de sonorité : le cas des auditeurs du mandarin confrontés à des clusters russes. *JEP 2022 (34e Journées d'Études sur la Parole)*, p. 183-192. DOI: 10.21437/JEP.2022-20.

CLEMENTS G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology*, pp. 283-333. Cambridge: Cambridge University Press.

DUANMU S. (2007). The Phonology of Standard Chinese. Oxford University Press. Oxford & New York.

GICK B., CAMPBELL F., OH S. & TAMBURRI-WATT L. (2006). Toward universals in the gestural organization of syllables: A crosslinguistic study of liquids. *Journal of Phonetics*, 34(1), p. 49–72. DOI: 10.1016/j.wocn.2005.03.005.

GREENBERG J.H. (1978) in GREENBERG J.H, FERGUSON C.A, MORAVCSIK E.A, Éds., *Universals of Human Language*, Vol 2, p. 243-279.

HALLÉ P., BEST C. (2007). Dental-to-velar perceptual assimilation: a cross-linguistic study of the perception of dental stop+/l/ clusters. *The Journal of the Acoustical Society of America*, 121, p. 2899-2914. DOI: 10.1121/1.2534656.

KARLGREN B. (1915). Études sur la phonologie chinoise. Thèse de doctorat, Université d'Upsala.

LEE W. (2015). A phonetic study of the "er-hua" rimes in Beijing Mandarin. *INTERSPEECH*, 2005, p. 1093-1096. DOI: 10.21437/Interspeech.2005-433.

LIN Y. (2007). The sounds of Chinese. Cambridge University Press. Cambridge.

LINDAU M. (1980). The story of /r/. *The Journal of the Acoustical Society of America*, 67, S27. DOI: 10.1121/1.2018134.

MASSARO D., COHEN M. (1983). Phonological context in speech perception. *Perception and Psychophysics*, 34, p. 338-348. DOI: 10.3758/BF03203046.

R CORE TEAM. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. http://www.R-project.org.

SPREAFICO, L., & VIETTI, A. (2013). Rhotics: New data and perspectives. Bozen-Bolzano University Press. DOI: 10.13124/9788860461025.

SMITH J. (2010). Acoustic properties of English /l/ and /r/ produced by Mandarin Chinese speakers. Mémoire de master, University of Toronto.

WIESE R. (2001). The phonology of /r/. In D.G.Mouton: *In Distinctive feature theory*, p. 335-368. DOI : 10.1515/9783110886672.335.

WIESE R. (2011). The representation of rhotics. *In The Blackwell Companion to Phonology: Vol. vol.1*, p. 711-729. DOI : 10.1002/9781444335262.wbtcp0030.

XING K. (2021). Phonetic and phonological perspectives on rhoticity in Mandarin. Thèse de doctorat, University of Manchester.

YE M. (1981). The investigation of acoustic origin of the syllable-initial /r/. 日母音值源流考. *In Chifeng Institute Journal 赤峰学院学报*, 1981(1), p. 20-63.

ZHAO X., BERENT I. (2016). Universal restrictions on syllable structure: evidence from mandarin Chinese. *Journal of Psycholinguistic Research*, 01 Aug 2016, 45(4), p. 795-811. DOI : 10.1007/s10936-015-9375-1.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Le rythme : un marqueur d'atteinte du nerf laryngé supérieur ?

Hélène Massis[1], Marie-Hélène Degombert[2], Juliette Dindart[3], Diane Lazard[4], Christophe Trésallet[5], Frédérique Frouin[3], Claire Pillot-Loiseau[1],
(1) Laboratoire de Phonétique et Phonologie UMR 7018, 4 rue des Irlandais, 75005 Paris
(2) Cabinet Libéral, 78ter rue Guynemer, 92400 Courbevoie
(3) LITO, Inserm, Institut Curie, Université Paris Saclay, 91400 Orsay
(4) Service ORL et Maxillo-faciale, Centre hospitalier Princesse Grace, Monaco
(5) Service de chirurgie digestive, bariatrique et endocrinienne, hôpital universitaire Paris-Seine-Saint-Denis, hôpital Avicenne, 125 rue de Stalingrad, 93000 Bobigny
[helene.massis, claire.pillot]@sorbonne-nouvelle.fr, orthodegombert@gmail.com, [juliette.dindart, frederique.frouin]@curie.fr, dianelazard@yahoo.fr, christophe.tresallet@aphp.fr

RESUME

Après thyroïdectomie totale, la plainte vocale des patients, hors paralysie récurrentielle, est attribuée à l'atteinte du nerf laryngé supérieur (NLS) difficilement objectivable. Cette étude détermine si des paramètres rythmiques (ici temporels) de la parole de ces patients peuvent servir à son diagnostic. Elle a été menée chez 28 femmes avec suspicion d'atteinte de la branche crico-thyroïdienne du NLS (CT-), comparées à 27 autres sans dommage (CT+) après thyroïdectomie, au regard d'une population témoin (T). Les paramètres rythmiques étudiés montrent une diminution de la proportion des intervalles vocaliques sur tout l'énoncé, ainsi qu'une augmentation de la durée des intervalles consonantiques, moins variables, chez les CT-. La mobilisation des plis vocaux serait plus complexe et la réalisation des consonnes semblerait mettre en difficulté les CT-. Cette étude innovante sur le rythme des CT-, a tout son intérêt pour répondre aux attentes des patients et des professionnels de la voix.

ABSTRACT

Rhythm as a marker of upper laryngeal nerve damage?
After total thyroidectomy, patients' vocal complaints, excluding recurrent paralysis, are attributed to damage to the superior laryngeal nerve (SLN), which is difficult to objectify. The aim of this study was to determine whether rhythmic parameters (here temporal) of these patients' speech could be used in their diagnosis. It was carried out in 28 women with suspected damage to the crico-thyroid branch of the SLN (CT-), compared with 27 others without damage (CT+) after thyroidectomy, against a control population (T). The rhythmic parameters studied showed a decrease in the proportion of vowel intervals throughout the utterance, and an increase in the duration of the less variable consonant intervals in CT-. Mobilization of the vocal folds was more complex, and consonant realization appeared to be more difficult for CT- speakers. This innovative study of CT- rhythm is of great interest in meeting the expectations of patients and voice professionals.

MOTS-CLES : rythme, parole, nerf laryngé supérieur, thyroïdectomie, qualité vocale.
KEYWORDS : rhythm, speech, superior laryngeal nerve, thyroidectomy, voice quality.

1 Introduction

Les différents trajets anatomiques possibles de la branche externe du nerf laryngé supérieur (NLS) la rendent vulnérable aux pathologies de la structure la plus proche, à savoir la thyroïde, et aux interventions chirurgicales avec leurs possibles conséquences lésionnelles. La prévalence de l'atteinte du NLS oscille entre 14% et 58% des cas après chirurgie thyroïdienne (Barczyński et al., 2013 ; Neri et al., 2011). La lésion du NLS altère la contraction du muscle crico-thyroïdien (CT) provoquant un défaut d'élongation et de tension du pli vocal, et donc atteint la voix (Orestes & Chhetri, 2014). La prévalence de cette lésion est difficile à caractériser du fait d'un diagnostic encore incertain malgré tout le panel d'examens possibles. Sans outils diagnostiques fiables, cette lésion et la plainte étaient jusqu'à récemment encore peu étudiées et sous-estimées. Le terme de « suspicion d'atteinte » est d'ailleurs préférablement employé. Il convient alors d'écouter les patients et d'évaluer leur voix.

L'atteinte vocale a d'abord été particulièrement observée en voix chantée, mais les patients atteints d'une paralysie du NLS témoignent aussi communément d'une baisse de la fréquence fondamentale (f_o, Orestes & Chhetri, 2014 ; Roy et al., 2016), une difficulté à projeter la voix et l'impossibilité d'atteindre les sons aigus (Orestes & Chhetri, 2014). En conversation, la voix est décrite comme affaiblie, soufflée et monotone (Roy et al., 2016), mais aussi fatigable (Neri et al., 2011). Selon plusieurs auteurs, la fatigue vocale s'explique par l'installation de mécanismes compensatoires permettant d'augmenter la pression sous-glottique et la résistance laryngée (Orestes & Chhetri, 2014). Par des mesures aérodynamiques, Barczyński et al. (2013) objectivent l'augmentation de la pression sous-glottique et le débit d'air réduit lorsque le NLS est atteint, d'où un contrôle de l'air plus difficile au niveau des plis vocaux. Outre les phénomènes décrits précédemment, des interruptions inopinées, des pauses impromptues ou encore une accélération du débit de la parole comme phénomène de compensation sont attendues. La gestion vocale au quotidien de ces patients en proie à des difficultés pneumo-phonatoires pose question. Au-delà du trouble vocal lui-même, le rythme de la parole est donc ici étudié.

Le rythme est un élément incontournable pour le traitement de la parole. Du fait d'une notion faisant débat, s'accorder sur une définition du rythme est en soit un enjeu. Par rythme, nous évoquons ici uniquement des aspects temporels de la parole faisant appel au principe de Di Cristo (2013, 2016 dans Pillot-Loiseau et Xie, 2018) soit la construction d'« une alternance plus ou moins régulière de temps forts et de temps faibles ». Pillot-Loiseau et Xie (2018, p. 3) avancent qu'« une bonne maîtrise de l'aspect temporel d'une langue favorise l'intercompréhension entre interlocuteurs et ceux-ci tendent à considérer ce genre de production plus agréable et plus naturelle ». Le rythme soutient la production langagière et aide à la clarté du discours transmis (Arvaniti, 2012). Kohler (2009) insiste sur l'importance de son étude, très peu analysé dans la littérature. Ce paramètre est pourtant un argument majeur dans la classification des langues. Ainsi, on trouve un intérêt dans l'utilisation d'une telle méthode de catégorisation lorsqu'on l'applique à une même langue (Arvaniti, 2012). « Il semblait donc probable que les mesures [rythmiques] seraient en mesure d'identifier les déviations par rapport au rythme normal et serviraient donc d'outil de diagnostic » (Lowit, 2014, p. 2-3). L'auteur pense également que les paramètres rythmiques pourraient même servir à évaluer la sévérité d'un trouble et mesurer l'efficacité d'un traitement thérapeutique. Cette littérature démontre donc que l'étude du rythme chez une population pathologique est envisageable et pourrait mettre en lumière des altérations reflétant une certaine co-morbidité dans des pathologies d'origine vocale. Mener une telle étude sur la voix pathologique permet d'explorer un domaine peu étudié.

Plusieurs mesures rythmiques existent et reposent sur différents éléments de la production (voyelle, consonne, syllabe, segment). Le tableau 1 reprend ces mesures utilisées dans la littérature (Mairano & Romano, 2009 ; Pillot-Loiseau & Xie, 2018 ; Ramus, 1999b).

%V	Proportion des intervalles vocaliques de l'énoncé
ΔC	Ecart-type des intervalles consonantiques de l'énoncé
ΔV	Ecart-type de la durée des intervalles vocaliques de l'énoncé
Vmean	Moyenne de la durée des voyelles
Cmean	Moyenne de la durée des consonnes
VarcoΔC	Coefficient de variation de ΔC
VarcoΔV	Coefficient de variation de ΔV
rPVI-C	Variabilité brute des intervalles consonantiques successifs
rPVI-V	Variabilité brute des intervalles vocaliques successifs
nPVI-C	Variabilité normalisée des intervalles consonantiques successifs
nPVI-V	Variabilité normalisée des intervalles vocaliques successifs

TABLE 1 : Définition des paramètres temporels locaux (Mairano & Romano, 2009 ; Pillot-Loiseau & Xie, 2018 ; Ramus, 1999b)

L'objectif de cette étude[1] est de montrer que les mesures rythmiques peuvent distinguer des voix pathologiques après thyroïdectomie et en cas de suspicion d'atteinte du NLS, de voix non pathologiques. A partir d'épreuves de parole et d'une comparaison de sujets avec suspicion d'atteinte du NLS (CT-), sans atteinte (CT+) suite à une thyroïdectomie totale et de témoins (T), nous cherchons à définir des variables rythmiques permettant de discriminer la voix pathologique de la voix sans dommage. Il s'agit d'identifier un ou des marqueurs d'altération rythmique de la parole chez ces patientes, pour confirmer l'hypothèse du retentissement de l'atteinte sur le rythme de la parole et aider à l'objectivation de la plainte et au diagnostic en pratique clinique.

2 Matériels et méthode

2.1 Population et matériel

Cette recherche cible des femmes ayant subi une thyroïdectomie totale avec suspicion d'atteinte de la branche crico-thyroïdienne du NLS (population CT-). Ces patientes sont comparées avec des femmes ayant subi la même chirurgie mais n'ayant aucun dommage suspecté (population CT+), et avec une population témoin (T) appariée en âge à ces deux populations. Cette étude s'est déroulée en collaboration avec un chirurgien expert, ayant opéré toutes les patientes des populations CT- et CT+, et un chirurgien ORL expert du même service qui pose seul le diagnostic de suspicion d'atteinte du NLS sur la base de l'examen fibroscopique laryngé mais aussi des plaintes, le Voice Handicap Index (Jacobson et al., 1997) étant plus élevé chez les CT- (24,7/120) que les CT+ (11,2/120) et les témoins (6,4/120, voir Le Pape et al,. 2021 pour plus de détails). Les patientes sont âgées entre 24 ans et 85 ans et ont toutes été opérées entre 2016 et 2019. Elles ont été enregistrées entre 1 à 20 mois après leur date d'opération. Les examens cliniques et laryngoscopiques, en post-opératoire, ont éliminé toute atteinte de la mobilité des plis vocaux, toute lésion du nerf récurrent, myasthénie ou dystonie. Dans chaque groupe, les critères de non-inclusion comprennent toute personne non-

[1] Ayant bénéficié de l'autorisation CNIL n°2217748

francophone, toute personne atteinte de troubles neurologiques et/ou neuromusculaires, de maladie auto-immune, de troubles auditifs et cognitifs, toute personne avec antécédents de troubles vocaux et sans traitement hormonal de substitution. De même, n'ont pas été incluses les personnes bilingues avec un accent non francophone, influençant le rythme, ainsi que les personnes avec un faible niveau de lecture et les patientes avec une prise en charge orthophonique. Cette étude fait suite à celle de Le Pape et al. (2021) : ont été repris une partie de leurs enregistrements et de nouveaux sujets (témoins et patientes) ont été recrutés. Au total, a été constituée une cohorte de 82 participants. Elle comprend 28 patientes pour la population CT- (âge moyen : 56,6 ans ; 26-78 ans), 27 patientes pour la population CT+ (âge moyen : 57 ans ; 24-85 ans) et 27 sujets pour la population T (âge moyen : 55,7 ans ; 26-81 ans). Pour la population CT-, le délai entre la chirurgie et l'enregistrement est de 1 à 6 mois pour 11 patientes, de 7 à 12 mois pour 4 patientes et de 13 à 20 mois pour les 13 dernières. Ce délai est compris entre 1 à 6 mois pour la population CT+. Le matériel utilisé reprend le dispositif mis en place dans Le Pape et al. (2021) qui a suivi les recommandations du guide du *Committee on Phoniatrics* (Dejonckere et al., 2001). Les captations vocales ont été enregistrées grâce à un micro casque AKG C520 situé à deux doigts de la commissure des lèvres (angle de 45°). La voix des sujets a été recueillie avec une carte son Audient ID-4 sur le logiciel Audacity® via un ordinateur avec le système d'exploitation macOS Mojave version 10.14.3. Les enregistrements ont pu être effectués à l'hôpital ou chez le patient.

2.2 Corpus et mesures effectuées

Avant la passation des épreuves, une note d'informations a été donnée et le formulaire de consentement a été recueilli. Le corpus comprend la lecture d'un extrait de « La poupée rouge » de Pierre Gripari (1994). Pour cette étude, les enregistrements, identiques pour les trois populations, ont été annotés à l'aide du logiciel PRAAT (Boersma & Weenink, 2022) par deux examinateurs formés dans le cadre de ce protocole (47% des découpages effectués par le premier examinateur et 53% par le second examinateur). Toutes les annotations ont été reprises par un seul examinateur (premier examinateur) afin de respecter la conformité des découpages. Un accord intra-juge a été calculé afin de vérifier la fiabilité des découpages effectués. Un coefficient alpha de α=0,876 a pu être calculé et permet de conclure à une bonne fiabilité interne. Ces découpages comportent plusieurs niveaux : phrases, mots, syllabes et phonèmes. Les frontières droites des voyelles ont été positionnées à la disparition des premier et second formants. Le logiciel *Correlatore* (Mairano & Romano, 2009) a été utilisé pour extraire des fichiers TextGrid des variables rythmiques locales présentées au tableau 1. Les mesures temporelles globales (durée, débits de parole et articulatoire, durée des pauses, nombre et proportion des pauses) ont également été relevées. Notre analyse statistique compare ces mesures extraites des découpages des enregistrements entre les groupes CT+, CT- et T pour déterminer quelles mesures sont les plus altérées chez les patientes appartenant au groupe CT-. Elle a utilisé le test de Student avec le logiciel R.

3 Résultats

Les mesures temporelles globales ne sont pas significativement différentes entre les trois populations (figure 1). Concernant les mesures rythmiques locales (figure 2 et tableau 2), *%V* montre des différences significatives des moyennes entre les trois populations. La moyenne de *%V* des CT- est significativement inférieure à celle des CT+ et à celle de la population T. Cependant, les moyennes de CT+ et T ne se distinguent pas. *Cmean* des CT- est significativement supérieur à celle des CT+ mais aussi avec celle des témoins. On ne retrouve pas cette distinction entre les CT+ et les témoins T. La mesure de variabilité *nPVI-C* des patientes CT- est significativement inférieure à celle des CT+ et T. Aucune différence n'est trouvée entre les CT+ et les T. La mesure de variation *VarcoΔC*

des CT- est significativement inférieure à celle des patientes sans atteinte du NLS. Cette même différence se retrouve avec les témoins avec une meilleure significativité. On ne retrouve pas de différences notables entre les moyennes des CT+ et des T. Concernant les autres paramètres temporels locaux, aucune différence n'est retrouvée entre les scores des trois populations.

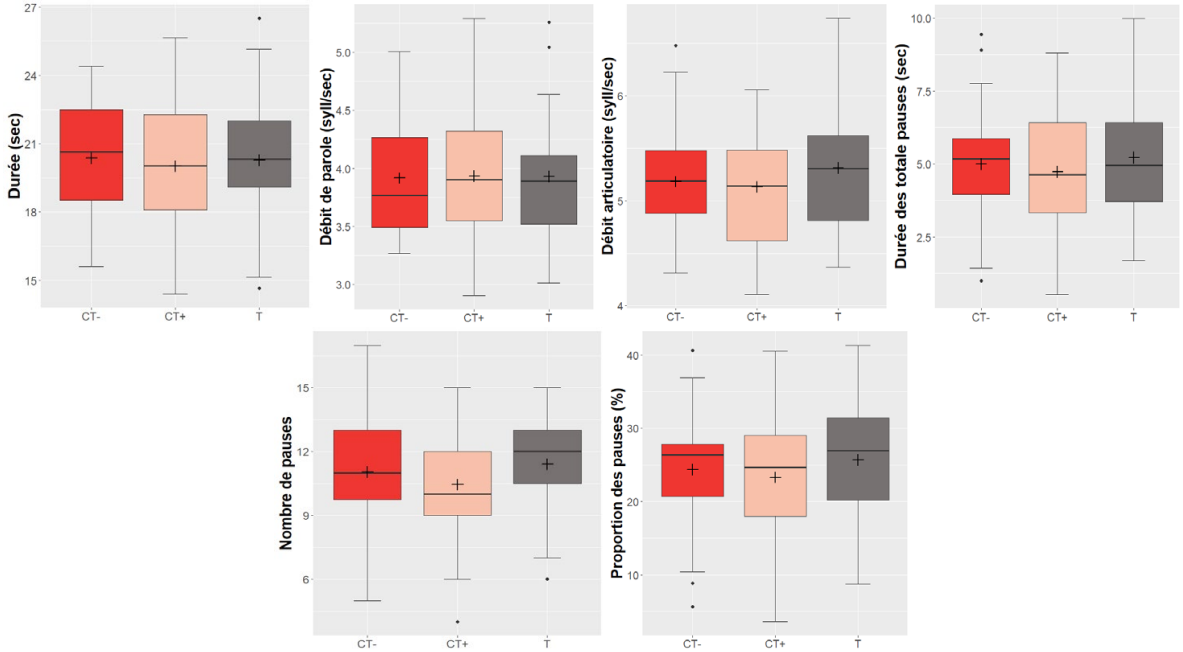


FIGURE 1 : Moyenne (+), médiane (trait horizontal) et variabilité de la durée de parole (Durée, en haut à gauche), du débit de parole (en haut au milieu), du débit articulatoire (en haut à droite), de la durée des pauses (en bas à gauche), du nombre de pauses (en bas au milieu) et de la proportion des pauses (en bas à gauche) des patientes post-thyroïdectomie avec suspicion d'atteinte du NLS (CT-, rouge), sans suspicion (CT+, saumon) et des femmes témoins (T, gris).

Test de Student

	%V (ddl 52)			*ΔC* (ddl 50)			*ΔV* (ddl 52)		
	M (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**
CT-	**51,9** (*2,52*)			**52,3** (*5,92*)			**55,3** (*14,7*)		
CT+	**55,2** (*4,02*)	0,0007		**50,1** (*5,34*)	0,165		**55,8** (*5,30*)	0,272	
T	**55,2** (*4,48*)	0,002	0,973	**51,6** (*8,15*)	0,728	0,433	**59,8** (*15,5*)	0,276	0,920
	Vmean (ddl 53)			*Cmean* (ddl 52)			*VarcoΔC* (ddl 53)		
	M (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**
CT-	**104** (*11,7*)			**101** (*9,45*)			**52,6** (*3,76*)		
CT+	**111** (*13,6*)	0,056		**93,7** (*12,3*)	0.022		**55,8** (*5,30*)	0,012	
T	**107** (*18,9*)	0,475	0,424	**89,5** (*7,59*)	0.00001	0,149	**57,6** (*7.39*)	0,002	0,304
	VarcoΔV (ddl 52)			*rPVI-C* (ddl 52)			*rPVI-V* (ddl 53)		
	M (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**
CT-	**52,7** (*10,2*)			**63,4** (*6,3*)			**52,8** (*11,2*)		
CT+	**54,4** (*9,8*)	0,546		**63,2** (*10*)	0,956		**57,8** (*14*)	0,152	
T	**54,7** (*8,9*)	0,441	0,887	**61,4** (*9*)	0,352	0,477	**56** (*12,8*)	0,330	0,627
	nPVI-C (ddl 53)			*nPVI-V* (ddl53)					
	M (*ET*)	**CT-**	**CT+**	**M** (*ET*)	**CT-**	**CT+**			
CT-	**62,3** (*4,22*)			**45** (*5,2*)					
CT+	**68** (*5,98*)	0,0001		**45,2** (*6,1*)	0,870				
T	**67,5** (*6,53*)	0,0009	0,755	**45,4** (*5,7*)	0,755	0,897			

TABLE 2 : Moyenne (**M**), écart-type (*ET*) et résultats du test de Student (p) pour les paramètres temporels locaux des patientes CT-, CT+ et des témoins (T).

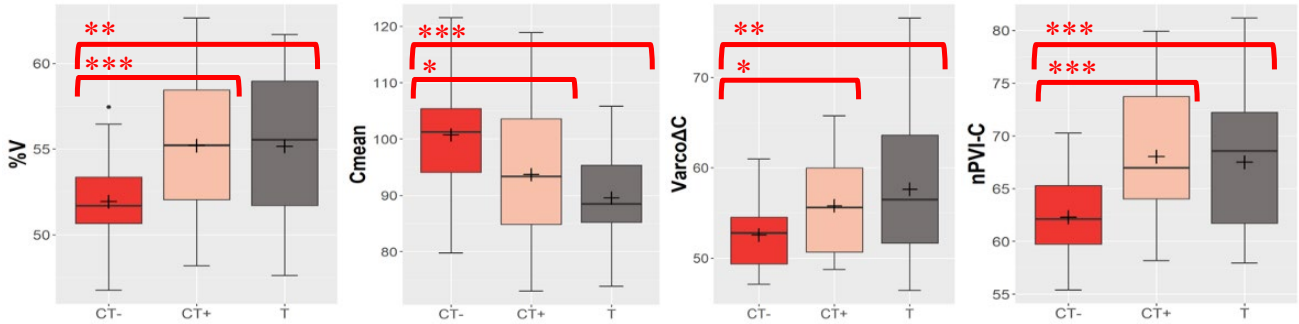


FIGURE 2 : Moyenne (+), médiane (trait horizontal) et variabilité du pourcentage d'intervalles vocaliques (*%V*, gauche), de la durée des intervalles consonantiques (*Cmean*, milieu gauche), du coefficient de variation de *ΔC* (*VarcoΔC*, milieu droite) et de la variabilité normalisée des intervalles consonantiques (*nPVI-C*, droite) des patientes post-thyroïdectomie avec suspicion d'atteinte du NLS (CT-, rouge), sans suspicion (CT+, saumon) et des femmes témoins (T, gris).

4 Discussion

Seules les mesures rythmiques locales *%V*, *Cmean*, *VarcoΔC*, et *nPVI-C* permettent de distinguer significativement les productions des patientes atteintes du NLS des deux autres populations. Le rythme dans son aspect temporel de la parole de ces patientes est donc impacté par la lésion suspectée. Du fait de la limitation des capacités de modulation vocale, une compensation sur le rythme est présente dans la parole de ces patientes.

Nous avons notamment identifié *%V* comme marqueur d'altération du rythme dans la parole issue de voix pathologiques, avec un score abaissé chez les CT-. Comme il n'est pas attendu en français qu'il y ait une grande variabilité de la durée des voyelles du fait de l'isochronie syllabique (Pillot-Loiseau & Xie, 2018), la différence des moyennes au-delà du seuil de significativité interroge sur les modifications de parole de notre population. Il semble que la réalisation du voisement et donc la mise en vibration des plis vocaux mette en difficulté les CT-. Cette observation est en lien avec les constatations déjà faites sur les difficultés rencontrées par les patients avec suspicion d'atteinte du NLS, autour de la réduction de la plage de variation de la fréquence fondamentale et de la fatigue vocale (Neri et al., 2011 ; Orestes & Chhetri, 2014). A la lumière de nos résultats, nous pourrions parler de l'apparition d'une certaine isochronie dans la production de parole des CT-.

En outre, des moyennes significativement plus élevées de *Cmean* (durée des consonnes) et plus basses de *VarcoΔC* (Coefficient de variation des intervalles consonantiques de l'énoncé) et de *nPVI-C* (Variabilité normalisée des intervalles consonantiques successifs) pour les CT- laissent penser à des modifications majoritairement localisées au niveau consonantique. Du fait d'une variabilité déjà observée au niveau des voyelles, il serait intéressant d'explorer plus en avant une possible modification au niveau du rapport des consonnes voisées et non-voisées. Les patientes avec suspicion d'atteinte du NLS allongent les consonnes, réduisent les intervalles vocaliques et varient peu la production des intervalles consonantiques. L'altération de ces paramètres témoigne de troubles rythmiques chez les patientes avec suspicion d'atteinte du NLS.

Le français étant une langue syllabique, il est également important de mettre en regard les données rythmiques globales avec le nombre de syllabes et la durée de la parole, dont il a été démontré que

les moyennes n'étaient pas significativement différentes entre nos trois populations. L'absence de variation significative démontre l'intérêt à pousser l'investigation rythmique au niveau des mesures locales.

Notre étude ouvre une nouvelle piste sur l'interaction entre la voix pathologique et la parole. Nos résultats indiquent qu'un trouble vocal est susceptible d'induire des altérations dans la parole continue. Les perturbations retrouvées se retrouvent principalement autour de la modification de la durée des consonnes. Elles reflèteraient une hyperarticulation tributaire des troubles vocaux. Cette observation induit la notion d'adaptation du discours au sein de l'*hypospeech* selon la théorie de Lindblom (1990). Les troubles vocaux acquis pourraient entraîner une production soudainement plus coûteuse sur le plan des mouvements articulatoires requis. La personne souffrant d'un trouble vocal pourrait ainsi ressentir l'importance d'insister sur la production consonantique afin d'assurer l'intelligibilité de son message, augmentant ainsi leur attente en termes de cible phonémique. Lindblom (1990) démontre déjà que lorsque le discours se veut plus « clair » -soit comprenant une surarticulation- cela induit une réduction de la durée des voyelles. Ainsi la théorie *H&H* de Lindblom (1990) permet de mieux appréhender les modifications observées sur nos données rythmiques. La réduction des durées vocaliques associée à une complexité dans la production des consonnes seraient un début de réponse expliquant les compensations retrouvées chez les patientes avec suspicion d'atteinte du NLS suite à une thyroïdectomie.

En outre, la majorité des études relatives à ce sujet sont faites en langue anglaise : Liss et al. (2009) et Lowit (2014) utilisent les variables rythmiques dans un but de distinction de populations d'une même langue. Cependant, aucune des études à notre connaissance ne porte sur l'étude du rythme de la parole dans le contexte particulier de troubles vocaux. De plus, Liss et al. (2009) ne font qu'une distinction intra-pathologie (pour des troubles de parole) et non population pathologique versus population témoin. Notre étude est donc une première dans l'utilisation des valeurs rythmiques dans l'objectif de diagnostiquer une pathologie vocale.

L'une des difficultés majeures de cette recherche se trouve dans le diagnostic initial de la pathologie étudiée. Aucun examen à ce jour ne permet de confirmer objectivement l'atteinte du NLS (Orestes & Chhetri, 2014). Cette observation est à la fois ce qui motive cette étude et ce qui la questionne.

En vue d'améliorer la validité de cette étude et pour de recherches futures, nous préconisons d'apparier les groupes de manière plus satisfaisante en termes de délai d'enregistrement. Nous avons inclus les patientes avec un délai entre l'opération et l'enregistrement de 20 mois maximum. Or, chez certains patients, une régression spontanée de la paralysie de la NLS suspectée peut être observée (Orestes & Chhetri, 2014). Barczyński et al. (2013) évoquent des récupérations spontanées mais aucune étude à notre connaissance ne permet de connaître avec précision ce délai de récupération. Il serait donc plus rigoureux de recruter une population de patientes à maximum 6 mois du geste opératoire afin de mieux garantir que la récupération spontanée n'a pas eu lieu et comparer ainsi des voix de patientes au plus proche de la chirurgie. La récupération possible de certaines patientes a pu influer nos résultats même si notre cohorte a permis de limiter cet écueil en partie, comme l'attestent nos résultats.

Par ailleurs, les conditions d'enregistrement n'ont pas été strictement identiques pour tous. Les gains réglés sur la carte son ont été conservés à l'identique mais nous n'avons pas pu procéder à un calibrage de l'intensité par sonomètre. De plus, certains sujets ont été enregistrés à leur domicile et non à l'hôpital comme la plupart des patientes de notre cohorte. Ainsi les enregistrements ont pu être impactés par ces différentes conditions.

5 Conclusion

Notre étude montre de manière inédite des paramètres rythmiques modifiés de la parole de patientes présentant une dysphonie particulière : *%V*, *Cmean*, *VarcoΔC* et *nPVI-C*. Au niveau des voyelles, une réduction vocalique témoigne de la difficulté de la mise en vibration des plis vocaux chez les CT-. Au niveau des consonnes, les paramètres altérés renvoient vers une augmentation de l'allongement dans la production des consonnes, une réduction des intervalles consonantiques et leur moindre variation. Le nombre de syllabes et la durée de parole étant proches entre nos populations, les variables mesurées sont robustes. Notre étude confirme l'hypothèse selon laquelle certaines composantes du rythme sont altérées dans la parole de patientes avec suspicion d'atteinte du NLS.

Malgré des opinions très diverses parmi les auteurs, les mesures rythmiques semblent donc pouvoir être investies d'une perspective diagnostique, non seulement pour quantifier, mais aussi pour qualifier la perturbation (Lowit, 2014). Afin de valider les mesures extraites des corpus, il serait intéressant de les comparer à des résultats perceptifs (Arvaniti, 2012 ; Lowit, 2014). Les études s'accordent sur l'idée que la comparaison des corpus avec une évaluation perceptive de locuteurs de la même langue maternelle permettrait de donner plus de valeur aux mesures. Du fait de la mise en évidence d'altérations particulièrement importantes autour de la production des consonnes, il serait également intéressant de poursuivre l'exploration des résultats sur cette piste. Nous pourrions constater une production particulière chez les patientes avec suspicion d'atteinte du NLS et les caractériser plus précisément.

Remerciements

Cette recherche a été financée par l'ANR "VOCALISE" (ANR-22-CE19-0035) et par le LabEx EFL(ANR-10-LABX-0083) qui contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

ARVANITI A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351-373. DOI : 10.1016/j.wocn.2012.02.003.

BARCZYŃSKI M., RANDOLPH G. W., CERNEA C. R., DRALLE H., DIONIGI G., ALESINA P. F., MIHAI R., FINCK C., LOMBARDI D., HARTL D. M., MIYAUCHI A., SERPELL J., SNYDER S., VOLPI E., WOODSON G., KRAIMPS J. L., HISHAM A. N. & THE INTERNATIONAL NEURAL MONITORING STUDY GROUP. (2013). External branch of the superior laryngeal nerve monitoring during thyroid and parathyroid surgery: International Neural Monitoring Study Group standards guideline statement: IONM During Thyroid Surgery. *The Laryngoscope*, 123, S1-S14. DOI : 10.1002/lary.24301.

BOERSMA P. & WEENINK D. (1992–2022) Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 23 January 2020 from https://www.praat.org.

DEJONCKERE P. H., BRADLEY P., CLEMENTE P., CORNUT G., CREVIER-BUCHMAN L., FRIEDRICH G., VAN DE HEYNING P., REMACLE M. & WOISARD V. (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology*, 258(2), 77-82. DOI : 10.1007/s004050000299.

KOHLER K. J. (2009). Rhythm in Speech and Language. *Phonetica*, 66(1-2), 29-45. DOI : 10.1159/000208929.

LE PAPE G., LAZARD D.-S., GATIGNOL P., TRESALLET C. & PILLOT-LOISEAU C. (2021). Modulation vocale, ressenti et branche motrice du nerf laryngé supérieur. *Annales françaises d'Oto-rhino-laryngologie et de Pathologie Cervico-faciale*, 138(4), 249-254. DOI : 10.1016/j.aforl.2020.05.018.

LINDBLOM B. (1990). Explaining Phonetic Variation : A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Éds.), *Speech Production and Speech Modelling* (p. 403-439). Springer Netherlands. DOI : 10.1007/978-94-009-2037-8_16.

LISS J. M., WHITE L., MATTYS S. L., LANSFORD K., LOTTO A. J., SPITZER S. M. & CAVINESS J. N. (2009). Quantifying Speech Rhythm Abnormalities in the Dysarthrias. *Journal of Speech, Language, and Hearing Research*, 52(5), 1334-1352. DOI : 10.1044/1092-4388(2009/08-0208).

LOWIT A. (2014). Quantification of rhythm problems in disordered speech: A re-evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130404. DOI : 10.1098/rstb.2013.0404.

MAIRANO P. & ROMANO A. (2009). Un Confronto Tra Diverse Metriche Ritmiche Usando CORRELATORE. In S. Schmid, M. Schwarzenbach, & D. Studer, *La dimensione temporale del parlato* (p. 79-100). EDK.

NERI G., CASTELLIO F., VITULLO F., DE ROSA M., CIAMMETTI G. & CROCE A. (2011). Post-thyroidectomy dysphonia in patients with bilateral resection of the superior laryngeal nerve: A comparative spectrographic study. *Acta Otorhinolaryngol Ital.*, 31(4), 228-234. PMID : 22065652; PMCID : PMC3203714.

ORESTES M. I. & CHHETRI D. K. (2014). Superior laryngeal nerve injury: Effects, clinical findings, prognosis, and management options. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 22(6), 439-443. DOI : 10.1097/MOO.0000000000000097.

PILLOT-LOISEAU C. & XIE H. (2018). Transfert rythmique du chinois mandarin au français dans l'apprentissage du Français Langue Étrangère : Acoustique et perception. *SHS Web of Conferences*, 46, 09001. DOI : 10.1051/shsconf/20184609001.

RAMUS F. (1999a). La discrimination des langues par la prosodie : Modélisation linguistique et études comportementales. *De la caractérisation à l'identification des langues, Actes de la 1ère journée d'étude sur l'identification automatique des langues*, 186-201.

RAMUS F. (1999b). *Rythme des langues et acquisition du langage* [École des Hautes Études en Sciences Sociales (EHESS)]. HAL : tel-00242452.

ROY N., FETROW R. A., MERRILL R. M. & DROMEY C. (2016). Exploring the Clinical Utility of Relative Fundamental Frequency as an Objective Measure of Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research*, 59(5), 1002-1017. DOI : 10.1044/2016_JSLHR-S-15-0354.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Nouvelle tâche sémantique pour le corpus de compréhension de parole en français MEDIA

Nadège Alavoine[1] Gaëlle Laperrière[2] Christophe Servan[1,2]
Sahar Ghannay[1] Sophie Rosset[1]
[1]Université Paris-Saclay, CNRS, LISN, Campus Universitaire bât,507 - Rue du Belvédère - 91405 Orsay
[2]Avignon Université, LIA, 339 chemin des Meinajariés - BP 1228 - 84911 Avignon Cedex 9
[3]QWANT, 10 bvd Haussmann - Paris 75009
{firstname.lastname}@lisn.upsaclay.fr

RÉSUMÉ

La détection d'intention et de concepts sont des tâches essentielles de la compréhension de la parole (SLU). Or il n'existe que peu de données annotées en français permettant d'effectuer ces deux tâches conjointement. Cependant, il existe des ensembles de données annotées en concept, dont le corpus MEDIA. Ce corpus est considéré comme l'un des plus difficiles. Néanmoins, il ne comporte que des annotations en concepts et pas en intentions. Dans cet article, nous proposons une version étendue de MEDIA annotée en intentions pour étendre son utilisation. Cet article présente une méthode semi-automatique pour obtenir cette version étendue. De plus, nous présentons les premiers résultats des expériences menées sur cet ensemble de données en utilisant des modèles joints pour la classification des intentions et la détection de concepts.

ABSTRACT

New Semantic Task for the French Spoken Language Understanding MEDIA Benchmark

Intention and concepts detection are essential tasks in speech understanding (SLU). There are a few annotated data sets in French, both in concepts and intention. However, there are some French datasets annotated in concept, including MEDIA. This French dataset, distributed since 2005 by ELRA, is one of the top SLU task. Unfortunately, it is only annotated in concepts and not in intent. In this article, we propose an improved version of MEDIA annotated with intentions to extend its use. This article presents the semi-automatic methodology used to obtain this improved version. In addition, we present the first results of experiments on this improved dataset using joint models for intention classification and concepts detection.

MOTS-CLÉS : Données d'évaluation, compréhension de la parole, détection jointe de l'intention et de concepts, tri-apprentissage.

KEYWORDS: Benchmark Dataset, Spoken Language Understanding, Joint Intent Detection And Slot-filling, Tri-training.

1 Introduction

Le module de compréhension de la parole (en anglais *Spoken Language Understanding* – SLU) est un élément crucial d'un système de dialogue oral. Les tâches de SLU regroupent trois sous-tâches : la classification de domaine, la détection d'intentions, et l'annotation de séquences en concepts sémantique (ou détection de concepts) (Tur & Mori, 2011). Dans cette étude, nous nous sommes

intéressés à la détection d'intentions et à la tâche de remplissage de formulaire. Cette dernière tâche peut également être considérée comme une tâche de détection de concepts (Bonneau-Maynard *et al.*, 2006).

La plupart des systèmes de dialogue traitent ces tâches séparément en développant des modules indépendants insérés dans un pipeline (Hakkani-Tür *et al.*, 2016). Ces approches pipeline souffrent généralement de la propagation d'erreurs en raison de leurs modèles indépendants. Ainsi, des modèles joints de classification des intentions et de détection de concepts ont été proposés pour résoudre ce problème et pour améliorer mutuellement ces deux tâches (Weld *et al.*, 2022). Pour ces modèles joints, plusieurs approches ont été explorées, telles que, des modèles fondés sur des champs conditionnels aléatoires (CRF) (Jeong & Lee, 2008), des réseaux de neurones convolutionels (Xu & Sarikaya, 2013), récurrents (Guo *et al.*, 2014; Hakkani-Tür *et al.*, 2016; Liu & Lane, 2016), avec des portes de remplissages (*slot-gated models*) (Goo *et al.*, 2018), avec mécanisme d'attention (Chen *et al.*, 2016; Hakkani-Tür *et al.*, 2016; Liu & Lane, 2016), et des modèles Transformers pré-entraînés (Chen *et al.*, 2019; Castellucci *et al.*, 2019; Wang *et al.*, 2020; Qin *et al.*, 2021; Han *et al.*, 2021) ou des modèles convolutionels graphiques (Tang *et al.*, 2020). Pour l'anglais, les modèles joints sont traditionnellement évalués sur des tâches annotées avec des intentions et de concepts : ATIS (Hemphill *et al.*, 1990) et SNIPS (Coucke *et al.*, 2018). Cependant, en français, moins de ressources sont disponibles. Il existe des données ATIS étendues au français dans le corpus MultiATIS++ (Xu *et al.*, 2020) par traduction. La version étendue MultiATIS++ ne dispose malheureusement pas des supports audios correspondants aux nouvelles langues. Les données MEDIA comportent les supports audios. Malheureusement, ces données ne sont annotées qu'en concepts et pas en intention, même si en l'état ce corpus est considéré comme l'un des plus difficiles (Béchet & Raymond, 2019).

Cet article présente une version mise-à-jour des données MEDIA avec des annotations d'intention à l'aide d'une approche semi-automatique. De plus, nous présentons les premiers résultats des expériences de compréhension sur cet ensemble de données amélioré à l'aide de modèles conjoints pour la classification des intentions et la détection de concepts.

2 Annotation du corpus MEDIA en intentions

Le corpus MEDIA est composé d'appels téléphoniques enregistrés pour la réservation d'hôtel. Il est dédié à l'extraction sémantique de l'information de la parole dans le contexte des dialogues homme-machine recueillis en utilisant la méthode Wizard-of-Oz (Bonneau-Maynard *et al.*, 2005). L'ensemble de données représente 1258 dialogues enregistrés de 250 différents locuteurs et environ 70 heures de conversations. Seuls les tours de parole des utilisateurs sont annotés avec des transcriptions manuelles et des annotations sémantiques complexes (concepts), et utilisés dans ce travail. Le corpus MEDIA est disponible en version *full* ou *relax*. Dans le second, les attributs sont simplifiés en excluant les spécifieurs. Récemment, Laperrière *et al.* (2022) ont proposé une version mise-à-jour de l'ensemble de données MEDIA. Le travail présenté ici est fondé sur cette version de MEDIA que nous notons MEDIA 2022 par la suite.

À notre connaissance, le corpus MEDIA n'a jamais été annoté avec des intentions. Contrairement à d'autres ensembles de données de référence comme ATIS (Hemphill *et al.*, 1990) ou SNIPS (Coucke *et al.*, 2018), seule la détection de concepts a été prise en compte. Nous proposons ici une version étendue du corpus MEDIA 2022 annotée en intentions. Pour cela, nous avons défini une liste de 11 intentions après avoir examiné attentivement le contenu de l'ensemble de données. Un tour de parole peut recevoir plusieurs intentions. Dans un tel cas, les labels sont séparées par le signe #. Les détails de cette liste, des exemples et des contre-exemples sont disponibles dans un guide d'annotation.

2.1 Méthodologie

L'annotation d'un ensemble de données peut être une tâche chronophage. Pour réduire le coût et le temps d'annotation, nous avons utilisé une méthode de tri-apprentissage (Zhou & Li, 2005) pour augmenter la taille des données d'apprentissage. Le tri-apprentissage est une méthode inductive épisodique semi-supervisée (van Engelen & Hoos, 2020) visant à améliorer les performances de n'importe quel type de système en ajoutant des données non-étiquetées. Il utilise un trio de classifieurs formés sur différents ensembles de données d'apprentissage. A chaque épisode de l'algorithme, ces classifieurs attribuent une *pseudo-étiquette* (Chen *et al.*, 2019) à des données non-étiquetées. Lorsque deux classifieurs du trio s'accordent sur une *pseudo-étiquette*, les données correspondantes *pseudo-étiquetées* sont ajoutées à l'ensemble d'apprentissage du troisième modèle. Lors de l'épisode suivant, les classifieurs pourront poursuivre leur apprentissage sur les ensembles de données mis à jour. L'algorithme de tri-apprentissage s'arrête quand aucun changement ne peut être observé dans l'apprentissage de tous les classifieurs du trio.

Récemment, Boulanger *et al.* (2022) ont montré que le tri-apprentissage pouvait être utilisé dans un cadre de faible ressource sur une tâche de reconnaissance d'entité nommée (REN). Nous avons décidé d'utiliser un système similaire pour entrainer et évaluer des trios de classifieurs. Le meilleur trio sera retenu pour annoter l'ensemble du corpus MEDIA en intention.

2.1.1 Ensembles de données pour le tri-apprentissage

Une partie des données annotées manuellement est nécessaire pour l'apprentissage et l'évaluation des trios de classifieurs. À cette fin, nous avons utilisé une version transcrite de l'ensemble de données MEDIA résultant du système de segmentation le plus couramment utilisé, avec des mots tronqués conservés comme mots entiers ("mer(ci)" est écrit "merci"). Cette version sera notée "MEDIA original". Un sous-ensemble d'énoncés choisis au hasard à partir de l'ensemble du corpus d'apprentissage original et d'autres choisis explicitement pour leur contenu ont été annotés manuellement en suivant notre guide d'annotation (disponible en ligne dans le dépôt). Cette annotation a été réalisée hors contexte : chaque déclaration a été traitée sans tenir compte des précédents énoncés dans le dialogue. 1551 énoncés ont été annotés manuellement pour le tri-apprentissage, 1240 constituant le corpus d'apprentissage, 124 pour le corpus de développement et 187 pour le corpus d'évaluation.

	Tri-apprentissage				**MEDIA**			
Intentions	**Appr.**	**Val.**	**Éval.**	**Total**	**Appr.**	**Val.**	**Éval.**	**Total**
annulation	15	1	1	17	32	1	15	48
incompréhension	6	1	4	11	273	30	94	397
marqueur_de_discours	38	6	5	49	282	40	113	435
modification	7	1	1	9	115	10	31	156
merci	47	5	6	58	713	100	200	1013
information	114	11	19	144	1611	159	401	2171
réponse_oui	392	42	52	486	4325	419	1190	5934
réponse_indécis	9	1	3	13	37	5	9	51
réponse_non	362	35	57	454	1315	88	344	1747
réservation	352	30	48	430	5437	522	1410	7369
salutation	43	8	6	57	717	101	206	1024

TABLE 1 – Distribution des étiquettes dans un sous-échantillonage du corpus MEDIA utilisé pour le tri-apprentissage (partie de gauche) et sur l'ensemble de données MEDIA (partie de droite). Ces sous-ensembles sont découpés en apprentissage (Appr.), validation (Val.) et évaluation (Éval.).

2.1.2 Protocole expérimental

Notre cas d'utilisation diffère du travail de Boulanger *et al.* (2022), car nous avons beaucoup de données non annotées. Nous avons adapté leur code en désactivant la génération de données synthétiques

et en modifiant le classificateur pour réaliser une pseudo-annotation *multi-label*. Il utilise l'état caché final du jeton spécial [CLS] combiné avec une couche sigmoïde et un seuil de $0,5$ pour déterminer l'intention à associer avec la phrase.

Nous avons utilisé deux modèles français de Transformers (Vaswani *et al.*, 2017) : CamemBERT (Martin *et al.*, 2020), un modèle dérivé de RoBERTA (Zhuang *et al.*, 2021), et FrALBERT (Cattan *et al.*, 2021), un modèle compact dérivé d'ALBERT (Lan *et al.*, 2020). Nous avons évalué deux versions comparables, formées sur 4 gigabytes (GB) de texte à partir du site Web de Wikipedia : CamemBERT-base-Wikipedia-4GB [1]et FrALBERT-base [2]. Cattan *et al.* (2022) ont montré que les classifieurs basés sur ces modèles avaient de bonnes performances en SLU sur l'ensemble de données d'évaluation MEDIA pour la tâche de détection de concepts (*slot-filling*).

Avant l'algorithme de tri-apprentissage, un échantillonnage aléatoire de 1000 énoncés parmi les 1240 qui constituent notre corpus d'apprentissage de tri-apprentissage est effectué pour chaque modèle de trio. Le réglage fin (*fine-tuning*) de nos modèles sur cette portion de données diminue les chances que les trois classifieurs produisent les mêmes résultats. L'algorithme est entraîné sur un maximum de 30 époques, bien qu'il s'arrête une fois qu'aucune variation du score d'évaluation n'est observée sur le corpus de validation. Les hyperparamètres sont fixés avec un taux d'apprentissage de 1e-5, une taille de lot d'apprentissage de 16 et une valeur de dropout de $0,1$. Le nombre maximal d'époques par épisode est de 1000, avec une méthode d'arrêt précoce de 20 époques.

Les performances des classifieurs sont évaluées au cours de l'apprentissage avec le rapport de correspondance exact (*Exact Match Ratio*, abrégé EMR) (Sorower, 2010) d'intention sur l'ensemble de validation. Une fois l'algorithme de tri-apprentissage arrêté, EMR, précision, rappel et F-mesure (ou score F1) sont évalués sur l'ensemble d'évaluation présenté dans le Tableau 1 (partie de gauche). Ces performances sont calculées sur les votes majoritaires des prédictions de trios de modèles.

2.1.3 Évaluation

Les résultats de nos expériences utilisant les données présentées dans la partie gauche du Tableau 1 sont présentés dans le Tableau 2. La plupart des expériences se sont arrêtées après 3 ou 4 époques de tri-apprentissage. Les trios utilisant le modèle CamemBERT obtiennent de meilleurs résultats que les trios utilisant FrALBERT en les dépassant de $7,17$ points sur le EMR et de $5,09$ points sur la F-mesure. Ils ont également moins de variabilité dans leurs résultats, avec un écart type oscillant entre $0,33$ et $0,70$ sur les différentes mesures par rapport à $0,81$ à $1,62$ pour FrALBERT.

Transformer	**EMR**	**Précision**	**Rappel**	**F1**
CamemBERT	92,09 $\pm 0,45$	95,29 $\pm 0,70$	93,48 $\pm 0,36$	93,73 $\pm 0,33$
FrALBERT	84,92 $\pm 1,62$	90,86 $\pm 0,81$	87,97 $\pm 1,44$	88,64 $\pm 1,37$

TABLE 2 – Résultats de classification en intentions avec l'algorithme de tri-apprentissage en utilisant une partie de l'ensemble de données MEDIA.

Suite à ces résultats, nous avons examiné de plus près les performances de notre meilleur trio de modèles de CamemBERT. Ce trio sera retenu pour annoter automatiquement l'ensemble de données MEDIA. Le trio obtient un EMR de $92,51$ points et une F-mesure de $93,85$ points par échantillon. En ce qui concerne la performance de la macro F-mesure, elle est à $58,99$ points. Cette macro F-mesure semble fortement influencée par une proportion importante de faux négatifs dans certaines

1. https ://huggingface.co/CamemBERT/CamemBERT-base-Wikipedia-4GB
2. https ://huggingface.co/qwant/FrALBERT-base

étiquettes, avec un macro rappel de $60,98$ points. En revanche, les faux positifs sont rares, avec une macro précision de $93,77$ points. Les faux négatifs concernent principalement les étiquettes avec peu d'exemples dans notre jeu d'évaluation présenté dans le Tableau 1 car ils n'affectent pas la moyenne de l'échantillon de rappel et de F-mesure autant.

2.1.4 Discussion, annotations, et corrections

Ce travail représente une première approche vers l'utilisation d'un algorithme de tri-apprentissage avec des classifieurs basés sur les Transformers pour annoter un ensemble de données. Puisque notre but était d'accélérer l'annotation, nous avons gardé la *pseudo-étiquette* pour lequel notre meilleure trio a obtenu un consensus. Pour chaque combinaison de *pseudo-étiquettes*, les énoncés correspondants ont été présentés à l'annotateur, qui a dû invalider les tentatives erronées. Les énoncés avec *pseudo-étiquettes* erronées ou sans aucune *pseudo-étiquette* attribuée, faute de consensus du trio de modèles, ont été annotés manuellement. Il y a eu 3122 intentions totalement ou partiellement erronées ($19,51$ % des données de 16005 *pseudo-étiquetées* et 137 de non-*pseudo-étiquetées*). L'ensemble des intentions étiquetées sur le corpus MEDIA sont présentées dans la partie de droite du Tableau 1.

2.2 Annotation de la version MEDIA 2022

La version MEDIA 2022 a également été annotée. Pour les ensembles d'apprentissage, validation et d'évaluation, la méthodologie utilisée diffère de celle décrite dans la section 2.1 puisque nous avions déjà les intentions associées à chaque énoncé. Une correspondance sur le contenu textuel des énoncés a été réalisée pour récupérer les annotations dans la mesure du possible.

3 Expériences sur les transcriptions manuelles

En utilisant l'ensemble enrichi des données MEDIA, nous présentons une première évaluation sur les transcriptions manuelles, en effectuant l'apprentissage joint des tâches de détection d'intentions et de concepts.

3.1 Architecture neuronale

Le modèle joint de détection d'intention et de de concepts utilisé dans nos expériences est le modèle JointBERT (Chen *et al.*, 2019). Pour la tâche de détection des intentions, ce modèle combine l'état caché final du jeton [CLS] à une couche *softmax*. Pour la détection de concepts, il détermine quel concept peut être associé à chaque mot en fournissant l'état final de chaque première sous-position d'un mot à une couche *softmax*. Le modèle est affiné en optimisant la somme des pertes d'entropie croisée pour les deux tâches.

La probabilité P_i pour qu'une phrase soit associée à une intention i passant $h_{[CLS]}$ (le dernier état caché du transformeur pour le jeton[CLS]) à une couche de poids W^i et de biais b^i est définie comme suit : $P_i = sigmoide(W^i h_{[CLS]} + b^i) > 0,5$. Une perte d'entropie croisée binaire remplace la perte d'entropie croisée précédemment utilisée pour la classification des intentions. Le modèle est optimisé sur la somme des pertes d'entropie croisée binaires et non binaires pour la classification d'intentions et la détection de concepts respectivement.

3.2 Protocole expérimental

Pour la tâche de détection de concepts, nous avons utilisé un format BIO. Les performances sont évaluées en termes de micro F-mesure, couramment utilisée pour les modèles joints (Weld *et al.*, 2022), et de Concept Error Rate (CER) qui est la mesure officielle utilisée dans la campagne MEDIA (Bonneau-Maynard *et al.*, 2006). Pour une comparaison avec les expériences sur les sorties de reconnaissance automatique de la parole (RAP), nous utilisons la micro F-mesure calculée sur des vecteurs *multi-hot* (abrégée F1mh) de concepts présents dans les annotations attendues et obtenues.

Pour la classification des intentions, lorsqu'il y a plusieurs intentions, nous les concaténons en utilisant le caractère (#). Dans la plupart des modèles conjoints, la performance de cette tâche est évaluée à l'aide de l'exactitude (*accuracy* en anglais, abrégée Exa.) (Weld *et al.*, 2022). Comme nous utilisons un système de classification *multi-label*, l'exactitude proposée par Godbole & Sarawagi (2004) et EMR ont été évaluées. L'exactitude du cadre sémantique de la phrase (*sentence-level semantic frame accuracy*, abrégée SFA) - correspondant au nombre d'énoncés avec une intention et des concepts correctement trouvés divisés par le nombre de phrases - couramment utilisée pour les modèles joints (Weld *et al.*, 2022), est également évaluée.

Nous avons choisi le modèle de base CamemBERT entrainé sur 135 GB de texte de CCNET (CamemBERT-base-CCNET) (Martin *et al.*, 2020) ainsi que les modèles de base CamemBERT-base-Wikipedia-4GB et FrALBERT utilisés précédemment. Ces modèles ont montré des résultats à l'état-de-l'art , ou proches de ceux-ci, pour la détection de concepts sur les transcriptions manuelles MEDIA (Ghannay *et al.*, 2020; Cattan *et al.*, 2022). Nous avons également choisi un modèle français BERT, FlauBERT, optimisé pour quelques époques sur des données transcrites par RAP (FlauBERT-oral-ft) (Hervé *et al.*, 2022) qui a également obtenu des performances proches de l'état-de-l'art sur les sorties MEDIA ASR (Pelloin *et al.*, 2022).

De la même manière que Cattan *et al.* (2022), nous avons utilisé un algorithme génétique, le *population based training* (abrégé PBT) (Jaderberg *et al.*, 2017), pour déterminer les meilleurs hyperparamètres.

3.3 Résultats sur les transcriptions manuelles

Les performances sur la version originale et relax de MEDIA, sur la version relax de MEDIA 2022, et la version full de MEDIA 2022 sont affichées dans la partie gauche du Tableau 3.

En ce qui concerne les scores de la tâche de détection de concepts, nous pouvons logiquement observer que les modèles obtiennent de meilleurs résultats sur la version relax que la version full. Par exemple, sur MEDIA 2022, il y a une différence de $2,42$ points de F-mesure pour les meilleurs résultats obtenus entre la version relax et la version full, en faveur de la version relax. Plus surprenamment, tous les modèles fonctionnent mieux sur les deux tâches avec la version relax d'origine que sur la version MEDIA 2022 relax. Cela pourrait s'expliquer par la conservation de mots tronqués dans MEDIA 2022, qui compliquerait ces tâches.

4 Expériences sur la reconnaissance automatique de la parole

Nous évaluons les performances des deux approches à l'aide des mesures utilisées dans la section 3.2, à l'exception de la F-mesure de la détection de concepts.

	Transcriptions manuelles						Cascade avec RAP				
	Intentions		Concepts				Intentions		Concepts		
Modèle	**Exa.**	**EMR**	**F1**	**F1mh**	**CER**	**SFA**	**Exa.**	**EMR**	**F1mh**	**CER**	**SFA**
MEDIA original, relax											
CamemBERT-base-CCNET	93,87	91,79	88,52	**95,97**	8,68	**76,26**	92,07	89,82	**93,82**	13,93	**65,69**
CamemBERT-base-Wikipedia-4GB	**93,98**	**91,84**	87,93	95,41	9,34	75,58	**92,43**	**90,08**	93,29	15,03	65,49
FlauBERT-oral-ft	93,66	91,19	87,93	95,63	8,95	76,04	92,28	89,77	93,12	14,19	65,55
Base FrALBERT	92,27	89,88	84,24	93,66	13,14	72,12	90,81	88,18	91,14	19,97	62,91
MEDIA 2022**, relax**											
CamemBERT-base-CCNET	91,87	**89,78**	86,95	94,66	10,33	72,68	90,16	**88,00**	92,74	12,78	64,43
CamemBERT-base-Wikipedia-4GB	91,25	88,66	86,88	94,88	10,24	72,60	89,39	86,75	92,90	13,19	64,00
FlauBERT-oral-ft	**92,10**	89,73	**87,75**	**95,41**	**9,18**	**73,29**	**90,40**	87,95	**93,40**	**11,93**	**64,96**
FrALBERT-base	90,71	88,37	82,48	92,94	14,71	69,18	89,29	86,86	91,00	17,81	62,01
MEDIA 2022**, full**											
CamemBERT-base-CCNET	92,28	89,73	**85,33**	**92,87**	**11,61**	72,13	**90,86**	**88,29**	**91,06**	**14,18**	**64,14**
CamemBERT-base-Wikipedia-4GB	91,81	89,25	85,24	92,42	12,11	**72,15**	90,62	88,13	90,60	15,19	63,66
FlauBERT-oral-ft	**92,31**	**89,97**	84,26	92,34	12,68	71,54	90,23	87,76	90,68	15,11	63,10
FrALBERT-base	90,64	88,29	80,10	90,38	17,40	68,04	88,89	86,25	88,22	20,16	61,24

TABLE 3 – Meilleures performances de nos modèles optimisés par PBT sur le jeu d'évaluation de MEDIA. Les résultats ont été obtenus à partie des transcriptions manuelles (partie gauche du tableau) ou des sorties RAP (approche en cascade, partie droite).

4.1 Approche en cascade

L'approche en cascade (ou séquentielle) consiste à utiliser deux modules séparés et se suivant pour résoudre des problèmes spécifiques. Ici, un module de RAP est suivi d'un module de compréhension.

Le modèle RAP utilisé pour l'approche en cascade est constitué d'un encodeur de parole (LeBenchmark FR 3k large (Evain *et al.*, 2021)), suivi de 3 couches de bi-LSTM et d'une couche linéaire de 1024 neurones. Le modèle utilise l'optimiseur Adam avec un taux d'apprentissage de $0,0001$, tandis que la couche de sortie linéaire utilise un optimiseur Adadelta avec un taux d'apprentissage de $1,0$. La fonction de coût de la classification temporelle connexionniste (CTC) est optimisée pour 100 époques, visant le meilleur taux d'erreur de mots (WER pour *Word Error Rate* en anglais) possible. Nous obtenons un score de $9,49\%$ de WER sur MEDIA 2022 et de $10,51\%$ sur l'ensemble de données MEDIA original avec ce système.

Les sorties du module de RAP sont ensuite transmises au modèle joint présenté dans la section 3.1. Ce second modèle prédit donc les informations sémantiques (concepts et intentions) à partir des transcriptions automatiques. Les résultats du système cascade sont présentés dans la partie droite du Tableau 3.

4.2 Approche bout-en-bout

Une approche de bout-en-bout (*end-to-end* en anglais) vise à développer un système unique, directement optimisé pour extraire des informations sémantiques de la parole sans utiliser de transcriptions intermédiaires. Notre modèle de bout-en-bout est composé de l'encodeur de parole SAMU-XLSR original (Khurana *et al.*, 2022), ou SAMU-XLSR$_{IT \oplus FR}$ spécialisé (Laperrière *et al.*, 2023), suivi de deux blocs de décodage différents de 3 couches bi-LSTM de 1024 neurones. Chaque bloc est suivi d'une couche entièrement connectée de la même dimension, activée avec *Leaky ReLU* et une fonction *softmax*. Un bloc est optimisé pour produire les intentions des segments audio, tandis que l'autre exécute la tâche de détection de concepts de MEDIA.

Nous avons optimisé les fonctions de coût (*loss* en anglais) CTC sur 100 époques avec les mêmes optimiseurs que ceux utilisés dans l'approche en cascade, à l'exception de la couche linéaire de

l'optimiseur de classification des intentions dont le taux d'apprentissage est fixé à $0,1$. La somme des deux *loss* est définie comme suit : $loss = \frac{1}{4} * loss(intent) + loss(slot)$.

	Intent		Slot-filling		
Modèle	**Accuracy**	**EMR**	**F1mh**	**CER**	**SFA**
MEDIA original, relax					
SAMU-XLSR	**92,02**	**90,14**	91,68	16,01	71,57
SAMU-XLSR$_{IT \oplus FR}$	91,74	90,02	**92,35**	15,44	**72,51**
LeBenchmark FR 3k large	91,75	89,91	91,98	15,51	71,12
MEDIA 2022**, relax**					
SAMU-XLSR	**90,74**	**88,85**	90,01	15,28	68,50
SAMU-XLSR$_{IT \oplus FR}$	90,53	88,64	90,65	15,16	70,83
LeBenchmark FR 3k large	90,18	87,98	**90,77**	**15,08**	**71,12**
MEDIA 2022**, full**					
SAMU-XLSR	90,52	88,69	88,88	18,50	69,72
SAMU-XLSR$_{IT \oplus FR}$	**90,98**	**88,93**	**89,14**	**18,30**	**70,63**
LeBenchmark FR 3k large	90,07	88,02	87,99	19,67	69,06

TABLE 4 – Résultats pour l'approche bout-en-bout sur les différents corpus MEDIA.

Le Tableau 4 présente les résultats sur les tâches de classification d'intention et de détection de concepts avec cette approche de bout-en-bout. Pour la tâche de classification en intentions, l'approche de bout-en-bout obtient globalement de meilleurs résultats sur MEDIA 2022 que l'approche en cascade. Pour la tâche de détection des concepts, la tendance semble s'inverser. Cependant, les écarts entre les résultats du Tableau 4 et du Tableau 3 pourraient ne pas être suffisamment importants pour déterminer si une approche est favorable à l'autre. Néanmoins, nous obtenons de meilleurs scores de SFA avec notre approche de bout-en-bout pour toutes les versions du corpus MEDIA.

5 Conclusion

Dans cet article, nous avons présenté une version de l'ensemble de données de référence MEDIA enrichi avec des annotations de l'intention. Nous espérons ainsi ouvrir plus largement l'utilisation de ce corpus à la communauté internationale. Nous avons présenté également les premiers résultats expérimentaux sur cet ensemble de données enrichi en utilisant des modèles joints pour la classification des intentions et la détection de concepts.

Nous avons présenté différents systèmes réalisant conjointement la classification des intentions et la détection des concepts, que ce soit sur les transcriptions manuelles, les transcriptions automatiques (cascade) ou les signaux de parole (bout-en-bout). Les résultats expérimentaux sur les transcriptions manuelles et automatiques n'ont pas pu atteindre les résultats antérieurs de l'état-de-l'art pour la tâche détection de concepts, mais sont toujours compétitifs. Les modèles de bout-en-bout faisant l'optimisation jointe semblent obtenir de meilleurs résultats sur les deux tâches que les modèles en cascade.

Les annotations en intention sont librement disponibles dans un dépôt public [3] incluant le manuel d'annotation.

3. https://gitlab.lisn.upsaclay.fr/nlp/corpora/media-benchmark-intent-annotations

Références

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. & VILLANEAU J. (2006). Results of the French evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the French media dialog corpus. In *Proceedings Interspeech 2005*, p. 3457–3460. DOI : 10.21437/Interspeech.2005-312.

BOULANGER H., LAVERGNE T. & ROSSET S. (2022). Generating unlabelled data for a tri-training approach in a low resourced NER task. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, p. 30–37, Hybrid : Association for Computational Linguistics. DOI : 10.18653/v1/2022.deeplo-1.4.

BÉCHET F. & RAYMOND C. (2019). Benchmarking Benchmarks : Introducing New Automatic Indicators for Benchmarking Spoken Language Understanding Corpora. In *Proceedings Interspeech 2019*, p. 4145–4149. DOI : 10.21437/Interspeech.2019-3033.

CASTELLUCCI G., BELLOMARIA V., FAVALLI A. & ROMAGNOLI R. (2019). Multi-lingual intent detection and slot filling in a joint bert-based model. *ArXiv*, **abs/1907.02884**.

CATTAN O., GHANNAY S., SERVAN C. & ROSSET S. (2022). Benchmarking Transformers-based models on French Spoken Language Understanding tasks. In *Proceedings Interspeech 2022*, p. 1238–1242. DOI : 10.21437/Interspeech.2022-385.

CATTAN O., SERVAN C. & ROSSET S. (2021). On the Usability of Transformers-based Models for a French Question-Answering Task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 244–255, Held Online : INCOMA Ltd.

CHEN Q., ZHUO Z. & WANG W. (2019). Bert for joint intent classification and slot filling. *ArXiv*, **abs/1902.10909**.

CHEN Y.-N., HAKANNI-TÜR D., TUR G., CELIKYILMAZ A., GUO J. & DENG L. (2016). Syntax or semantics ? knowledge-guided joint semantic frame parsing. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, p. 348–355. DOI : 10.1109/SLT.2016.7846288.

COUCKE A., SAADE A., BALL A., BLUCHE T., CAULIER A., LEROY D., DOUMOURO C., GISSELBRECHT T., CALTAGIRONE F., LAVRIL T., PRIMET M. & DUREAU J. (2018). Snips Voice Platform : an embedded Spoken Language Understanding system for private-by-design voice interfaces. *ArXiv*, **abs/1805.10190**.

EVAIN S., NGUYEN H., LE H., BOITO M. Z., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N. A., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). Lebenchmark : A reproducible framework for assessing self-supervised representation learning from speech. In *Interspeech*, p. 1439–1443.

GHANNAY S., SERVAN C. & ROSSET S. (2020). Neural networks approaches focused on French spoken language understanding : application to the MEDIA evaluation task. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2722–2727, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.245.

GODBOLE S. & SARAWAGI S. (2004). Discriminative methods for multi-labeled classification. In H. DAI, R. SRIKANT & C. ZHANG, Éds., *Advances in Knowledge Discovery and Data Mining*, p. 22–30, Berlin, Heidelberg : Springer Berlin Heidelberg.

GOO C.-W., GAO G., HSU Y.-K., HUO C.-L., CHEN T.-C., HSU K.-W. & CHEN Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, p. 753–757, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2118.

GUO D., TUR G., YIH W.-T. & ZWEIG G. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, p. 554–559. DOI : 10.1109/SLT.2014.7078634.

HAKKANI-TÜR D., TUR G., CELIKYILMAZ A., CHEN Y.-N., GAO J., DENG L. & WANG Y.-Y. (2016). Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In *Proceedings Interspeech 2016*, p. 715–719. DOI : 10.21437/Interspeech.2016-402.

HAN S. C., LONG S., LI H., WELD H. & POON J. (2021). Bi-directional joint neural networks for intent classification and slot filling. In *Interspeech*.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

HERVÉ N., PELLOIN V., FAVRE B., DARY F., LAURENT A., MEIGNIER S. & BESACIER L. (2022). Using ASR-generated text for spoken language modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 17–25, virtual+Dublin : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.2.

JADERBERG M., DALIBARD V., OSINDERO S., CZARNECKI W. M., DONAHUE J., RAZAVI A., VINYALS O., GREEN T., DUNNING I., SIMONYAN K., FERNANDO C. & KAVUKCUOGLU K. (2017). Population based training of neural networks. *ArXiv*, **abs/1711.09846**.

JEONG M. & LEE G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(7), 1287–1302. DOI : 10.1109/TASL.2008.925143.

KHURANA S., LAURENT A. & GLASS J. (2022). Samu-xlsr : Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1493–1504. DOI : 10.1109/JSTSP.2022.3192714.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

LAPERRIÈRE G., NGUYEN H., GHANNAY S., JABAIAN B. & ESTÈVE Y. (2023). Semantic enrichment towards efficient speech representations. In *Interspeech*, p. 705–709.

LAPERRIÈRE G., PELLOIN V., CAUBRIÈRE A., MDHAFFAR S., CAMELIN N., GHANNAY S., JABAIAN B. & ESTÈVE Y. (2022). The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning : data updates, training and evaluation tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1595–1602, Marseille, France : European Language Resources Association.

LIU B. & LANE I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. *ArXiv*, **abs/1609.01454**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE E., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

PELLOIN V., DARY F., HERVÉ N., FAVRE B., CAMELIN N., LAURENT A. & BESACIER L. (2022). ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks. In *Proceedings Interspeech 2022*, p. 3453–3457. DOI : 10.21437/Interspeech.2022-352.

QIN L., LIU T., CHE W., KANG B., ZHAO S. & LIU T. (2021). A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE. DOI : 10.1109/icassp39728.2021.9414110.

SOROWER M. S. (2010). A literature survey on algorithms for multi-label learning.

TANG H., JI D. & ZHOU Q. (2020). End-to-end masked graph-based crf for joint slot filling and intent detection. *Neurocomputing*, **413**, 348–359. DOI : https ://doi.org/10.1016/j.neucom.2020.06.113.

TUR G. & MORI R. D. (2011). *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.

VAN ENGELEN J. E. & HOOS H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, **109**(2), 373–440. DOI : 10.1007/s10994-019-05855-6.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG C., HUANG Z. & HU M. (2020). Sasgbc : Improving sequence labeling performance for joint learning of slot filling and intent detection. In *Proceedings of 2020 6th International Conference on Computing and Data Engineering*, ICCDE '20, p. 29–33, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3379247.3379266.

WELD H., HUANG X., LONG S., POON J. & HAN S. C. (2022). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, **55**(8). DOI : 10.1145/3547138.

XU P. & SARIKAYA R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 78–83. DOI : 10.1109/ASRU.2013.6707709.

XU W., HAIDER B. & MANSOUR S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5052–5063, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.410.

ZHOU Z.-H. & LI M. (2005). Tri-training : exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, **17**(11), 1529–1541. DOI : 10.1109/TKDE.2005.186.

ZHUANG L., WAYNE L., YA S. & JUN Z. (2021). A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, p. 1218–1227, Huhhot, China : Chinese Information Processing Society of China.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Perception des frontières prosodiques intonatives du français par des natifs : Études comportementale et électroencéphalographique

Lei Xi[1] Rachid Ridouane[1] Frédéric Isel[1, 2]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)
(2) Modèles, Dynamiques, Corpus (CNRS & Université Paris Nanterre)
lei.xi@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr, frederic.isel@parisnanterre.fr,

RESUME

Deux expériences sur la désambiguïsation syntaxique par prosodie sont exposées dans cette étude. Dans l'expérience perceptive, nous avons demandé à 20 francophones natifs de compléter des phrases localement ambiguës afin de déterminer leur capacité à assigner correctement les mots cibles à leurs fonctions syntaxiques sur la base des indices prosodiques disponibles. Dans l'expérience électroencéphalographique (EEG), le signal continu de 20 francophones natifs a été enregistré pendant qu'ils écoutaient les mêmes phrases ambiguës. Les résultats perceptifs ont montré que les participants, bien que natifs du français, ont eu des difficultés à établir la fonction syntaxique du mot cible par l'indice prosodique. En revanche, les données neurocognitives suggèrent que les frontières intonatives ont été analysées comme l'atteste la *Closure Positive Shift* (CPS), présentant un maximum autour de 400 à 500 ms après l'onset de la dernière syllabe qui précède la frontière prosodique. Nos données soulignent l'importance du contexte prosodique complet et informatif en perception de la parole.

ABSTRACT

Perception of French prosodic boundary by native speakers: behavioral and electroencephalographic studies

Two experiments exploring the role of prosody in syntactic disambiguation are presented in this study. In a perceptual experiment, we asked 20 French native speakers to complete locally ambiguous sentences. The aim was to determine their ability to correctly assign target words to their syntactic functions based on available prosodic cues. In an electroencephalography (EEG) experiment, the EEG signal of 20 French natives was recorded while they listened to the same ambiguous sentences. The results showed that, behaviorally, listeners, even if they are native speakers of French, had difficulty establishing the correct syntactic relationship between the target word and the preceding verb. In contrast, at the neurocognitive level, prosodic boundary processing was associated with a component thought to be related to intonation boundary processing, i.e., a Closure Positive Shift (CPS), peaking around 400-500 ms after the onset of the last syllable preceding the intonational phrase boundary. We discussed our results in two frameworks of sentence processing.

MOTS-CLES : Perception de la frontière prosodique, EEG, ERP, CPS, français
KEYWORDS : Perception of prosodic boundary, EEG, ERP, CPS, French

Actes des 35èmes Journées d'Études sur la Parole, pages 481–491.

1 Introduction

L'encodage et le décodage de la frontière prosodique jouent un rôle fondamental dans la production et dans la perception de la parole. Le rôle de la frontière prosodique est d'autant plus crucial que certaines structures syntaxiques ambiguës ne peuvent se distinguer que par l'intonation, utilisée très tôt dans le traitement par l'auditeur pour prédire la structure syntaxique, résoudre l'ambiguïté et ainsi décoder l'information émise par le locuteur (Beach, 1991 ; Price et al., 1991 ; Marslen-Wilson et al., 1992 ; Nagel et al., 1996 ; Speer et al., 1996 ; Kjelgaard & Speer, 1999). Le décodage des indices fournis par la prosodie permet ainsi à l'auditeur d'établir les bonnes relations entre les unités syntaxiques et de désambiguïser le sens des phrases en fonction de l'intention du locuteur (Lehiste, 1973 ; Schafer et al., 2000 ; Carlson et al., 2009a, 2009b).

La plupart des études qui ont examiné comment la prosodie aide à désambiguïser des phrases syntaxiquement ambiguës se sont basées sur des tâches de production et/ou de perception. Pour le français, il y a notamment les travaux de Millotte et al. (2007, 2008) qui ont examiné la désambiguïsation au niveau du syntagme phonologique, avec les paires de phrases comme « *Le petit chien # mord la laisse qui le retient.* » vs « *Le petit chien mort # sera enterré demain.* ». Leurs études de production et de perception ont montré que les francophones natifs ont spontanément produit les indices prosodiques (montée de F0 et allongement final, Di Cristo, 1998 ; Jun & Fougeron, 2002) qui différencient les paires de phrases ambiguës, sans même se rendre compte de l'ambiguïté. D'autre part, les auteurs ont également montré que les francophones natifs ont recouru à ces indices prosodiques pour prédire la structure syntaxique et résoudre l'ambiguïté syntaxique dans une tâche de perception.

La perception des frontières prosodiques peut aussi être étudiée en neuroimagerie, notamment à partir de l'examen électroencéphalographique (EEG) des potentiels évoqués. Un de ces potentiels évoqués, la *Closure Positive Shift* (CPS), découvert en allemand par Steinhauer et al. (1999), est supposé marquer le traitement de frontières d'intonation majeure (et non de pauses). Cette composante de polarité positive présente une latence du pic environ 500 ms après la frontière prosodique et est distribuée bilatéralement sur les électrodes centro-pariétales. Après avoir été montrée en allemand (voir aussi Isel et al., 2005), la CPS a été répliquée dans différentes langues : en chinois (Li et al., 2008 ; Li & Yang, 2010), en anglais (Peter et al., 2014), en néerlandais (Bögels et al., 2011), en coréen (Hwang & Steinhauer, 2011) et en portugais (Batista et al., 2023), entres autres. Ainsi, Bögels et al. (2011) concluent dans leur revue de questions sur la CPS que cette déflection positive constitue un marqueur fiable de frontière prosodique, indépendant de la modalité sensorielle. En effet, la CPS a été trouvée en lecture silencieuse (Steinhauer & Friederici, 2001 ; Drury et al., 2016), mais aussi lors du traitement de parole délexicalisée ou synthétisée (Pannekamp et al., 2005 ; Honbolygó et al., 2016) ou encore de phrasé musical (Knösche et al., 2005 ; Neuhaus et al., 2006 ; Glushko et al., 2016). Des études ultérieures (Kerkhofs et al., 2008 ; Glushko et al., 2016) ont apporté de nouvelles précisions : la CPS peut être déclenchée par des frontières prosodiques de tailles différentes, mais son amplitude peut varier en fonction de la saillance perceptive. Prises dans leur ensemble, ces études multimodales suggèrent que lorsqu'un auditeur ou un lecteur s'appuie sur les informations prosodiques pour segmenter et structurer un signal continu (auditif ou visuel) en entités discrètes de taille variable, une CPS est observée. Toutefois, à ce jour, ce résultat neurophysiologique est difficilement généralisable à toutes les langues, puisque certaines, comme le français, ont été peu étudiées. À notre connaissance, il n'existe qu'une seule étude EEG (Gilbert et al., 2023), portant sur le français canadien, qui a répliqué une CPS en réponse au traitement de frontières prosodiques par des auditeurs francophones. Afin d'apporter des données complémentaires, cette étude a été conçue dans le but d'examiner le traitement de frontières prosodiques situées dans deux positions différentes (clôture précoce (CP) versus clôture tardive (CT)) chez des natifs du français métropolitain. Pour ce faire, nous avons d'abord analysé nos

données en production et en perception, avant de procéder à l'analyse par EEG. L'objectif est double : 1) d'une part, au niveau comportemental, déterminer si les francophones natifs utilisent en temps réel les indices prosodiques pour résoudre des cas d'ambiguïté syntaxique en français ; 2) d'autre part, au niveau neurocognitif, vérifier que le traitement prosodique de clôtures précoces ou tardives dans des phrases localement ambiguës du français module la CPS. La contribution majeure de cette étude est d'examiner si la CPS est déclenchée lors de la perception de la frontière prosodique du français chez des locuteurs natifs en combinant des mesures comportementales perceptives et des mesures neurophysiologiques.

2 Partie expérimentale

2.1 Stimuli

Inspirés de Pauker (2013), les stimuli utilisés dans les deux expériences sont construits à partir de 50 paires de phrases localement ambiguës (cf. Ces phrases sont disponibles sur la plateforme OSF). Chaque paire de phrases a été construite de sorte à avoir les frontières prosodiques situées dans deux positions différentes, l'une précédant le syntagme nominal, compatible avec une clôture précoce (CP) et l'autre suivant le syntagme nominal, compatible avec une clôture tardive (CT). Un exemple de ces paires de phrases est présenté dans la TABLE 1. Le même syntagme nominal dans chaque paire (ici 'le rat') a deux fonctions syntaxiques différentes, qui sont *sujet* dans la condition CP et *complément d'objet direct* dans la condition CT. En plus de ces 100 phrases (50 paires de phrases), 50 autres phrases, ne présentant aucune ambiguïté, ont été ajoutées pour servir de distracteurs (par exemple, « Chaque fois que j'allais à la piscine, il y avait du monde. »).

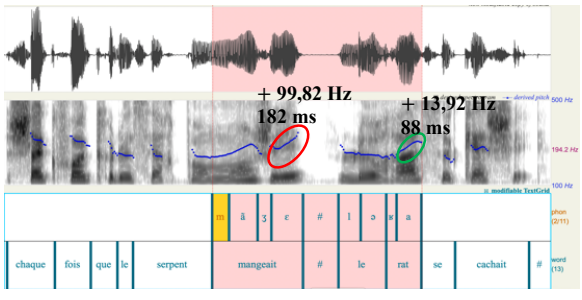
Condition prosodique	Exemple
Clôture précoce (CP) (*Early Closure*)	Chaque fois que le serpent mangeait, le rat se cachait.
Clôture tardive (CT) (*Late Closure*)	Chaque fois que le serpent mangeait le rat, le lapin se cachait.

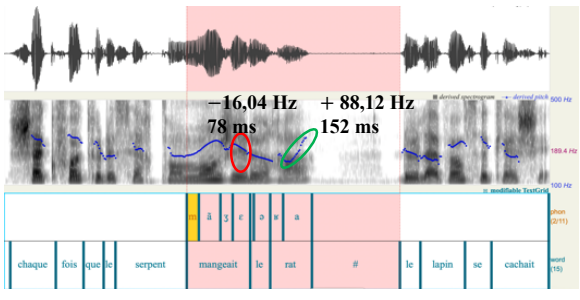
TABLE 1 : Exemple d'une paire de phrases utilisée dans cette étude

Les 150 phrases ont été lues dans un ordre aléatoire par une locutrice native du français dans la chambre sourde du Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle). Les phrases enregistrées ont été digitalisées à 16,000 Hz et à 16 Bits. Afin de mieux examiner les stimuli au niveau de la désambiguïsation syntaxique, nous avons procédé à deux analyses : acoustique et perceptive.

2.1.1 Caractéristiques acoustiques

Des analyses acoustiques et statistiques ont été effectuées pour déterminer les différences de F0 et de durée des voyelles finales[1] du syntagme verbal (SV) (en rouge dans la TABLE 1) et du syntagme nominal (SN) (en vert) dans les deux conditions CP et CT. Les analyses acoustiques et statistiques ont montré que la locutrice enregistrée avait bien réalisé les indices acoustiques attendus pour marquer la frontière du syntagme intonatif : la montée de F0, l'allongement final et la pause.





[1] Les T tests ont révélé un effet significatif de la Condition (2 modalités : CP et CT) sur chacune des variables dépendantes mesurées : la F0 et la durée, tant pour les SN (F0 : $t(49)=-6,72$, $p<.001$; durée : $t(49)=-10,89$, $p<.001$) que pour les SV (F0 : $t(49)=14,43$, $p<.001$; durée : $t(49)=26,3$, $p<.001$). Par ailleurs, une pause est systématiquement marquée à la frontière du syntagme intonatif de chaque phrase (FIGURE 1).

FIGURE 1 : Signaux acoustiques et contours mélodiques d'une paire de phrase ambiguës (CP : gauche vs CT : droite), avec les valeurs moyennes des différences de F0 et de durée de la même voyelle dans SN (entourée en rouge) et dans SN (entourée en vert)

2.1.2 Caractéristiques perceptives

Nous avons ensuite mené une étude perceptive afin de savoir si les auditeurs francophones natifs seront capables, au niveau comportemental, d'utiliser les indices acoustiques présents dans les stimuli pour désambiguïser les relations syntaxiques.

2.1.2.1 Participants et procédure

Vingt locuteurs francophones natifs (quatre hommes, seize femmes ; âge moyen : 23,1 ans ; écart-type = 4,4 ans) ont été recrutés pour prendre part à cette expérience. Tous les participants sont des locuteurs natifs du français et ne souffrent d'aucun trouble langagier ou auditif. Parmi les 150 phrases lues, 45 ont été choisies pour l'expérience perceptive (20 en condition CP, 20 en condition CT et 5 distracteurs sans ambiguïté). Nous avons tronqué ces phrases pour créer des stimuli identiques se distinguant uniquement sur le plan prosodique (e.g. « *Chaque fois que le serpent mangeait, le rat ~~se cachait.~~* » vs « *Chaque fois que le serpent mangeait le rat, ~~le lapin se cachait.~~* »). Les 40 phrases ambiguës ainsi obtenues ont été réparties dans deux blocs pour que deux phrases issues de la même paire n'apparaissent jamais dans le même bloc. À cela s'ajoutent dans chaque bloc les 5 mêmes phrases distractrices. Chaque participant a reçu la consigne suivante : « *Cette expérience consiste en une tâche de complétion de phrases, où vous devez écrire ce qui vous vient à l'esprit après chaque stimulus entendu, sous condition que la phrase complétée soit grammaticalement correcte.* ».

2.1.2.2 Résultats

Nous avons analysé un total de 500 phrases, complétées par les 20 participants. Chaque participant a évalué 20 conditions prosodiques ambiguës, comprenant 10 conditions CP et 10 conditions CT, en plus des 5 distracteurs, pour déterminer si les syntagmes nominaux ont été traités comme sujets ou comme complément d'objet direct. Les résultats obtenus sont présentés dans la FIGURE 2. L'analyse descriptive de ces résultats montre que les participants ont identifié correctement les syntagmes nominaux dans la CT, mais pas dans la CP. Autrement dit, le syntagme nominal est traité comme complément d'objet direct dans la CT alors qu'il est traité tantôt comme sujet et tantôt comme complément d'objet dans la CP.

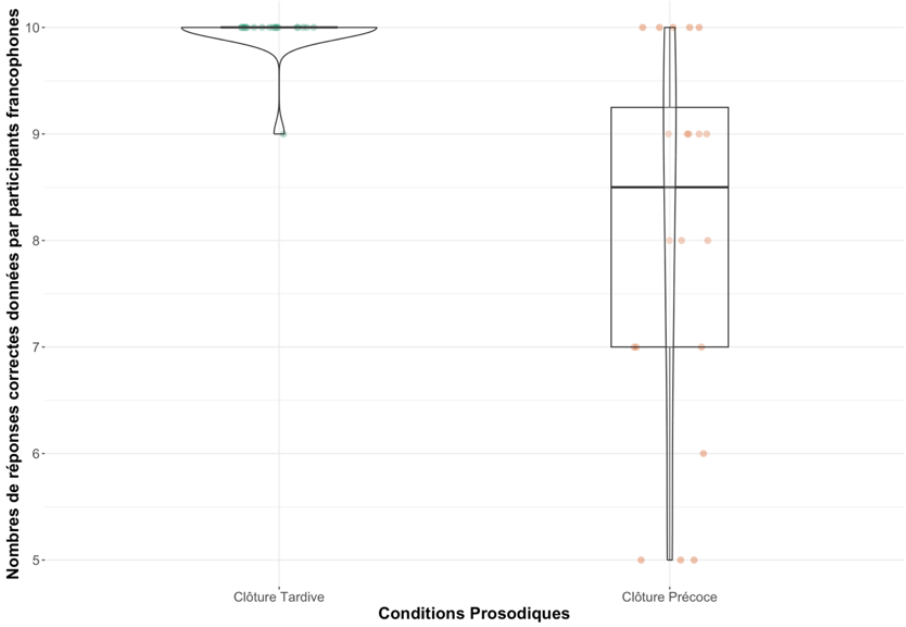


FIGURE 2 : Nombres de réponses correctes dans les conditions CP et CT pour les participants francophones natifs

Nous avons utilisé une analyse de covariance (ANCOVA) pour comparer le nombre de réponses correctes entre les conditions CP et CT, en prenant compte le bloc (1 ou 2) comme facteur covariable. Nous avons constaté un effet significatif pour la « condition prosodique » ($F(1, 37)=36,1$, $p<.001$), mais aucun effet significatif pour le « bloc » ($F(1, 37)=1,6$, $p=0,32$).

2.2 Expérience électroencéphalographique (EEG)

2.2.1 Stimuli et participants

Nous avons réalisé une deuxième expérience en utilisant la technique EEG pour examiner les processus neurocognitifs impliqués dans le traitement de frontières prosodiques en français. Les stimuli étaient constitués des 150 phrases non tronquées présentées dans la section 2.1. Pour chacune des 100 phrases localement ambiguës, deux marqueurs d'événement (*triggers*) ont été placés à l'onset de la dernière syllabe du syntagme verbal et du syntagme nominal. Cela nous a permis de comparer les réponses électriques du cerveau (potentiels évoqués) en fonction de la présence d'une frontière majeure, mineure ou en l'absence de la frontière prosodique.

Nous avons recruté 20 locuteurs ayant le français comme seule langue maternelle (6H, 14F ; âge moyen : 22,2 ans, ET=3,3 ans) pour participer à cette expérience. Tous les participants sont droitiers (Oldfield, 1971), et ont une vision normale ou corrigée. Aucun d'entre eux ne présente de signe de trouble langagier, auditif, neurologique ou psychiatrique. Par ailleurs, aucun des 20 participants n'a pris part à l'expérience perceptive, garantissant qu'ils n'ont jamais écouté les stimuli utilisés dans l'expérience EEG.

2.2.2 Procédure

Le protocole expérimental a été examiné et approuvé par le Comité d'Éthique de Recherche de la Sorbonne Nouvelle (Avis CER-USN-01-2023) et par le Service de Protection des Données du CNRS (2-22087). L'expérience s'est déroulée au Laboratoire MoDyCo (CNRS & Université Paris Nanterre). Les 150 stimuli ont été aléatoirement répartis en 5 blocs, avec 30 stimuli dans chaque bloc. Cette répartition a permis aux participants de faire une courte pause entre chaque bloc. Les participants ont été testés individuellement dans une pièce isolée de plans électrique et acoustique. Les participants étaient confortablement installés sur un fauteuil, face à un écran d'ordinateur (à 80 cm de leurs yeux) et un haut-parleur. L'expérimentateur a donné la consigne suivante : « *Vous allez écouter 150 phrases en français. Après chaque phrase, un mot s'affiche sur l'écran. Si ce mot apparaît dans la phrase que vous venez d'entendre, vous appuyez, avec la main droite, sur la touche « J » du clavier pour le choix « oui » ; s'il n'apparaît pas dans la phrase, vous appuyez sur la touche « F » du clavier pour le choix « non »* (tâche de vérification lexicale). *Il faut rester le plus immobile possible pendant l'expérience.* ». Après l'explication des consignes, un bloc d'entraînement constitué de 9 stimuli (qui ne font pas partie des 150 stimuli) a été proposé au participant afin qu'il se familiarise avec la tâche expérimentale. L'expérience était lancée une fois que le participant avait confirmé avoir bien compris la consigne. Les stimuli ont été diffusés un par un par le logiciel *Presentation* (version 24.0), avec le mot cible projeté sur l'écran à la fin de chaque stimulus. Entre chaque bloc (tous les 30 stimuli), une pause d'un temps illimité a été insérée pour que le participant puisse se reposer.

Les signaux EEG ont été enregistrés en continu à l'aide de 64 électrodes (Ag-AgCl électrodes ; Bisomei ActiveTwo system, Amesterdam, Pays-Bas) placées sur un bonnet, avec deux électrodes supplémentaires au niveau des mastoïdes. Afin de détecter les mouvements oculaires, quatre électrodes EOG (électrooculographie) ont été utilisées, une à gauche et une à droite de chaque œil, et une au-dessus et une au-dessous de l'œil gauche. Les signaux EEG étaient numérisés à une fréquence d'échantillonnage de 512 Hz. Un filtrage en temps réel a été appliqué avec une bande passante entre 0,05 Hz et 100 Hz.

2.2.3 Prétraitement des signaux EEG

Les données ont été prétraitées hors ligne à l'aide des logiciels Matlab (R2023b) et EEGLAB (Delorme & Makeig, 2004). Chaque ensemble de données EEG continues a été re-référencées aux deux électrodes mastoïdes, ré-échantillonnées à 256 Hz et filtrées dans une bande passante de 0,5 à 30 Hz. Ensuite, les données ont été visuellement inspectées pour éliminer les segments affectés par d'importants artéfacts de mouvements de la tête et remplacer les canaux de signal de mauvaise qualité par des données d'interpolation spatiale. Les artéfacts, comprenant les mouvements oculaires et clignements des yeux, ont été supprimés à l'aide d'une Analyse en Composantes Indépendantes (ICA). Des segments de données (*epochs*) correspondant aux deux conditions prosodiques (CP et CT) ont été extraits dans ERPLAB (Lopez-Calderon & Luck, 2014) : le segment temporel a été choisi entre 200 ms avant le trigger (ligne de base) et 1500 ms après (i.e. une époque de [-200, 1500]). Tout segment présentant un voltage dépassant 100 µV dans n'importe quel canal du scalp a été supprimé. Le taux de rejet de chaque participant était compris entre 1% et 30% (moyenne de rejet : 13,22%). Les signaux restants, exempts d'artéfacts, ont été inclus dans la procédure statistique de grand moyennage.

En nous basant sur les études antérieures utilisant la CPS comme marqueur de frontières prosodiques, nous avons défini trois régions d'intérêt (ROI), avec 14 électrodes dans chaque ROI : frontale (F1, F3, F5, FP1, AF3, AF7 ; FPz, Fz ; F2, F4, F6, FP2, AF4, AF8) ; centrale (FC1, FC3, FC5, C1, C3, C5 ; FCz, Cz ; FC2, FC4, FC6, C2, C4, C6) et pariétale (CP1, CP3, CP5, P1, P3, P5 ; CPz, Pz ; CP2, CP4, CP6, P2, P4, P6). Pour déterminer si la CPS est déclenchée par la perception des frontières prosodiques, les conditions CP et CT ont été examinées séparément. Après une inspection visuelle du signal ERP, trois fenêtres temporelles (FT) ont été sélectionnées pour l'analyse entre CP vs nonCP (syllabes en rouge dans TABLE 1) : [0-400], [400-800] et [800-1200], et ensuite pour CT vs nonCT (syllabes en vert) : [0-450], [450-1000] et [1000-1400].

Les valeurs de l'amplitude moyenne des électrodes ont ensuite été extraites à l'aide du logiciel EPRLAB (Lopez-Calderon & Luck, 2014) dans les trois FTs et pour toutes les conditions prosodiques. Les amplitudes moyennes ont été analysées à l'aide d'un modèle linéaire mixte pour chaque FT. Cela implique de tester l'effet de la frontière prosodique (deux niveaux : CP vs nonCP ; CT vs nonCT) et l'effet des ROIs (trois niveaux : frontale vs centrale vs pariétal). Les participants ont été considérés comme facteur à effet aléatoire. Le modèle établi était donc le suivant : [Amplitude moyenne ~ Condition prosodique * ROI + (1 | Participant)] et puis analysé dans JASP (Version 0.18.1).

2.2.4 Résultats

2.2.4.1 La condition CP vs nonCP

Les modèles ont révélé une différence significative pour toutes les FTs : avec une amplitude significativement plus positive dans la condition CP que nonCP dans la FT 400-800 ms. En revanche, dans les deux FTs [0-400 ms] et [800-1200 ms], l'amplitude de l'onde était significativement plus négative dans la CP que nonCP (FIGURE 2). Aucune interaction entre condition prosodique et ROI n'a été observée pour les trois FTs (TABLE 2). L'absence d'interaction indique que la CPS observée dans la FT [0-400 ms] n'a pas varié à travers les trois ROIs (FIGURE 2).

Fenêtre temporelle (FT) (ms)	Effet de condition prosodique	Interaction (Condition prosodique * ROI)
0-400	b= −0,29, SE=0,03, t= −8,74, p<.001	b= −0,04, SE=0,05, t= −0,77, p=0,44
400-800	b=0,37, SE=0,05, t=6,84, p<.001	b=0,007, SE=0,08, t=0,10, p=0,92
800-1200	b= −0,61, SE=0,07, t= −9,42, p<.001	b= −0,07, SE=0,09, t= −0,76, p=0,45

TABLE 2 : résultats du modèle linéaire mixte pour la condition CP vs nonCP dans les trois FTs

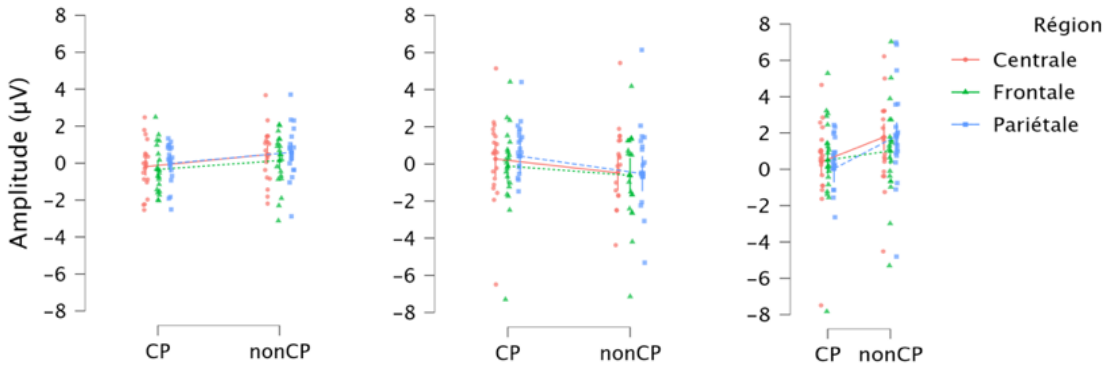


FIGURE 2 : Variation d'amplitudes moyennes de la condition CP vs nonCP pour toutes les électrodes des ROIs dans les trois FTs (0-400 ms : gauche, 400-800 ms : milieu, 800-1200 ms : droite)

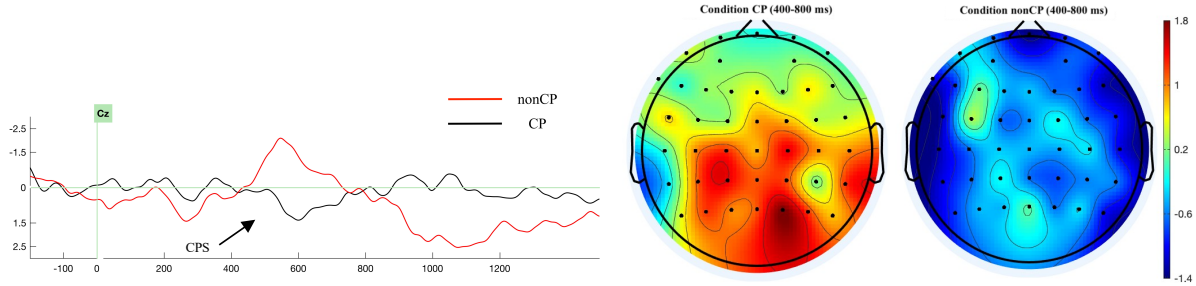


FIGURE 3 : Effet CPS sur l'électrode Cz et cartes topographiques pour la condition CP (gauche) vs nonCP (droite)

2.2.4.2 La condition CT vs nonCT

Le même schéma de résultat a été trouvé dans les trois FTs pour la condition CT vs nonCT : l'amplitude moyenne de la CPS est significativement plus positive dans la condition CT que nonCT, indépendamment de la ROI, mais uniquement dans la FT [450-1000 ms] (TABLE 3 & FIGURE 4).

Fenêtre temporelle (FT) (ms)	Effet de condition prosodique	Interaction (Condition prosodique * ROI)
0-450	b= −0,55, SE=0,04, t= −13,63, p<.001	b= −0,007, SE=0,06, t= −0,12, p=0,9
450-1000	b=0,16, SE=0,05, t=2,91, p=0,004	b=0,02, SE=0,08, t=0,22, p=0,83
1000-1400	b= −0,29, SE=0,06, t= −4,78, p<.001	b=0,09, SE=0,09, t=1,02, p=0,31

TABLE 3 : résultats du modèle linéaire mixte pour la condition CT vs nonCT dans les trois FTs

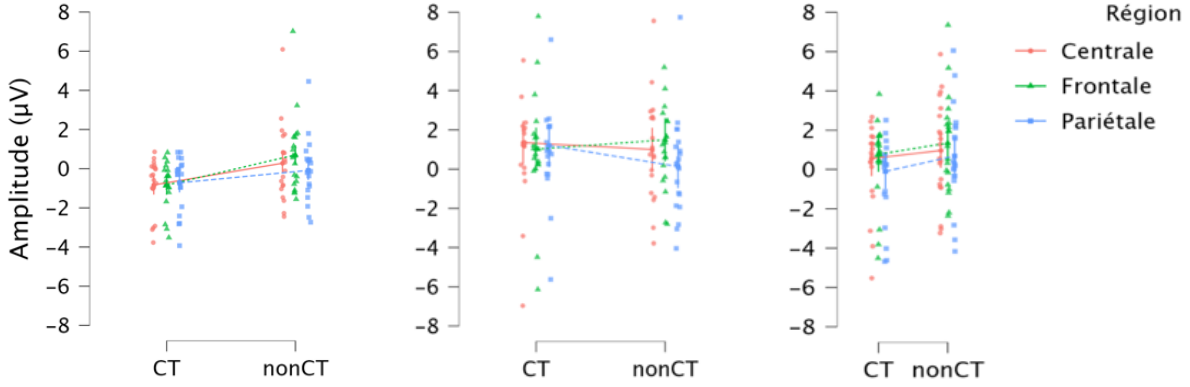


FIGURE 4 : Variation d'amplitudes moyennes de la condition CT vs nonCT pour toutes les électrodes des ROIs dans les trois FTs (0-450 ms : gauche, 450-1000 ms : milieu, 1000-1400 ms : droite)

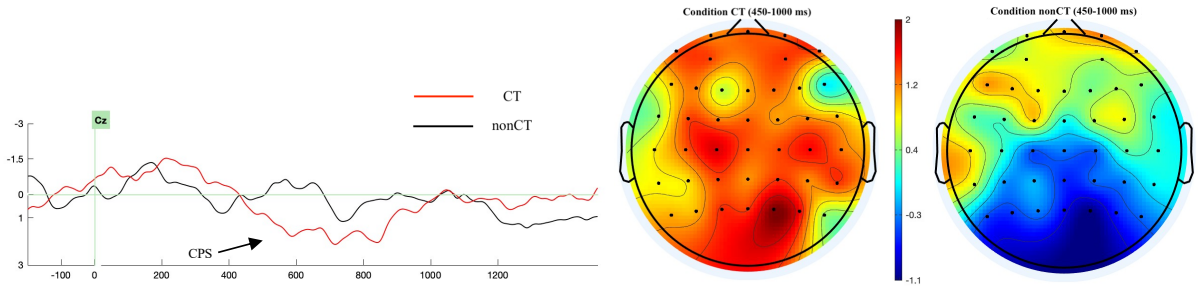


FIGURE 5 : Effet CPS sur l'électrode Cz et cartes topographiques pour la condition CT (gauche) vs nonCT (droite)

3 Discussion et conclusion

Dans ces deux expériences, nous avons étudié les performances comportementales et neurocognitives d'auditeurs francophones natifs lors de l'écoute de syntagmes nominaux inclus dans des phrases localement ambiguës. L'expérience perceptive a montré que les auditeurs avaient des difficultés à identifier correctement la fonction syntaxique des syntagmes nominaux lorsqu'ils agissent en tant que sujets (i.e. quand le syntagme nominal est situé après la frontière). Bien qu'ils aient donné plus de réponses CP pour la condition CP et plus de réponses CT pour la condition CT, le nombre de réponses correctes différait significativement entre les deux conditions. En effet, les auditeurs natifs ont préféré l'interprétation CT, ce qui a entraîné des difficultés de traitement dans les structures CP. Au contraire, les résultats de l'expérience EEG ont montré que sur la plan neurocognitif, les conditions CP et CT ont été différenciées. La CPS a été déclenchée environ 400-450 ms après la dernière syllabe accentuée avant la frontière, avec une large distribution bilatérale (FIGURES 3 & 5). Ces résultats confirment l'hypothèse selon laquelle l'analyse neurocognitive des frontières prosodiques par les locuteurs natifs du français est associée à une CPS. De plus, dans notre étude, la CPS est observée pour les frontières prosodiques situées dans deux syntagmes intonatifs différents (CP et CT). Ces données convergent avec celles d'études antérieures menées dans d'autres langues, et soutiennent l'idée que la CPS constitue un marqueur neurocognitif fiable et probablement universel de frontières prosodiques. Lorsque nous comparons les résultats des données perceptives à ceux des données EEG, une contradiction apparaît. En effet, l'expérience perceptive a révélé que le syntagme nominal est traité tantôt comme sujet tantôt comme complément d'objet dans la condition CP, suggérant ainsi une moindre saillance perceptive de la frontière prosodique pour CP par rapport à CT. Cependant, les résultats EEG montrent que le cerveau a été sensible aux frontières intonatives comme l'atteste la CPS observée quelle que soit la position de la clôture.

Cette apparente contradiction peut être attribuée à la différence entre les stimuli présentés lors des deux expériences. Dans l'étude perceptive, les phrases ont été tronquées après le syntagme nominal ambigu, privant ainsi les auditeurs d'une structure prosodique complète et contextuelle, les limitant à un traitement local (qui peut être d'autant plus influencé par la fréquence d'utilisation des verbes comme transitifs ou intransitifs). Cette observation est conforme aux prédictions de l'*Informative Boundary Hypothesis* (Clifton Jr. et al., 2002 ; Watson & Gibson, 2005 ; Carlson et al., 2009a, 2009b) et de *Late Closure Preference* (Frazier, 1979), selon lesquelles l'effet d'une frontière prosodique donnée dépend d'autres frontières apparues avant et après dans la phrase. Lorsque les informations prosodiques ne sont pas suffisamment saillantes pour établir la bonne relation syntaxique entre les syntagmes, les auditeurs ont tendance à attacher automatiquement le mot ambigu à la partie précédente de la phrase, ce qui explique pourquoi certaines conditions CP ont été interprétées comme CT. En revanche, dans l'étude EEG, aucune troncation de phrases n'a été effectuée, permettant ainsi un traitement de la structure syntaxico-prosodique globale (Lee & Garnsey, 2012). Ce contexte prosodique complet et suffisamment informatif a permis au cerveau des auditeurs d'établir et d'ajuster la hiérarchie prosodique et la relation syntaxique entre syntagmes, facilitant ainsi le traitement neurocognitif des deux conditions CP et CT.

Références

BATISTA A., CATRONAS D., FOLIA V. & SILVA S. (2023). Increased Pre-Boundary Lengthening Does Not Enhance Implicit Intonational Phrase Perception in European Portuguese: An EEG Study. *Brain Sciences*, 13, 441.

BEACH C. (1991). The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations. *Journal of memory and language*, 30, 644-663.

BOERSMA P. & WEENINK D. (2022). *Praat (version 6.2.15): doing phonetics by computer*.

BÖGELS S., SCHRIEFERS H., VONK W. & CHWILLA D. (2011). Prosodic Breaks in Sentence Processing Investigated by Event-Related Potentials. *Language and Linguistics Compass*, 5/7, 424-440.

CARLSON K., FRAZIER L. & CLIFTON JR C. (2009a). How prosody constrains comprehension: A limited effect of prosodic packaging. *Lingua*, 119(7), 1066-1082.

CARLSON K., CLIFTON JR C. & FRAZIER L. (2009b). Nonlocal effects of prosodic boundaries. *Memory & Cognition*, 37(7), 1014-1025.

CLIFTON JR C., CARLSON K. & FRAZIER L. (2002). Informative Prosodic Boundaries. *Language and Speech*, 45(2), 87-144.

DELORME A. & MAKEIG S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134, 9-21.

DI CRISTO A. (1998). Intonation in French. In D. HIRST & A. DI CRISTO, Éds., *Intonation Systems A Survey of Twenty Languages*, chapter 11, p. 195-218. Cambridge University Press.

DRURY J., BAUM S., VALERIOTE H. & STEINHAUER K. (2016). Punctuation and Implicit Prosody in Silent Reading: An ERP Study Investigating English Garden-Path Sentences. *Frontiers in Psychology*, Volume 7: 1375.

FRAZIER L. (1979). *On comprehending sentences: Syntactic parsing strategies*. Thèse de doctorat, University of Connecticut, États-Unis.

GILBERT A., LEE J., WOLPERT M. & BAUM S. (2023). Phrase parsing in a second language as indexed by the closure positive shift: The impact of language experience and acoustic cue salience. *European Journal of Neuroscience*, 58, 3838-3858.

GLUSHKO A., STEINHAUER K., DE PRIEST J. & KOELSCH S. (2016). Neurophysiological Correlates of Musical and Prosodic Phrasing: Shared Processing Mechanisms and Effects of Musical Expertise. *PLoS ONE*, 11(5): e0155300.

HONBOLYGÓ F., TÖRÖK Á., BÁNRÉTI Z., HUNYADI L. & CSÉPE V. (2016). ERP correlates of prosody and syntax interaction in case of embedded sentences. *Journal of Neurolinguistics*, 37, 22-33.

HWANG H. & STEINHAUER K. (2011). Phrase Length Matters: The Interplay between Implicit Prosody and Syntax in Korean "Garden Path" Sentences. *Journal of Cognitive Neuroscience*, 23:11, 3555-3575.

ISEL F., ALTER K. & FRIEDERICI A. (2005). Influence of Prosodic Information on the Processing of Split Particles: ERP Evidence from Spoken German. *Journal of Cognitive Neuroscience*, 17:1, 154-167.

JASP TEAM. (2023). *JASP (Version 0.18.1)*.

JUN S. & FOUGERON C. (2002). Realization of Accentual Phrase in French Intonation. *Pobus*, 14, 147-172.

KERKHOFS, R., VONK, W., SCHRIEFERS, H., CHWILLA, D. (2008). Sentence processing in the visual and auditory modality: Do comma and prosodic break have parallel functions? *Brain Research*, 1224, 102-118.

KJELGAARD M. & SPEER S. (1999). Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity. *Journal of Memory and Language*, 40, 153-194.

KNÖSCHE T., NEUHAUS C., HAUEISEN J., ALTER K., MAESS B., WITTE O. & FRIEDERICI A. (2005). Perception of Phrase Structure in Music. *Human Brain Mapping*, 24, 259-273.

LEE E. & GARNSEY S. (2012). Do contrastive accents modulate the effect of intonational phrase boundaries in parsing? *Lingua*, 122, 1763-1775.

LEHISTE I. (1973). Phonetic disambiguation of syntactic ambiguity. *The Journal of the Acoustical Society of America*, 53(1), 380.

LI W. & YANG Y. (2010). Perception of Chinese Poem and Its Electrophysiological Effects. *Neuroscience*, 168, 757-768.

LI W., WANG L., LI X. & YANG Y. (2008). Closure Positive Shifts Evoked by Different Prosodic Boundaries in Chinese Sentences. In R. WANG, E. SHEN & F. GU, Éds., *Advances in Cognitive Neurodynamics*, Chapter 88, p. 505-509. Springer.

LOPEZ-CALDERON J. & LUCK S. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, Volume 8, 213.

MARSLEN-WILSON W., TYLER L., WARREN P., GRENIER P. & LEE C. (1992). Prosodic Effects in Minimal Attachment. *The Quarterly Journal of Experimental Psychology*, 45A(1), 73-87.

MILLOTTE S., RENÉ A., WALES R. & CHRISTOPHE A. (2008). Phonological Phrase Boundaries Constrain the Online Syntactic Analysis of Spoken Sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 34 No. 4, 874-885.

MILLOTTE S., WALES R. & CHRISTOPHE A. (2007). Phrasal prosody disambiguates syntax. *Language and cognitive processes*, 22(6), 898-909.

NAGEL H., SHAPIRO L., TULLER B. & NAWY R. (1996). Prosodic Influences on the Resolution of Temporary Ambiguity During On-Line Sentence Processing. *Journal of Psycholinguistic Research*, Vol. 25 No. 2, 319-344.

NEUHAUS C., KNÖSCHE T. & FRIEDERICI A. (2006). Effects of Musical Expertise and Boundary Markers on Phrase Perception in Music. *Journal of Cognitive Neuroscience*, 18:3, 472-493.

OLDFIELD R. (1971). The Assessment and Analysis of Handedness: The Edinburgh Inventory. *Neuropsychologia*, Vol. 9, 97-113.

PANNEKAMP A., TOEPEL U., ALTER K., HAHNE A. & FRIEDERICI A. (2005). Prosody-driven Sentence Processing: An Event-related Brain Potential Study. *Journal of Cognitive Neuroscience*, 17:3, 407-421.

PAUKER E. (2013). *How multiple prosodic boundaries of varying sizes influence syntactic parsing: Behavioral and ERP evidence*. Thèse de doctorat, McGill University, Canada.

PETER V., MCARTHUR G. & CRAIN S. (2014). Using event-related potentials to measure phrase boundary perception in English. *BMC Neuroscience*, 15, 129.

PRICE P., OSTENDORF M., SHATTUCK-HUFNAGEL S. & FONG C. (1991). The use of prosody in syntactic disambiguation. *The journal of the Acoustical Society of America*, 90(6), 2956-2970.

SCHAFER A., SPEER S., WARREN P. & WHITE S. (2000). Intonational Disambiguation in Sentence Production and Comprehension. *Journal of Psycholinguistic Research*, Vol. 29 No. 2, 169-182.

SPEER S., KJELGAARD M. & DOBROTH K. (1996). The Influence of Prosodic Structure on the Resolution of Temporary Syntactic Closure Ambiguities. *Journal of Psycholinguistic Research*, Vol. 25 No.2, 249-271.

STEINHAUER K. & FRIEDERICI A. (2001). Prosodic Boundaries, Comma Rules, and Brain Responses: The Closure Positive Shift in EPRs as a Universal Marker for Prosodic Phrasing in Listeners and Readers. *Journal of Psycholinguistic Research*, Vol. 30 No. 3, 267-295.

STEINHAUER K., ALTER K. & FRIEDERICI A. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, volume 2 no 2, 191-196.

WATSON D. & GIBSON E. (2005). Intonational phrasing and constituency in language production and comprehension. *Studia linguistica*, 59(2-3), 279-300.

XI L. & RIDOUANE R. (2022). Quand la syntaxe a besoin de la prosodie : comment les indices prosodiques en français aident les apprenants sinophones à traiter l'information syntaxique – une étude perceptive. In *Actes des 34ème Journées d'Études sur la Parole – JEP 2022 « Parole, Geste, Musique : des unités à leur organisation »*, p. 173-182.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Peut-on évaluer la compréhensibilité de la parole sans référence quant aux intentions de communication du locuteur ? Une étude auprès d'apprenants germanophones de FLE

Verdiana De Fino[1, 2] Isabelle Ferrané[1] Julien Pinquier[1] Lionel Fontan[2]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) Archean LABS, Montauban, France
verdiana.defino@irit.fr, isabelle.ferrane@irit.fr, lfontan@archean.tech, julien.pinquier@irit.fr

RÉSUMÉ

En didactique des langues étrangères, la compréhensibilité des énoncés produits par les apprenants est le plus souvent évaluée de manière subjective, à l'aide d'échelles qualitatives. Très souvent, ces évaluations sont menées sans que l'évaluateur ne soit informé du contenu sémantique du message que souhaitait transmettre l'apprenant. L'évaluateur peut donc ignorer des divergences entre ce dernier et sa propre interprétation de l'énoncé – avec pour conséquence une surestimation de la compréhensibilité. Dans cette étude, nous vérifions l'existence d'un tel biais en demandant à 80 francophones natifs d'évaluer la compréhensibilité d'énoncés produits par neuf apprenants germanophones de français lors d'une tâche de traduction. L'évaluation est conduite sans référence (condition « *a priori* »), et en prenant connaissance d'une traduction de référence (condition « *a posteriori* »). Les résultats démontrent que les scores de compréhensibilité sont significativement plus élevés dans la condition *a priori* que dans la condition *a posteriori*, avec une taille d'effet moyenne.

ABSTRACT

Can speech comprehensibility be assessed without any reference as to the communicative intent of the speaker? A study in German speakers of French as a foreign language

The comprehensibility of speech utterances produced by non-native speakers of foreign languages is generally assessed using subjective ratings on qualitative scales. Often, raters are not informed about the semantic content of the messages that the speakers intended to transmit. The raters can thus overestimate comprehensibility, given that they cannot be aware of discrepancies between the communicative intents of the speakers and their own understanding of the utterances. The present study aimed at verifying the existence of this bias by asking 80 native-French speakers to assess the comprehensibility of utterances produced by nine German learners of French during a translation task. The assessment took place without providing to the rater any reference as to the communicative intent of the speaker (*a priori* assessment) and after providing a reference translation (*a posteriori* assessment). As was hypothesized, the results show that *a priori* ratings are significantly higher than *a posteriori* ratings, with a medium effect size.

MOTS-CLÉS : Compréhensibilité de la parole, évaluation, biais, L2.

KEYWORDS: Speech comprehensibility, assessment, bias, L2.

Actes des 35èmes Journées d'Études sur la Parole, pages 492–501.

1 Introduction

En didactique des langues étrangères (L2), il a pendant longtemps été considéré que l'objectif ultime de l'enseignement était, pour l'apprenant, d'atteindre le niveau de compétence d'un locuteur natif, et ce y compris au niveau phonético-phonologique – avec la production d'une parole parfaitement intelligible, sans qu'aucun accent étranger ne soit perceptible (Derwing, 2003). Cependant, il s'agit là d'un objectif que très peu d'apprenants parviennent à réaliser, en particulier lorsque le processus d'acquisition débute après l'enfance (Flege, 1988). Peu à peu, cette perspective a donc évolué et la majorité des didacticiens considèrent aujourd'hui qu'il est bien plus important pour l'apprenant de parvenir à une bonne *compréhensibilité* lors d'interactions avec les locuteurs de la communauté cible que de parvenir à s'exprimer « sans accent » et avec un niveau de performance égal à celui d'un locuteur natif (Derwing & Munro, 2009; Munro & Derwing, 2011).

Par conséquent, dans le domaine de l'acquisition/apprentissage des langues, la performance des locuteurs est souvent évaluée à travers la compréhensibilité des énoncés qu'ils produisent dans la langue cible. Pour cela, la méthode la plus utilisée (le « *gold standard* » actuel) consiste à demander à des auditeurs natifs de la L2 d'évaluer le degré de difficulté qu'ils éprouvent à comprendre lesdits énoncés (Derwing & Munro, 2005). De nombreuses études ont ainsi recours à des échelles perceptives d'évaluation de la compréhensibilité, appliquées à l'évaluation d'énoncés produits dans des tâches de production orale plus ou moins contraintes : de la lecture oralisée (Ludwig & Mora, 2017) à la parole recueillie durant des tâches de production plus spontanées comme des tâches d'argumentation (Suzuki & Kormos, 2020).

Demander à un auditeur d'estimer, sur une échelle, le degré de difficulté qu'il a éprouvé pour comprendre un message suppose que celui-ci connaisse les intentions de communication du locuteur, autrement dit, qu'il connaisse le contenu sémantique, discursif, voire pragmatique, que le locuteur souhaitait véhiculer avec son message (Fontan, 2012; Fontan *et al.*, 2013). Or, dans une grande partie des travaux en L2, les auditeurs évaluent la compréhensibilité sans aucune connaissance quant au contenu du message que le locuteur souhaitait transmettre. C'est en particulier le cas pour les études reposant sur des énoncés produits dans des tâches de production orale (semi-)spontanées, pour lesquelles il n'existe pas de « script » correspondant aux énoncés que l'apprenant devait réaliser. Cependant, cette méthode est aussi appliquée pour l'évaluation de la parole lue. Pour illustration, la table 1 recense des études relativement récentes dans lesquelles la compréhensibilité d'apprenants de L2 a été évaluée sur différentes échelles perceptives, sans qu'aucune information n'ait été donnée aux évaluateurs quant aux intentions de communication des apprenants.

Le fait que l'auditeur ne dispose d'aucune information quant au message que l'apprenant souhaitait transmettre peut représenter un biais pour l'évaluation de la compréhensibilité de la parole. En effet, les « erreurs » de prononciation, de choix lexicaux, ou de constructions morphosyntaxiques commises par les apprenants peuvent conduire à la production d'énoncés pouvant être perçus par les évaluateurs comme étant tout à fait corrects. Dans ce cas, ces énoncés peuvent recevoir des scores de compréhensibilité élevés, même s'ils ne correspondent pas aux messages que les apprenants souhaitaient transmettre. Pour prendre le seul exemple d'une erreur de prononciation, il est possible qu'un apprenant produise une phrase perçue comme « Je veux deux cafés », alors qu'il souhaitait dire « Je veux du café », mais que la réalisation du déterminant « du » (/dy/) soit perçue comme « deux » ([dø]) à cause de son système interphonologique. Dans ce cas, si l'évaluateur ne connaît pas le message que souhaitait transmettre l'apprenant, il pourra alors sous-estimer sa difficulté à comprendre l'énoncé.

Étude	Tâche de production orale	Outil d'évaluation de la compréhensibilité
Bergeron & Trofimovich (2017)	Parole spontanée (narration d'histoire imagée et entretien semi-dirigé)	Échelle continue
Crowther *et al.* (2015)	Parole spontanée (narration d'histoire imagée)	Échelle continue
Hansen Edwards *et al.* (2018)	Lecture de texte Parole spontanée (entretien semi-dirigé)	Échelle à 9 points
Kang (2010)	Parole spontanée (entretien semi-dirigé)	Échelle à 7 points
Kennedy & Trofimovich (2008)	Lecture oralisée de phrases	Échelle à 9 points
Ludwig & Mora (2017)	Lecture oralisée de phrases	Échelle à 7 points
Nagle & Huensch (2020)	Parole spontanée (entretien semi-dirigé)	Échelle continue
Saito *et al.* (2017)	Parole spontanée (narration d'histoire imagée)	Échelle continue
Saito *et al.* (2023)	Parole spontanée (description d'image)	Échelle à 9 points
Suzukida & Saito (2021)	Parole spontanée (narration d'histoire imagée)	Échelle à 9 points

TABLE 1 – Exemples d'études dans lesquelles la compréhensibilité d'apprenants de L2 a évaluée sans informer les évaluateurs du contenu sémantique des messages que les apprenants avaient pour intention de transmettre.

La présente étude a pour objet de vérifier l'existence d'un tel biais lors de l'évaluation de la compréhensibilité de la parole d'énoncés produits par des apprenants germanophones de français langue étrangère (FLE), recueillis pendant une tâche de traduction orale de phrases cibles. À cette fin, un protocole permettant d'évaluer la compréhensibilité des énoncés sans aucune connaissance du message cible, et avec la mise à disposition d'une traduction de référence en français a été mis en place. Nos hypothèses sont celles (i) d'une surestimation de la compréhensibilité des énoncés lorsqu'ils sont présentés sans aucune référence, et (ii) de meilleurs accords inter-évaluateurs après la mise à disposition des traductions de référence en comparaison des accords observés pour les évaluations réalisées *a priori*.

2 Méthode

2.1 Stimuli de parole

Neuf locuteurs germanophones ont participé à l'étude. Il s'agit de neuf apprenants de FLE (2 femmes, 7 hommes), dont les niveaux de compétences selon le Cadre Européen Commun de Référence pour

les Langues (CECRL, Conseil De l'Europe, 2001) s'échelonnaient de A1 à B1.

Les apprenants ont été enregistrés lors de la traduction orale de 40 phrases cibles présentées en allemand. Ces 40 phrases, élaborées avec l'aide de deux enseignants allemands de FLE, correspondaient à des phrases dans un registre de langue courant, et étaient de nature déclarative. Chaque phrase était destinée à susciter des erreurs lexicales, syntaxiques ou morphosyntaxiques fréquentes chez les apprenants germanophones de FLE. Ainsi, une erreur courante au niveau lexical est d'utiliser le verbe « recevoir » au lieu d'« obtenir » (*p. ex.* « J'ai reçu mon diplôme » au lieu de « J'ai obtenu mon diplôme »), au niveau syntaxique d'utiliser un syntagme nominal en lieu et place d'un syntagme prépositionnel comme complément au verbe « répondre » (*p. ex.* « Nous devons répondre cette question » au lieu de « Nous devons répondre à cette question ») et au niveau morphosyntaxique d'utiliser des formes comme « vieux » ou « beau » devant un nom masculin commençant par une voyelle (*p. ex.* « C'est un vieux homme » au lieu de « C'est un vieil homme »).

Les énoncés produits par les apprenants ont été enregistrés dans une salle calme de l'Université Ostfalia à Wolfenbüttel (Allemagne), à l'aide d'un microphone casque Jabra (Copenhague, Danemark) Evolve 20 HSC016. Le niveau sonore des 360 fichiers audio produits a été normalisé avant les évaluations.

2.2 Évaluation de la compréhensibilité de la parole

Quatre-vingts évaluateurs (26 femmes, 54 hommes) ont été recrutés pour évaluer le degré de compréhensibilité des énoncés produits par les apprenants. Les critères d'inclusion dans l'étude étaient les suivants : francophone natif, sans trouble auditif connu, et âgé entre 18 et 40 ans (pour limiter la probabilité de troubles auditifs liés à l'âge dont les évaluateurs n'auraient peut-être pas conscience, Cruickshanks *et al.*, 1998).

Les évaluations ont pris place de manière individuelle dans des pièces calmes. Les stimuli étaient diffusés dans un casque Jabra (Copenhague, Danemark) Evolve 20 HSC016, connecté à la carte son d'un ordinateur portable MacBook Pro 13 pouces. Sur l'écran d'ordinateur, une interface graphique développée en langage de programmation Python avec la librairie Streamlit 3 [1] était présentée (voir Figure 1).

Avant de démarrer la procédure, chaque évaluateur était familiarisé avec le concept de compréhensibilité de la parole. En accord avec la définition proposée par V. Woisard et collègues (« Capacité de l'auditeur à interpréter le sens du message oral produit par un locuteur, sans tenir compte de la précision ou de la justesse phonétique ou lexicale » ; Woisard *et al.*, 2013), nous avons en particulier insisté sur le fait que l'évaluation portait sur la capacité à *interpréter le sens* des énoncés, et qu'il ne fallait pas, pour cela, tenir compte d'« erreurs » phonétiques ou linguistiques (par exemple, une prononciation déficiente) qui n'entraveraient pas cet accès au sens. Chaque évaluateur était ensuite familiarisé avec l'échelle à 5 points utilisée pour l'évaluation de la compréhensibilité, échelle qui s'étendait de 1 (« compréhensibilité nulle ») jusqu'à 5 (« compréhensibilité totale »).

Scores de compréhensibilité *a priori*

Pour chaque énoncé, chaque participant devait d'abord évaluer la compréhensibilité de la parole sans aucune indication quant au contenu sémantique cible du message produit par l'apprenant (voir cadre 1

1. https://streamlit.io/

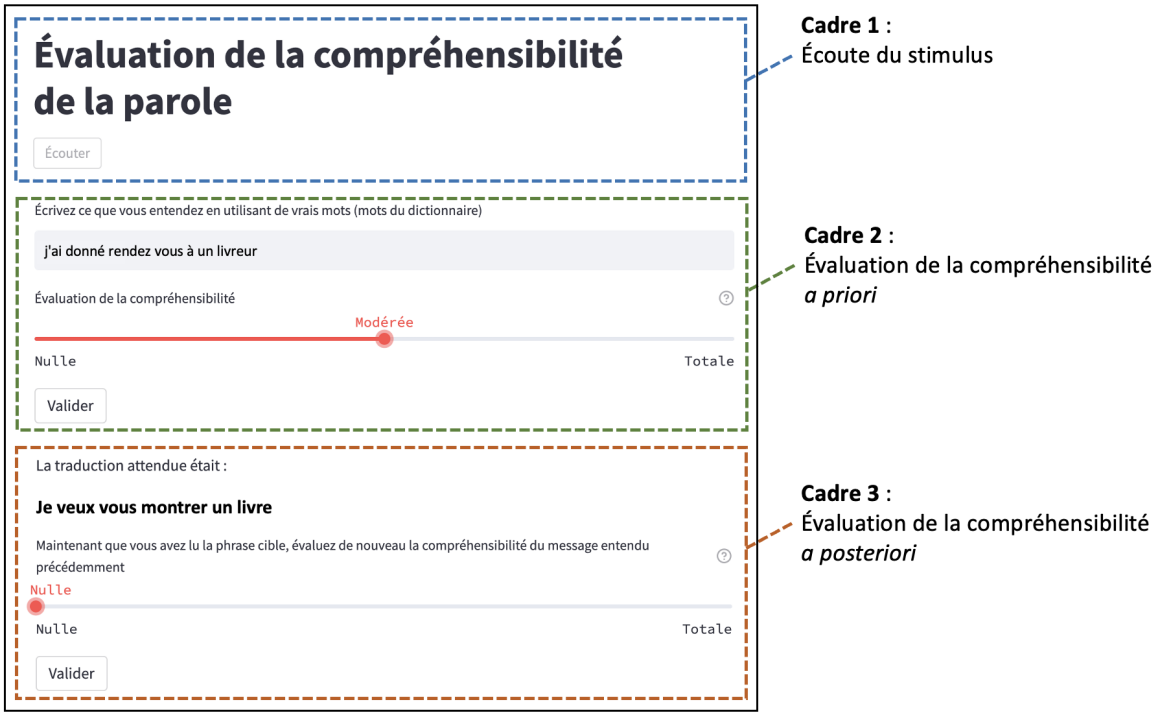


FIGURE 1 – Interface logicielle utilisée pour l'évaluation de la compréhensibilité de chaque enregistrement. L'exemple correspond à de vraies évaluations obtenues pour une traduction orale de la phrase « Ich möchte Ihnen ein Buch zeigen » (« Je veux vous montrer un livre »), perçue par l'évaluateur comme « j'ai donné rendez-vous à un livreur ».

de l'interface). Après l'écoute de la production orale, l'évaluateur avait pour consigne de retranscrire la phrase entendue [2], et d'attribuer un score de compréhensibilité sur l'échelle à 5 points (voir cadre 2 de l'interface). Nous nommons ces évaluations « Scores de compréhensibilité *a priori* ».

Scores de compréhensibilité *a posteriori*

Une traduction de référence de l'énoncé en français, établie par un consensus entre les enseignants qui ont défini le jeu de phrases à traduire, était ensuite présentée à l'évaluateur. Après la prise de connaissance de cette traduction de référence, l'évaluateur devait à nouveau attribuer un score de compréhensibilité sur l'échelle à 5 points (voir cadre 3 de l'interface). Nous nommons ces secondes évaluations « Scores de compréhensibilité *a posteriori* ».

Les évaluateurs ne pouvaient écouter qu'une seule fois les enregistrements, avant l'évaluation *a priori*. Aucune écoute supplémentaire entre les évaluations *a priori* et *a posteriori* n'était possible. De plus, afin d'éviter les effets d'entraînement (habituation au locuteur), chaque évaluateur a évalué neuf productions orales (une par apprenant). Cela a donné lieu à 18 scores de compréhensibilité : neuf *a priori*, neuf *a posteriori*. Enfin, l'attribution des fichiers audios aux différents évaluateurs a été réalisée de manière à ce que chaque couple d'évaluateurs ait à évaluer un même ensemble de neuf fichiers audio, permettant ainsi le calcul d'accords inter-évaluateurs.

Au total, 1440 scores de compréhensibilité ont ainsi été recueillis, correspondant aux 2 conditions

2. Donnée que nous n'exploitons pas pour la présente étude.

d'évaluation (*a priori vs. a posteriori*) × 9 fichiers audio × 80 évaluateurs.

3 Résultats

3.1 Scores de compréhensibilité *a priori* et *a posteriori*

La Figure 2 présente les distributions des scores de compréhensibilité *a priori* (à gauche) et *a posteriori* (à droite).

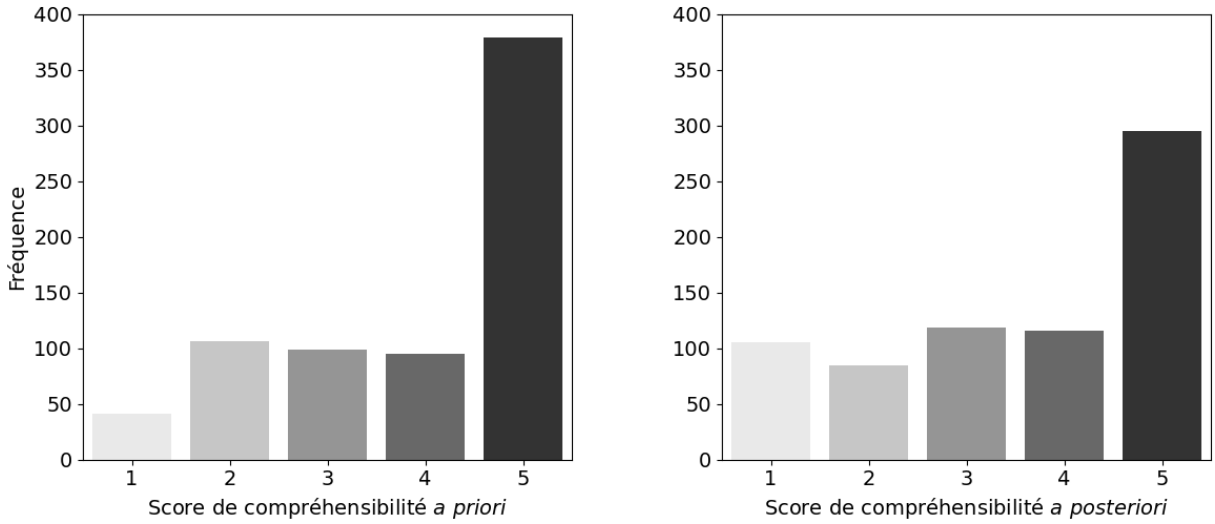


FIGURE 2 – Distribution des scores de compréhensibilité attribués *a priori* et *a posteriori*.

Comme nous pouvons le remarquer, les scores de compréhensibilité sont globalement plus élevés dans la condition *a priori* que dans la condition *a posteriori*, avec, en particulier, moins d'énoncés ayant été associés avec la valeur la plus faible de l'échelle (1), et davantage d'énoncés ayant obtenu la valeur maximale (5). Afin de vérifier la significativité de cette différence de distribution, nous avons utilisé le test non-paramétrique de Wilcoxon. Les résultats ont mis au jour une différence significative entre les deux conditions, les scores de compréhensibilité relevés *a priori* étant plus élevés que ceux observés dans la condition *a posteriori* ($Z = -10{,}4$; $p < 0{,}001$, test unilatéral). Afin d'estimer la taille de l'effet, le r de Cohen a été calculé ; sa valeur est de 0,39, ce qui dénote une taille d'effet moyenne.

3.2 Accords inter-évaluateurs lors des évaluations *a priori* et *a posteriori*

Les accords inter-évaluateurs ont été calculés, pour chaque paire d'évaluateurs, sous la forme de coefficients de corrélation de Spearman. Sur l'ensemble des 40 paires d'évaluateurs, ces coefficients s'échelonnent de 0,16 à 0,98 dans la condition *a priori* (moyenne : 0,73) et de 0,25 à 1 dans la condition *a posteriori* (moyenne : 0,78). La Figure 3 présente les distributions des accords inter-évaluateurs dans ces deux conditions.

Nous pouvons observer que la valeur médiane des accords inter-évaluateurs est plus élevée dans la condition *a posteriori* (accord médian : 0,83) que dans la condition *a priori* (accord médian : 0,71).

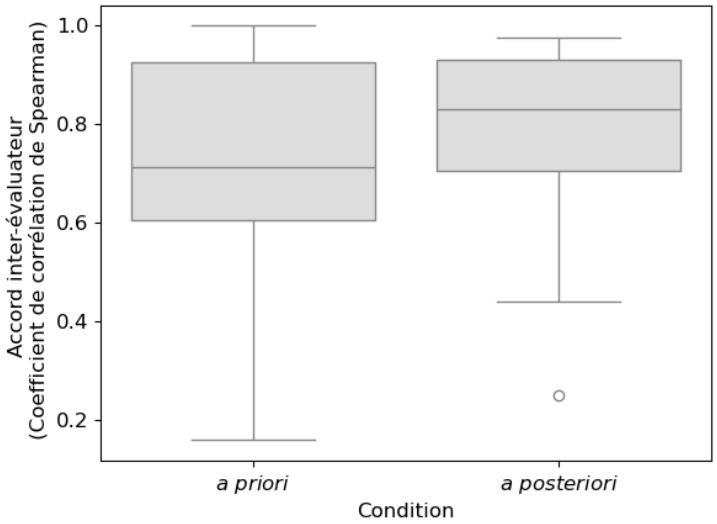


FIGURE 3 – Distribution des accords inter-évaluateurs pour les conditions *a priori* et *a posteriori*. Les barres d'erreur renvoient aux valeurs minimales et maximales, et les limites basses et hautes des boîtes aux 25ème et 75ème percentiles, respectivement. Les traits horizontaux à l'intérieur des boîtes représentent les valeurs médianes. Le cercle vide correspond à une donnée aberrante.

De même, l'étendue interquartile est plus faible dans la condition *a posteriori*, avec la moitié des valeurs comprises entre 0,70 et 0,93 (contre 0,6 et 0,92 pour la condition *a priori*). Afin de vérifier la significativité des différences d'accord observées entre les deux conditions, nous avons d'abord réalisé des tests de normalité de distribution à l'aide du test de Shapiro-Wilk. Les résultats ayant montré que les coefficients de corrélation ne sont pas normalement distribués ($p \leq 0{,}019$ dans les deux conditions), nous avons ensuite réalisé un test de Wilcoxon. Les résultats n'ont pas mis au jour de différence significative entre les accords inter-évaluateurs observés *a priori* et ceux relevés *a posteriori* ($Z = -1{,}3$; $p = 0{,}09$, test unilatéral).

4 Conclusion et discussion

Cette étude avait pour premier objectif de vérifier si l'absence de référence quant aux intentions de communication d'un locuteur pouvait constituer un biais lors de l'évaluation de la compréhensibilité de la parole d'énoncés correspondant à des traductions orales de phrases cibles. Plus précisément, nous supposions que, sans une telle référence, les évaluateurs auraient tendance à surestimer leur propre compréhension, dans la mesure où ils ne peuvent avoir conscience de divergences d'ordre sémantique entre le message perçu et le message que souhaitait transmettre le locuteur. Cette hypothèse a été confirmée par la mise au jour d'une différence significative entre les évaluations de compréhensibilité réalisées *a priori* et celles réalisées *a posteriori*, c'est-à-dire après la prise de connaissance d'une traduction de référence – les premières étant supérieures aux secondes. La taille (moyenne) de l'effet observé n'est pas négligeable, ce qui souligne d'autant plus l'enjeu que peut représenter un tel biais dans les études reposant sur des mesures de compréhensibilité obtenues par des méthodes d'évaluation *a priori*.

Il est néanmoins utile de rappeler que l'effet observé correspond à une tendance générale, observée à travers les différents énoncés produits par neuf apprenants germanophones, et dont la compréhensibilité a été évaluée par 80 auditeurs différents. Ceux-ci ont parfois également sous-évalué leur compréhensibilité lors de la condition *a priori*. Cette sous-évaluation représente toutefois une minorité de cas (18,8%) en comparaison des phénomènes de surévaluation (81,2%).

Notre seconde hypothèse était que les accords inter-évaluateurs seraient plus élevés dans la condition d'évaluation *a posteriori* que dans celle *a priori*, dans la mesure où disposer d'une référence commune devrait contribuer à l'harmonisation des évaluations. Malgré le calcul d'un nombre important d'accords (40), cette hypothèse n'a pas été confirmée par les statistiques inférentielles. Il est possible que le faible nombre de productions pour lesquels ces accords ont été calculés (9) n'ait pas permis d'obtenir des coefficients de corrélation suffisamment fiables pour mettre au jour une telle différence. Nous prévoyons, à terme, de mener une étude similaire incluant l'évaluation de 40 énoncés produits par 40 locuteurs japonophones de FLE (De Fino, 2024) afin de dépasser cette limite.

Remerciements

Ces travaux ont été financés par l'Agence Nationale de la Recherche dans le cadre du laboratoire commun ALAIA (ANR-18-LVC3-001), et par l'Association Nationale de la Recherche et de la Technologie *via* la thèse CIFRE de V. De Fino. Nous tenons à remercier le Pr. Gerndt pour nous avoir donné l'opportunité d'enregistrer des apprenants de FLE à l'université Ostfalia. Nous remercions également le Dr. Volmer et Mme Bizien pour l'apport de leur expertise lors de la création des énoncés à traduire, ainsi que les apprenants et évaluateurs qui ont participé à cette étude.

Références

BERGERON A. & TROFIMOVICH P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language french. *Foreign Language Annals*, **50**(3), 547–566. DOI : 10.1111/flan.12285.

CONSEIL DE L'EUROPE (2001). *Cadre Européen Commun de Référence pour les Langues : Apprendre, Enseigner, Évaluer*. Paris : Didier.

CROWTHER D., TROFIMOVICH P., SAITO K. & ISAACS T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, **49**(4), 814–837. DOI : 10.1002/tesq.203.

CRUICKSHANKS K. J., WILEY T. L., TWEED T. S., KLEIN B. E., KLEIN R., MARES-PERLMAN J. A. & NONDAHL D. M. (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. DOI : 10.1093/oxfordjournals.aje.a009713.

DE FINO V. (2024). *Caractérisation et mesure de la compéhensibilité de la parole de locuteurs non natifs dans le cadre de l'apprentissage des langues*. Thèse de doctorat, Université Toulouse III Paul Sabatier.

DERWING T. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, **59**(4), 547–567. DOI : 10.3138/cmlr.59.4.547.

DERWING T. M. & MUNRO M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, **39**(3), 379–397. DOI : 10.2307/3588486.

DERWING T. M. & MUNRO M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, **42**(4), 476–490. DOI : 10.1017/S026144480800551X.

FLEGE J. E. (1988). Factors affecting degree of perceived foreign accent in english sentences. *The Journal of the Acoustical Society of America*, **84**(1), 70–79. DOI : 10.1121/1.396876.

FONTAN L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication*. Thèse de doctorat, Université de Toulouse. HAL : tel-00797883.

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, Éds., *La voix et la parole perturbées*, p. 131–144. Mons: CIPA.

HANSEN EDWARDS J. G., ZAMPINI M. L. & CUNNINGHAM C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, **37**(4), 538–557. DOI : 10.1111/weng.12344.

KANG O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, **38**(2), 301–315. DOI : 10.1016/j.system.2010.01.005.

KENNEDY S. & TROFIMOVICH P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, **64**(3), 459–489. DOI : 10.3138/cmlr.64.3.459.

LUDWIG A. & MORA J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation*, **3**(2), 167–198. DOI : 10.1075/jslp.3.2.01lud.

MUNRO M. J. & DERWING T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, **44**(3), 316–327. DOI : 10.1017/S0261444811000103.

NAGLE C. L. & HUENSCH A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, **6**(3), 329–351. DOI : 10.1075/bct.121.04nag.

SAITO K., MACMILLAN K., KACHLICKA M., KUNIHARA T. & MINEMATSU N. (2023). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, **45**(1), 234–263. DOI : 10.1017/S0272263122000080.

SAITO K., TROFIMOVICH P. & ISAACS T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, **38**(4), 439–462. DOI : 10.1093/applin/amv047.

SUZUKI S. & KORMOS J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, **42**(1), 143–167. DOI : 10.1017/S0272263119000421.

SUZUKIDA Y. & SAITO K. (2021). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, **25**(3), 431–450. DOI : 10.1177/1362168819858246.

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*, **134**(1), 27–33. HAL : hal-01486715.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Premier système IRIT-MyFamillyUp pour la compétition sur la reconnaissance des émotions Odyssey 2024

Adrien Lafore[1, 2] Clément Pagés[1] Leila Moudjari[1] Sebastiao Quintas[1]
Isabelle Ferrané[1] Hervé Bredin[1] Thomas Pellegrini[1] Farah Benamara[1]
Jérôme Bertrand[2] Marie-Françoise Bertrand[2]
Véronique Moriceau[1] Jérôme Farinas[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) My Family Up, Toulouse, France
prenom.nom@irit.fr
mf.bertrand@myfamilyup.com j.bertrand@myfamilyup.com

RÉSUMÉ

Dans cet article, nous présentons notre contribution à la tâche de classification des émotions dans la parole dans le cadre de notre participation à la campagne d'évaluation Odyssey 2024. Nous proposons un système hybride qui tire parti à la fois des informations du signal audio et des informations sémantiques issues des transcriptions automatiques. Les résultats montrent que l'ajout de l'information sémantique permet de dépasser les systèmes uniquement audio.

ABSTRACT

IRIT-MyFamilyUp system for the Odyssey 2024 Emotion Recognition Challenge.

In this paper, we present our contribution to emotion classification in speech as part of our participation in the Odyssey 2024 challenge. We propose a hybrid system that takes advantage of both audio signal information and semantic information from automatic transcriptions. The results show that adding semantic information allows surpassing systems based solely on audio.

MOTS-CLÉS : modélisation des émotions, Compétition Odyssey 2024, fusion texte et audio.

KEYWORDS: emotion modelling, Odyssey 22024 challenge, text and audio fusion.

1 Introduction

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 article 51, 2015). 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 (Ministère de l'économie des finances et de la relance, 2021), les principales observations sont :

— leur âge moyen est de 49 ans et 37% des aidants sont âgés de 50 à 54 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 (YouGov, 2019) et des enquêtes de la DREES [1] 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. C'est dans ce contexte que nous nous intéressons à l'identification des sentiments et états émotionnels dans la communication orale afin de développer un détecteur d'états émotionnels dans la parole, qui permettrait d'aider au diagnostic psychologique des proches aidants. Notre objectif est d'exploiter des informations audio et sémantiques afin de modéliser les états émotionnels par la détection de ces états dans la sémantique textuelle (parole retranscrite) et la détection d'émotions dans la prosodie.

Les émotions ont largement été étudiées dans un cadre théorique. On peut citer notamment les modèles de représentation de Plutchik (Plutchik, 1980) ou d'Ekman (Ekman & Journet, 2002) pour les plus connus.

Dans le cadre du Traitement Automatique des Langues (TAL), la détection des états émotionnels ou psychologiques a surtout été abordée dans le cadre de la détection des maladies mentales comme la dépression, les troubles de l'alimentation ou les troubles bipolaires (cf. (Harrigian *et al.*, 2021) pour une revue des collections de données et des tâches existantes). Les campagnes d'évaluation Computational Linguistics and Clinical Psychology (CLPsych) (Zirikly *et al.*, 2022) ou eRisk (Parapar *et al.*, 2023) se sont penchées en particulier sur la tâche de détection automatique de la dépression chez des utilisateurs des réseaux sociaux, avec un focus sur la détection "au plus tôt". Les modèles d'apprentissage développés pour la détection automatique sont à base soit de traits pour les plus performants (utilisation de pronoms personnels, sentiment positif ou négatif, temps des verbes, etc.) (Bae *et al.*, 2021; Molina *et al.*, 2023), soit d'apprentissage profond (cf. (Ríssola *et al.*, 2021) pour un panorama des méthodes automatiques existantes pour la classification des états mentaux sur les réseaux sociaux). Les travaux actuels dans ce domaine portent ainsi quasi uniquement sur des données écrites par des utilisateurs plutôt jeunes des réseaux sociaux.

En ce qui concerne le Traitement Automatique de la Parole (TAP), la détection d'émotion est un domaine de recherche qui est issu de l'analyse automatique de la prosodie. En effet, le champ des informations non verbales constitue la source des informations pour caractériser les émotions. Des compétitions internationales ont eu lieu depuis 2009 afin de faire avancer la connaissance sur cette problématique : The Interspeech 2009 Emotion Challenge (Schuller *et al.*, 2009) et The Interspeech 2010 Paralinguistic challenge (Schuller *et al.*, 2010). Les premiers systèmes étaient basés sur des systèmes discriminants alimentés par de nombreux paramètres extraits du signal audio, comme la boîte à outil OpenSmile (Eyben *et al.*, 2010). Les systèmes actuels se basent sur des architectures de réseaux de neurones profonds et ont permis d'obtenir de bonnes performances en reconnaissance des émotions sous forme de catégories (Tripathi *et al.*, 2018; Yenigalla *et al.*, 2018; Atmaja *et al.*, 2019; Yoon *et al.*, 2019). L'enjeu consiste maintenant à projeter les émotions dans un espace continu, ce qui permettra l'étude des états mentaux cognitifs (Atmaja & Akagi, 2020).

Le travail présenté dans cet article détaille une première contribution dans le cadre de notre participa-

1. https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/les-enquetes-capacites-aides-et-ressources-des-seniors-care

tion à la campagne d'évaluation Odyssey 2024 [2] portant sur la reconnaissance des émotions. Nous proposons ici un travail conjoint qui vise à utiliser des techniques de l'état de l'art pour combiner l'exploitation de données issues des retranscriptions d'extraits vidéo collectés sur internet et des données issues de la piste audio.

Dans les sections suivantes, nous présentons d'abord le corpus mis à notre disposition ainsi que la tâche de classification proposée. Ensuite, nous présentons en détail l'architecture de notre système. Enfin, nous détaillons et analysons les résultats obtenus.

2 Protocole expérimental

Nous présentons ici la campagne d'évaluation Odyssey 2024, à savoir : les données, la tâche proposée ainsi que les métriques d'évaluation utilisées.

2.1 Corpus

Les données mises à disposition sont des enregistrements en anglais issus du corpus MSP-Podcast (Lotfian & Busso, 2019), qui contient des segments audio provenant de podcasts en ligne. Les tours de parole ont été annotés par au moins 5 annotateurs selon les catégories d'émotion et leurs dimensions.

Les catégories d'émotions annotées dans ce corpus sont : Anger (colère), Contempt (mépris), Disgust (dégoût), Fear (peur), Happiness (bonheur), Neutral (neutre), Sadness (tristesse), Surprise, Other (autre), et No agreement (pas d'accord inter-annotateurs).

Les dimensions pour chaque émotion sont la *valence* (état positif ou négatif de l'individu), l'*arousal* (activité ou passivité de l'individu) et la *dominance* (contrôle faible à fort). Chacune de ces dimensions est annotée sur une échelle de 1 à 7.

Les données d'entraînement et de développement sont composées respectivement de 68 360 et 19 815 tours de parole annotés. Pour ces données, les transcriptions sont fournies ainsi que le genre des intervenants. Les données de l'ensemble de test sont constituées de 2 347 segments de parole venant de 187 personnes. Pour ces dernières, aucune transcription n'est fournie. De plus, les classes "Other" (O) et "No agreement" (X) ont été supprimées et la distribution des catégories d'émotion est équilibrée. Ainsi, nous avons aussi retiré ces deux classes du jeu de données d'entraînement et de développement. Le tableau 1 montre la distribution des données d'entraînement et de développement.

2.2 Tâche et métrique d'évaluation

La tâche à laquelle nous avons participé est celle de classification des émotions en 8 catégories : colère (A), mépris (C), dégoût (D), peur (F), bonheur (H), neutre (N), tristesse (S), et surprise (U).

La campagne d'évaluation n'autorise pas l'utilisation de modèles existants entraînés pour la détection d'émotions.

Les systèmes participants sont évalués selon les mesures classiques de précision, rappel, F1-score et

2. https://www.odyssey2024.org/emotion-recognition-challenge

Catégorie	Entraînement			Développement		
	Nombre	**%**	**Durée totale**	**Nombre**	**%**	**Durée totale**
Neutral (N)	25 106	36,72	39h43	5 667	28,60	08h47
No agreement (X)	13 709	20,05	22h02	4 013	20,25	06h32
Happiness (H)	13 440	19,66	22h09	3 340	16,86	05h14
Sadness (S)	3 882	5,68	06h13	1 101	5,56	01h44
Anger (A)	3 053	4,47	05h05	2 413	12,18	04h01
Surprise (U)	2 897	4,24	04h38	729	3,68	01h04
Contempt (C)	2 443	3,57	04h09	1 323	6,67	02h17
Disgust (D)	1 426	2,09	02h24	486	2,45	00h52
Other (O)	1 265	1,85	02h05	461	2,33	00h46
Fear (F)	1 139	1,67	01h46	282	1,42	00h26
TOTAL	68 360		110h14	19 815		32h06

TABLE 1 – Distribution des catégories d'émotion dans les données d'entraînement et de développement

accuracy. La distribution des classes dans les données de test étant équilibrée, la macro-F1 est utilisée pour classer les systèmes ; la macro-F1 étant la moyenne des F1-scores pour chacune des 8 classes (voir formule 1).

Pour comparer les performances de nos systèmes, nous avons durant ce challenge utilisé un jeu de données équilibré issu du jeu de développement fourni par les organisateurs de ce challenge. Nous l'avons construit par échantillonage de façon à calculer les performances de nos systèmes dans le même cadre que le jeu de Test qui lui aussi est équilibré.

$$\text{Macro F1} = \frac{1}{8}\sum_{i=1}^{8} 2 \times \frac{\text{précision}_i \times \text{rappel}_i}{\text{précision}_i + \text{rappel}_i} \tag{1}$$

3 Système proposé

Dans le cadre de la tâche de classification des émotions, nous proposons un premier système combinant les informations prosodiques et sémantiques à notre disposition. Dans ce premier système hybride, le but est de calculer les probabilités d'émotions pour chaque fichier en entrée (audio et texte) et d'utiliser la moyenne de ces probabilités pour notre prédiction. Les données de test pour ce challenge étant uniquement audio, nous avons dû utiliser un premier système de reconnaissance de la parole (cf. section 3.2) qui fournit les transcriptions au modèle sémantique entraîné pour cette tâche. En parallèle, le modèle audio (cf. section 3.1), lui aussi entraîné pour cette tâche, prend les segments audio fournis (d'une durée entre 3 et 11 secondes) pour calculer de son côté les probabilités d'émotion.

3.1 Modélisation acoustique

Le système dédié à la tâche de classification d'émotions en se basant uniquement sur l'audio a été développé et entraîné à partir de la librairie pyannote.audio (Bredin, 2023; Plaquet & Bredin, 2023), et

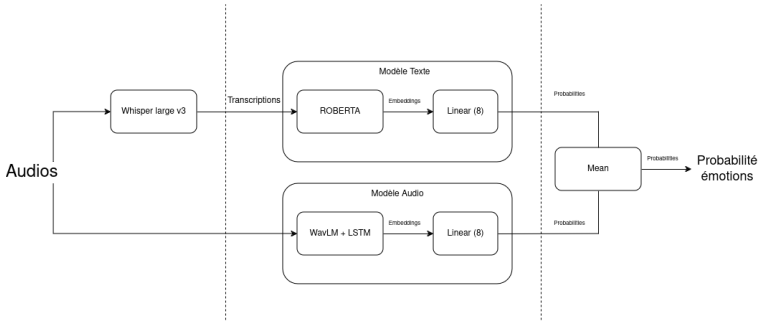


FIGURE 1 – Premier système hybride proposé

est inspiré de l'architecture de SSerRiouSS [3]. Le système prend en entrée des morceaux d'audio d'une durée de 10 secondes qui sont tout d'abord traités par un modèle WavLM-large (Chen *et al.*, 2022) pré-entraîné sur le corpus de LibriSpeech. La sortie fournie par ce module correspond à la moyenne pondérée des sorties de chacune des 12 couches composant le modèle WavLM-large, les poids utilisés pour cette moyenne étant appris durant l'entraînement. Cette sortie est ensuite injectée dans une pile de LSTM bidirectionnels (BLSTM) parcourant la séquence de caractéristiques produite par le modèle dans les deux sens. La séquence de trames issue de cette pile de BLSTM est ensuite passée à travers une couche linéaire dont le but est de faire la classification des émotions au niveau de chaque trame. L'étape suivante consiste à calculer la moyenne de ces classifications à l'aide d'une couche de mise en commun (mean pooling), la sortie de celle-ci étant finalement donnée à une fonction d'activation de type log-softmax, associant à chaque émotion possible une probabilité.

Ce modèle a été entraîné sur la partition d'entraînement du corpus MSP-PODCAST. Le nombre de BLSTM a été fixé à 2. Ce nombre correspond à la valeur par défaut dans l'architecture de SSeRiousSS, et également à celle pour laquelle les meilleurs résultats ont été obtenus. L'entraînement du système a été effectué à l'aide de l'optimiseur Adam, avec un taux d'apprentissage initial de 10e-3, ce dernier étant divisé par deux sans amélioration du F1-score macro sur 10 epochs consécutives, jusqu'à une valeur minimale de 10e-8. Le modèle WavLM-large a été gelé. La fonction de perte utilisée pour l'entraînement est une Vraisemblance logarithmique négative, pondérée avec des poids inversement proportionnels au nombre de représentants de chaque classe. La taille des lots a été fixée expérimentalement à 32.

3.2 Transcription de la parole

Des transcriptions manuelles ont été fournies pour les sous-ensembles d'entraînement et de développement, mais pas pour le jeu de test. Nous avons donc eu besoin d'utiliser un système de reconnaissance automatique de la parole (RAP) afin de fournir des transcriptions au modèle texte de classification d'émotions. Nous avons utilisé le système Whisper (Radford *et al.*, 2022), et en particulier le modèle whisper-large-v3. Ce modèle a été entraîné avec 680 000 heures de données supervisées multilingues et multitâches collectées sur le Web, un aspect qui le rend robuste à différents types d'accents et conditions acoustiques.

Le système de RAP génère des transcriptions forcément différentes des transcriptions manuelles fournies, et il peut en résulter une perte de performance si trop d'erreurs de transcription sont

3. Le code de cette architecture est accessible ici : https://github.com/pyannote/pyannote-audio/blob/develop/pyannote/audio/models/segmentation/SSeRiouSS.py

commises. Pour tenter d'évaluer la similarité entre les deux types de transcription, nous avons mesuré des taux d'erreur mot (TEM) sur les jeux d'entraînement et de développement. Avec le modèle whisper-large-v3, ce taux est d'environ 32% sur les deux jeux. Notons que nous avons testé d'autres modèles Whisper, plus petits (de *tiny* à *medium*), et le TEM augmentait inversement avec la taille du modèle.

Whisper fournit des transcriptions qui contiennent de la ponctuation et des majuscules pour les noms propres et autres acronymes. Si nous les normalisons (suppression des signes de ponctuation et de la casse), le TEM est drastiquement réduit à 12-13% sur les deux sous-ensembles. Nous avons choisi de garder la ponctuation et la casse cependant pour la modélisation des émotions, car a priori ces éléments donnent des informations probablement pertinentes pour cette tâche, comme par exemple un point d'exclamation qui peut exprimer de la surprise ou de la colère.

3.3 Modélisation des transcriptions

N'étant pas autorisés à utiliser des modèles dédiés à la détection d'émotions, nous avons entraîné et évalué plusieurs modèles de langue de l'état de l'art sur les transcriptions de référence des données d'entraînement et de développement : BERT (base, multilingual, large, etc.) (Devlin *et al.*, 2019) et RoBERTa-base (Liu *et al.*, 2019). Les meilleures performances ont été obtenues avec RoBERTa, c'est pourquoi nous avons fait le choix de ce modèle pour notre système hybride.

Le modèle RoBERTa-base est composé de 12 couches avec une taille cachée de 768 et un nombre de têtes d'attention de 12. Les couches intermédiaires (feedforward) ont une taille de 3072. Le modèle est construit sur un vocabulaire de 50 265 mots, y compris les jetons spéciaux pour le début et la fin de séquence (CLS et SEP). De plus, le modèle est sensible à la casse et utilise des jetons de masquage aléatoire (MLM) pour l'entraînement.

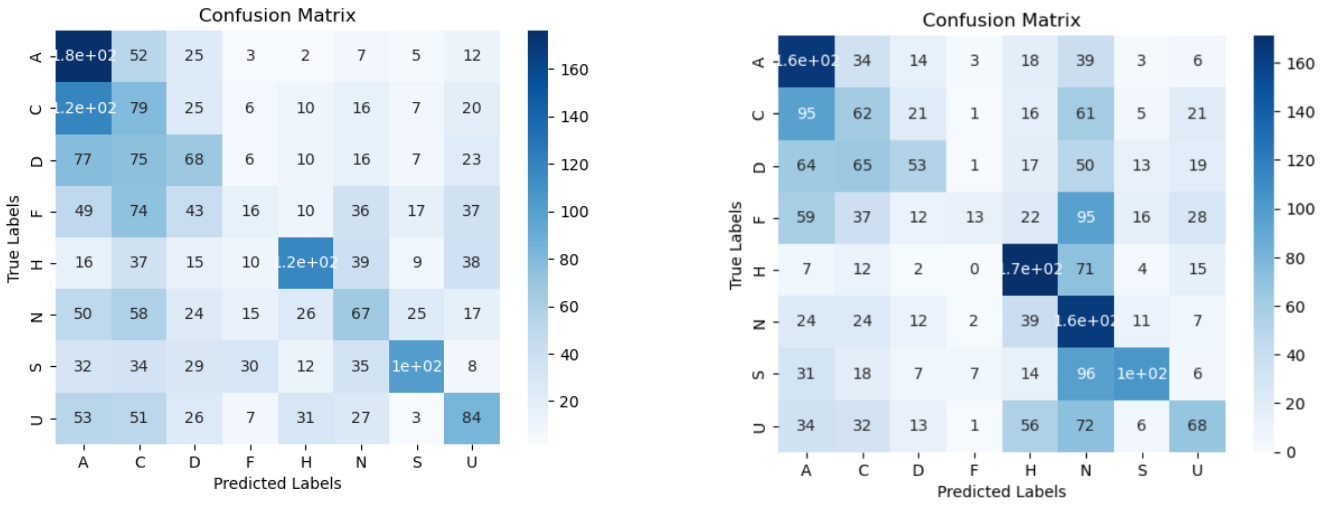
Nous avons adapté RoBERTa-base (125 millions de paramètres) aux données d'entraînement. Pour ceci, nous avons utilisé l'optimiseur Adam avec un taux d'apprentissage de $2e-5$ pendant 4 epochs. Les tailles de lots (batch sizes) ont été expérimentalement fixées à 64, et nous avons utilisé une fonction de perte d'entropie croisée pondérée (crossentropyweighted). Enfin, une couche linéaire a été ajoutée pour la tâche de classification des émotions. Comme pour le modèle acoustique, la sortie de cette couche est ensuite soumise à une fonction d'activation de type log-softmax, attribuant à chaque classe une probabilité.

3.4 Fusion des informations

Premièrement, comparons les résultats (sur notre jeu de développement équilibré) des systèmes acoustiques et textuels indépendamment pour établir plus tard du bénéfice de la fusion des deux informations pour notre tâche de classification.

	F1 Macro	**F1 Micro**	**Accuracy**
Système Audio	0.3073	0.3147	0.3147
Système Texte	0.2602	0.2836	0.2836

TABLE 2 – Résultats des deux systèmes sur notre jeu de développement équilibré



(a) Système audio

(b) Système texte

FIGURE 2 – Matrice de confusion des systèmes

Nous pouvons noter les différences entre les deux systèmes : l'émotion neutre est beaucoup mieux reconnue avec l'information textuelle tandis que le système audio semble plus performant en général.

Comme première expérimentation, nous proposons donc ici une méthode simple pour la fusion des résultats des deux modèles prosodique et sémantique. Nous départageons les deux modèles en faisant la moyenne de leurs résultats respectifs (probabilité pour chaque classe) afin de lisser les faiblesses ponctuelles de chacun (notamment lorsqu'un des deux systèmes est indécis).

4 Résultats et discussion

Le tableau 3 présente les résultats obtenus sur les ensembles de développement et de test.

À titre de comparaison, la campagne d'évaluation met à disposition une baseline (Goncalves *et al.*, 2024) présentée dans la figure 3. Leur méthode s'appuie uniquement sur un modèle audio constitué d'un encodeur WavLM (Chen *et al.*, 2022) pré-entraîné avec un taux d'apprentissage de $1e-5$, 20 epochs et une taille de lots de 32.

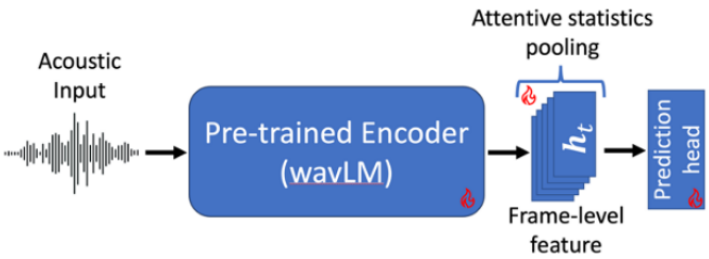


FIGURE 3 – Baseline de la campagne Odyssey 2024

Les meilleurs résultats obtenus par notre système se situent sur les émotions joie, tristesse et colère qui

Catégorie	Développement			Test		
	Précision	**Rappel**	**F1-score**	**Précision**	**Rappel**	**F1-score**
Neutral (N)	0.25	0.58	0.35			
Happiness (H)	0.48	0.61	0.54			
Sadness (S)	0.64	0.37	0.47			
Anger (A)	0.34	0.59	0.43			
Surprise (U)	0.40	0.24	0.30			
Contempt (C)	0.22	0.22	0.22			
Disgust (D)	0.40	0.19	0.25			
Fear (F)	0.46	0.05	0.08			
Accuracy	0,3537			**0.3511**		
Macro F1	0,3308			**0.3335**		
Baseline Accuracy	-			0,3272		
Baseline Macro F1	-			0,3113		

TABLE 3 – Résultats obtenus sur les données de développement et de test

ont des marqueurs sémantiques et prosodiques plus facilement identifiables et le neutre. La surprise, le mépris, le dégoût semblent elles difficilement identifiables par notre système. La peur, avec un F1-score de 0,08 est le point faible de notre système de classification.

Nous pouvons voir une petite baisse de précision lors de la prédiction sur l'ensemble de test (dont les labels ne sont encore pas publics) mais une amélioration majeure de la macro F1 permettant de dépasser la baseline de cette campagne.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une première expérimentation pour la classification des émotions dans la parole. Le système hybride proposé repose sur un modèle acoustique et un modèle sémantique tous deux entraînés pour la tâche. Les résultats obtenus montrent qu'une fusion simple (moyenne des probabilités des deux modèles pour chaque classe) permet d'atteindre des résultats qui dépassent ceux d'un modèle acoustique seul. À court terme, nous envisageons de tester d'autres méthodes de fusion de l'audio et du texte, par exemple une concaténation des vecteurs de sortie des systèmes audio et texte, suivi d'une ou plusieurs couches permettant la classification. Cela pourrait nous permettre de mieux gérer les faiblesses de chaque système pour profiter au maximum de la dualité d'information que nous utilisons.

Remerciements

Ces travaux ont bénéficié d'un accès aux calculateur Jean Zay de l'IDRIS au travers des allocations de ressources AD011014274 et AD011013612R1 attribuées par GENCI.

Références

ATMAJA B. & AKAGI M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, **9**(1).

ATMAJA B., SHIRAI K. & AKAGI M. (2019). Speech emotion recognition using speech feature and word embedding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*.

BAE Y. J., SHIM M. & LEE W. H. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors*, **21**(17), 5924.

BREDIN H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., WU J., ZENG M., YU X. & WEI F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1505–1518. arXiv :2110.13900 [cs, eess], DOI : 10.1109/JSTSP.2022.3188113.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*.

EKMAN P. & JOURNET N. (2002). *De l'universel au particulier*, In N. JOURNET, Éd., *La culture*, chapitre Le langage naturel des émotions, p. 29–37. Éditions Sciences Humaines : Auxerre.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM), ACM*.

GONCALVES L., SALMAN A., REDDY A., VELAZQUEZ L. M., THEBAUD T., GARCIA L. P., DEHAK N., SISMAN B. & BUSSO C. (2024). Odyssey 2024 - emotion recognition challenge. https://github.com/MSP-UTD/MSP-Podcast_Challenge.

HARRIGIAN K., AGUIRRE C. & DREDZE M. (2021). On the state of social media data for mental health research. In N. GOHARIAN, P. RESNIK, A. YATES, M. IRELAND, K. NIEDERHOFFER & R. RESNIK, Éds., *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology : Improving Access*, p. 15–24, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.clpsych-1.2.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv :1907.11692.

Loi n°2015-1776 article 51 (2015). Loi n° 2015-1776 du 28 décembre 2015 relative à l'adaptation de la société au vieillissement (article 51). JORF https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000031706430.

LOTFIAN R. & BUSSO C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, **10, no. 4**.

MINISTÈRE DE L'ÉCONOMIE DES FINANCES ET DE LA RELANCE (2021). Guide ministériel du proche aidant. https://www.economie.gouv.fr/files/files/2021/guide_proche-aidant.pdf.

MOLINA A., HUANG X., HURTADO L.-F. & PLA F. (2023). ELiRF-UPV at eRisk 2023 : Early detection of pathological gambling using SVM. In *CLEF - CEUR-WS Working Notes*.

PARAPAR J., MARTIN-RODILLA, PATRICIA ANS LOSADA D. E. & CRESTANI F. (2023). Overview of eRisk at CLEF 2023 : Early Risk Prediction on the Internet. In *CLEF - CEUR-WS Working Notes*.

PLAQUET A. & BREDIN H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

PLUTCHIK R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, p. 3–33. Elsevier.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. DOI : 10.48550/ARXIV.2212.04356.

RÍSSOLA E. A., LOSADA D. E. & CRESTANI F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, **2**(2), 1–31.

SCHULLER B., STEIDL S. & BATLINER A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. Interspeech 2009*, p. 312–315. DOI : 10.21437/Interspeech.2009-103.

SCHULLER B., STEIDL S., BATLINER A., BURKHARDT F., DEVILLERS L., MÜLLER C. & NARAYANAN S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Proc. Interspeech 2010*, p. 2794–2797. DOI : 10.21437/Interspeech.2010-739.

TRIPATHI S., SAMARTH S. & HOMAYOON B. (2018). Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv preprint arXiv :1804.05788.

YENIGALLA P., KUMAR A., TRIPATHI S., SINGH C., KAR S. & VEPA J. (2018). Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In *Proc. Interspeech 2018*.

YOON S., BYUN S., DEY S. & JUNG K. (2019). Speech Emotion Recognition Using Multi-hop Attention Mechanism. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

YOUGOV I. (2019). Sondage Le Huffigton Post. Publié le 16/09/2019 dans le Huffington Post, https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/20u9aq6czp/Copy%20of%20Results%20for%20YouGov%20%28Huf%20Post%20Psy%29%20159%2013.9.2019.pdf.

ZIRIKLY A., ATZIL-SLONIM D., LIAKATA M., BEDRICK S., DESMET B., IRELAND M., LEE A., MACAVANEY S., PURVER M., RESNIK R. & YATES A., Éds. (2022). *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, Seattle, USA. Association for Computational Linguistics.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Preuve de concept d'un système de génération automatique en Langue française Parlée Complétée

Brigitte Bigi, Núria Gala
LPL, CNRS, Aix-Marseille Univ, 5 avenue Pasteur, 13100 Aix-en-Provence
brigitte.bigi@cnrs.fr, nuria.gala@univ-amu.fr

RÉSUMÉ

La Langue française Parlée Complétée (LfPC) est un système de communication développé pour les personnes sourdes afin de compléter la lecture labiale avec une main, au niveau phonétique. Il est utilisé par les enfants pour acquérir des compétences en lecture, en lecture labiale et en communication orale. L'objectif principal est de permettre aux enfants sourds de devenir des lecteurs et des locuteurs compétents en langue française. Nous proposons une preuve de concept (PoC) d'un système de réalité augmentée qui place automatiquement la représentation d'une *main codeuse* sur la vidéo pré-enregistrée d'un locuteur. Le PoC prédit la forme et la position de la main, le moment durant lequel elle doit être affichée, et ses coordonnées relativement au visage dans la vidéo. Des photos de mains sont ensuite juxtaposées à la vidéo. Des vidéos annotées automatiquement par le PoC ont été montrées à des personnes sourdes qui l'ont accueilli et évalué favorablement.

ABSTRACT

Toward an Automatic Cued Speech System for French Language

Cued Speech is a communication system developed for deaf people to complement speechreading at the phonetic level with hands. It is used by children to acquire skills in reading, in lip reading and oral communication. The main goal is to allow deaf children to become proficient readers and speakers of an oral language. We propose a Proof of Concept (PoC) of an augmented reality system that automatically places the representation of a coding hand on a video of a pre-recorded speaker. The PoC is predicting the key to be coded (shape and position), when it has to be coded relatively to the audio and its coordinates relatively to the face in the video. Photos of human hands are then juxtaposed to the video. Videos automatically encoded with this system have been shown to deaf people who have welcomed and positively evaluated.

MOTS-CLÉS : LfPC, automatisation, PoC, surdité, vidéo, annotation.

KEYWORDS: Cued Speech, automatic, PoC, deaf, video, annotation.

1 Introduction

Lorsque la LSF n'est pas utilisée, la lecture labiale est l'une des principales modalités visuelles qui permet l'accès à la parole pour les personnes sourdes ou malentendantes. Elle est utilisée en conjonction avec d'autres stratégies de communication, comme les aides auditives, et/ou des solutions visuelles. Parmi ces dernières, en 1966, R. Orin Cornett a inventé le « Cued Speech » (CS), un codage qui ajoute des informations visuelles sur les sons qui ne sont pas différentiables sur les lèvres (Cornett, 1967). Ce codage CS représente chaque son avec une forme de main pour une consonne et

une position autour du visage pour une voyelle. Leur combinaison forme une *clé*. Lorsque les sons se ressemblent sur les lèvres, ils sont codés différemment ; la combinaison entre forme labiale et clé implique un percept unique de ce qui est prononcé. Par exemple, "bi" et "mi" qui sont identiques sur les lèvres sont codés avec deux formes différentes de la main. Le CS est souvent utilisé dans les milieux éducatifs, en particulier pour les jeunes enfants ayant une déficience auditive, car il leur donne accès à la langue orale *via* l'information phonémique qu'ils pourraient manquer par des moyens auditifs traditionnels. L'objectif majeur du CS est ainsi de faire en sorte que les enfants sourds puissent accéder plus facilement à la communication orale. L'efficacité de ce codage pour améliorer la perception et la production de la parole a été démontrée dans un grand nombre d'études, notamment (Kaplan, 1975; Neef & Iwata, 1985; Leybaert *et al.*, 2010).

L'automatisation du « Cued Speech », c-à-d l'utilisation d'un système automatisé pour coder et/ou décoder les sons, concerne essentiellement deux domaines : la synthèse – codage, dont cet article fait l'objet, et la reconnaissance –décodage. Avec un système de codage automatique, toutes sortes de vidéos codées pourraient être élaborées et diffusées pour tous les types d'utilisations. Disposer d'outils permettant de s'entraîner à la pratique du code constituerait un bénéfice important pour les parents d'enfants sourds, ainsi que pour les centres d'éducation spécialisée, par exemple. Cela permettrait entre autres de réduire les inégalités d'accès à la LfPC sur le territoire, d'apporter une aide à l'acquisition de la langue orale par les enfants sourds, d'améliorer la communication entre les personnes sourdes ou malentendantes et les membres de leur famille entendants, ou d'aider à développer des compétences de lecture labiale. Dans ce domaine, le premier système *AutoCuer* avait été proposé par l'inventeur du codage (Cornett *et al.*, 1977). Par la suite, dans les années 1995-2000, plusieurs recherches ont été conduites au *Massachusetts Institute of Technology* (MIT) pour automatiser le codage (Bratakos, 1995; Sexton, 1997; Bratakos *et al.*, 1998; Duchnowski *et al.*, 2000). Ces travaux ont consisté à vérifier la faisabilité d'automatiser la génération des clés, c'est-à-dire à déterminer la séquence de clés qui doit être produite puis générer une vidéo augmentée d'une main codeuse. Un locuteur était filmé pendant qu'il parlait, sans coder. Un système de reconnaissance automatique de la parole permettait d'obtenir la séquence des phonèmes à partir desquels le système déterminait les clés à incruster dans la vidéo. Dans une autre pièce, se trouvait une personne qui devait décoder la vidéo ainsi générée en temps réel. Dans les versions successives de ce système, les mains étaient représentées par des *cliparts*. Les différentes évaluations ont toujours montré *a minima* un petit avantage du décodage avec ajout de l'image de la main codeuse par rapport à la lecture labiale seule.

Malgré les nombreux travaux démontrant ses avantages, et l'intérêt grandissant qu'il suscite, il n'existe aucun système de génération automatique des clés. Les études récentes relatives à l'automatisation du CS se concentrent, en effet, sur la reconnaissance (Sankar *et al.*, 2023). En outre, il n'existe que très peu d'études qui décrivent le fonctionnement du codage, notamment en ce qui concerne l'organisation temporelle et spatiale du code dans sa co-production avec la parole (Attina, 2005).

2 Méthodologie proposée

La figure 1 illustre le processus que nous proposons pour implémenter un système de codage automatisé, c-à-d un système qui permettra d'ajouter une main codeuse artificielle à la vidéo d'un locuteur. En amont, l'utilisateur doit préparer un enregistrement audio-vidéo et sa transcription orthographique, alignée dans des unités courtes, telles que les Unités Inter-Pausales (IPUs). Le

système utilise le logiciel libre SPPAS (Bigi, 2015) pour obtenir les annotations audio-vidéo requises en entrée. Pour chaque IPU, SPPAS détermine 1/ la séquence des phonèmes et leur alignement temporel avec l'audio, 2/ les coordonnées de 68 points spécifiques du visage du locuteur dans chacune des images de la vidéo. A partir de ces informations, l'objectif est de produire des annotations qui permettront d'augmenter la vidéo automatiquement avec la main codeuse.

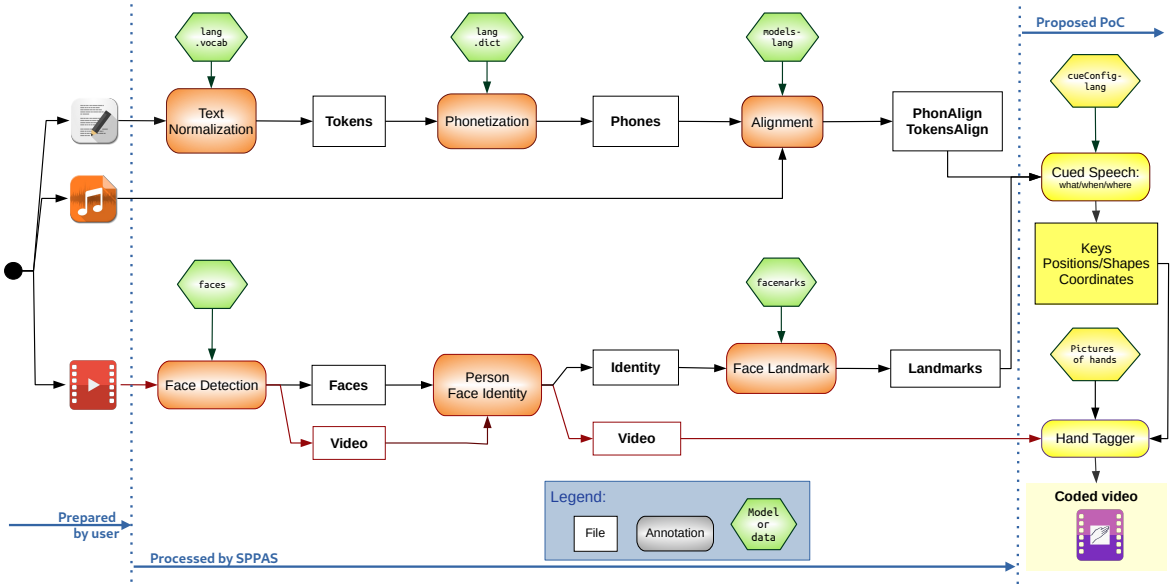


FIGURE 1 – Processus proposé pour la génération automatique du codage

Il est prévu que le système final soit implémenté avec des approches hybrides, donc partiellement basé sur des modèles à base de connaissance et partiellement sur des modèles issus de méthodes empiriques. Il sera développé pour le français ; dans ce cas, on utilise l'acronyme francophone LfPC, pour Langue française Parlée Complétée, plutôt que « Cued Speech ».

Pour ce faire, nous avons d'ores et déjà collecté un corpus, le Corpus de Lecture en LfPC (CLeLfPC) accessible sous licence libre (Bigi *et al.*, 2022). Il se compose de 4 heures d'enregistrements audio-vidéo, de 23 locuteurs codant en LfPC. Son enrichissement avec des annotations permettra les analyses requises pour la création de modèles de synchronisation audio-main basés sur des apprentissages supervisés, en répondant à quatre sous-problématiques que nous avons définies, et qui déterminent :

1. **quoi** : la séquence des clés à produire à partir des phonèmes,
2. **quand** : les moments de présentations et de transitions des formes et positions de la main,
3. **où** : les coordonnées, angle et taille de la main par rapport au visage, et,
4. **comment** : l'incrustation de représentations d'une main dans chaque image de la vidéo

Compte tenu du coût de la collecte et de l'annotation d'un corpus, avant d'aller plus avant dans la création de ce système, il nous a semblé indispensable de connaître l'intérêt réel que le système pourrait apporter, et d'en étudier la faisabilité.

Parallèlement à la création du corpus, nous avons donc élaboré une preuve de concept (PoC), qui est une phase déterminante pour **décider si le système peut et doit être implémenté**. La création des annotations du corpus, leur analyse et, d'une manière générale, l'ensemble des recherches liées à la mise en oeuvre des modèles et du système sont conditionnés **par l'approbation du PoC**. Il se compose de quatre modules, chacun impliquant de répondre à des problématiques spécifiques, afin de

déterminer la séquence de clés à produire à partir d'un signal d'entrée audio vidéo (lecture à haute voix) et de sa transcription orthographique, l'organisation temporelle entre la main et les phonèmes, le positionnement de la main par rapport au visage, et enfin le marquage de la vidéo avec la main.

3 Description de la preuve de concept

La preuve de concept est un système qui produit automatiquement des fichiers XML contenant les informations relatives au codage en LfPC. Ces fichiers contiennent les annotations indiquant quelles sont les formes et positions de la main qu'il faut intégrer, à quel moment et où il faut les placer dans la vidéo. D'autre part, le PoC crée un fichier vidéo augmenté avec la représentation de la main codeuse. En entrée, le PoC nécessite de connaître la séquence des phonèmes prononcés et leur position temporelle par rapport à l'audio, ainsi que les coordonnées du visage du locuteur pour chacune des images de la vidéo, comme indiqué dans la figure 1.

3.1 Quelles sont les clés ?

Afin de déterminer la séquence des clés à produire à partir des phonèmes, le PoC implémente un système à base de règles de productions, élaborées en collaboration avec des experts du codage.

Pour l'implémenter, dans un premier temps nous avons assigné un numéro à chacune des 8 formes de la main ainsi qu'à la forme neutre (voir figure 2), et nous avons assigné une lettre à chacune des 5 positions que compte la LfPC, autour du visage, et une lettre pour la position neutre sur la poitrine. Le système a donc pour tâche de proposer la séquence de clés qui correspond à la séquence de phonèmes, comme dans l'exemple suivant, dont les phonèmes sont codés en X-SAMPA :

```
entrée: 9~ d @ m i p o d H i l d @ k o k o
sortie: 5t.1s.5m.1s.1s.4m.6s.1s.2s.2s
```

Nous avons manuellement annoté une partie du corpus CLeLfPC (5 locuteurs) afin de déterminer les clés produites par les locuteurs, et nous les avons comparées aux clés prédites par le PoC (Auteur, 2023). Cette évaluation a permis de valider le modèle à base de règles que nous proposons. Cependant, la variabilité dans la production orale, notamment liée aux accents ou au contexte de la production, implique une difficulté dans cette tâche et des améliorations sont donc possibles et envisagées, par exemple en laissant le choix à l'utilisateur de modifier les règles de production.

3.2 Quand présenter les clés ?

Dans le codage CS, une clé correspond à un groupe spécifique de phonèmes (consonne + voyelle, consonne seule ou voyelle seule). Les mouvements de la main, forme pour la consonne et position pour la voyelle, doivent coïncider avec les phonèmes produits. Cette coordination précise est essentielle pour transmettre avec précision les nuances du langage parlé et combler le fossé entre la communication visuelle et auditive. (Cornett, 1967) avait déjà indiqué que les lèvres et les mouvements de la main n'apparaissent pas en même temps. Par la suite, (Bratakos *et al.*, 1998) indique que la main doit être en avance sur le son, qu'un retard de 33ms n'a que peu de conséquences et que le retard maximal acceptable est de 100ms. (Duchnowski *et al.*, 1998) démontrent ensuite que les scores de décodages sont meilleurs si la main est présentée 100ms avant le mouvement labial. Enfin, (Duchnowski *et al.*,

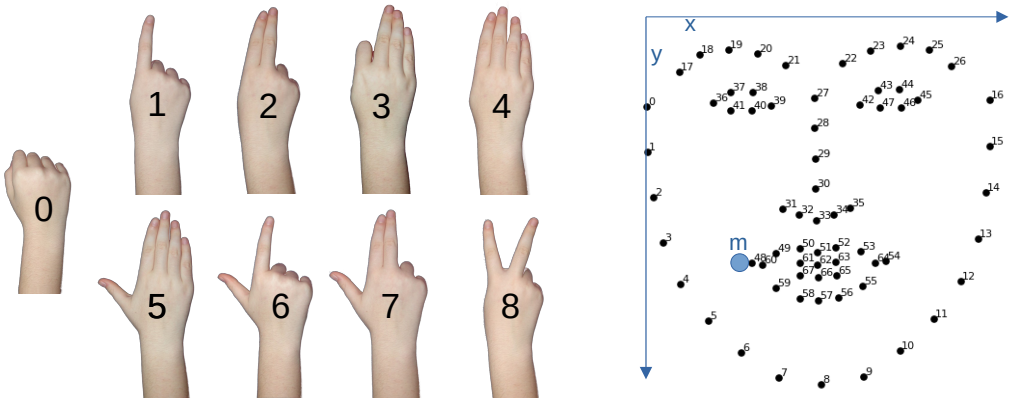


FIGURE 2 – Représentation du codage en LfPC

2000) indiquent qu'une transition de 150ms doit être opérée pour changer de position. D'autres études ont ensuite été menées pour la langue française, notamment dans (Cathiard *et al.*, 2003; Attina, 2005; Aboutabit, 2007) et ont abouti à la proposition de modèles de synchronisation main-lèvres-son, comme par exemple celui simplifié dans la figure 3. Cette figure représente le modèle que nous avons implémenté dans la preuve de concept. M1 indique le moment durant lequel la main commence à changer de position et M2 le moment où la main arrive à la position cible ; D1 indique le début du changement de forme de la main et D2 son accomplissement. Cependant, il ne couvre pas toutes les situations, et nous l'avons complété de règles à l'aide d'experts du codage. Entre-autres, nous avons traité les cas particuliers relatifs à la transition depuis la position neutre vers une position du visage (anticipée), et la transition depuis une position du visage vers la position neutre (retardée).

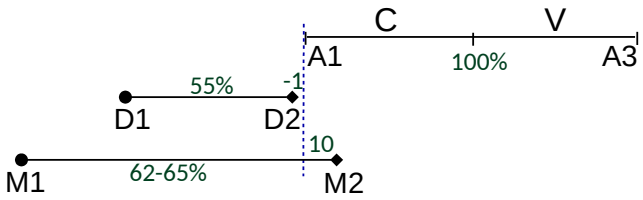


FIGURE 3 – Reproduction partielle de la synchronisation main-son, (Attina, 2005), figure 32 pp. 136. A1et A3 sont respectivement les moments de début et de fin des phonèmes consonne C, et voyelle V.

3.3 Où placer la main par rapport au visage ?

Il faut également modéliser la trajectoire de la main en proposant des modèles qui peuvent prédire à tout moment l'emplacement, l'angle et la taille de la main par rapport au visage du locuteur. Dans ce domaine, nous n'avons trouvé aucune étude antérieure publiée. Dans un premier temps, nous nous sommes appuyés sur les experts du codage pour placer les positions des voyelles par rapport à un visage théorique. Nous avons ensuite décrit ces positions par rapport aux 68 points de ce visage qui sont obtenus avec un système de "Face landmark". Par exemple, les coordonnées de la position 'm' qui représente les sons /i/, /a˜/ et /O˜/ et se situe proche du coin de la bouche, se calculent avec : $x = x_{48} - |(x_{54} - x_{48})/4|$, $y = y_{60}$. Cette position est illustrée dans la figure 2.

Nous n'avons introduit aucune variabilité dans l'estimation de ces coordonnées. Ainsi, si la main doit se trouver à cette position, le doigt cible (bout de l'index ou bout du majeur, selon la forme) est placé à ces coordonnées sur l'image correspondante dans la vidéo. Par ailleurs, une valeur d'angle de la main a été fixée pour chacune des positions par les experts du codage. Là aussi, nous n'avons introduit aucune variabilité : pour une position donnée, la main se place toujours avec le même angle. De même pour la taille de la main qui est proportionnelle à la hauteur du visage. Enfin, la trajectoire suivie par la main entre deux positions suit une ligne droite, à vitesse constante. Des analyses devront être conduites sur le corpus codé afin de rendre ce mouvement plus naturel.

3.4 Comment représenter le codage dans la vidéo ?

Une fois que le système a pu prédire quelle clé doit être codée, à quel moment et à quel endroit, le PoC peut augmenter la vidéo avec la main codeuse. Contrairement aux systèmes proposés par le MIT, pour le PoC, nous avons choisi d'utiliser des photos (figure 2), plutôt que des représentations imagées. Nous avons utilisé les fonctions de floutage (blur) et de transparence (fade in/fade out) pour indiquer respectivement les transitions de position et de forme.

4 Évaluations

4.1 Protocole

Pour évaluer la pertinence de la preuve de concept, nous avons mis en place un protocole d'évaluation en créant des séries de vidéos à décoder, *sans audio*. Le premier auteur de cet article a été filmé en lisant 4 sessions du corpus CLeLfPC, sans coder. Nous avons ensuite annoté le corpus avec SPPAS, en suivant le processus proposé dans la figure 1 : détection automatique des IPUs, transcription manuelle, segmentation automatique en phonèmes, détections des points du visage. Le PoC a ensuite généré les annotations et les vidéos codées automatiquement. Les vidéos des 4 sessions enregistrées ont été divisées en 4 séries différentes pour former un ensemble de 16 expériences avec des vidéos différentes. Chacune des 16 expériences se compose de 34 vidéos codées dont 10 servent de contrôle et 24 de test, avec un recouvrement des vidéos contrôle/test sur différentes expériences. Chaque vidéo ne contient qu'un mot ou une expression, que le participant ne peut visionner qu'une seule fois. Parmi les vidéos de contrôle, 8 ont été extraites du corpus CLeLfPC, codées par des codeurs professionnels. Les deux autres ont été codées automatiquement par le PoC et avaient été validées par des experts comme étant correctes.

Durant le stage annuel de l'Association pour la Langue française Parlée Complétée (ALPC), *des personnes sourdes connaissant le code se sont portées volontaires* pour décoder les vidéos. Pour participer, un texte de consentement devait être approuvé. Les expériences étaient anonymes, aucune donnée personnelle n'a été recueillie. Une vidéo qui décrit le protocole a été présentée à chaque participant pour s'assurer que tous ont reçu la même information. Durant l'expérience, pour chacune des 34 vidéos codées, le participant devant remplir les 3 champs suivants d'un formulaire :

- `J'ai décodé : ..............................`
- `J'ai correctement décodé:` *non ... peut-être ... oui* (barre de progression)
- `J'ai un commentaire sur cette vidéo (optionnel)`

4.2 Résultats quantitatifs

Les évaluations ont été réalisées avec l'outil Sclite, un programme inclut dans SCTK, le *Nist Scoring Toolkit*, habituellement utilisé pour estimer les résultats des systèmes de reconnaissance automatique de la parole. Cet outil permet de comparer des phrases de référence - les phrases à trouver, avec les phrases dites hypothèses - les phrases produites par le système automatisé. Il utilise un algorithme permettant d'estimer le pourcentage de mots correctement reconnus, ainsi que le taux de mots substitués, supprimés et insérés dans l'hypothèse. Parmi les 19 participants, 14 ont été sélectionnés après vérification du taux de décodage des vidéos de contrôle. Effectivement, nous avons estimé que si un participant n'est pas en mesure de décoder les mots du contrôle au moins à hauteur de 40%, il n'est pas suffisamment qualifié pour évaluer notre système. La table 1 résume les scores, obtenus sur les 14 réponses sélectionnées, dans les deux conditions (vidéos contrôles et vidéos du PoC). Les scores de décodage avec un codeur professionnel sont nettement meilleurs que ceux obtenus avec le PoC. Ils correspondent en fait au taux maximal qu'il est possible d'obtenir par les participants, dans la condition de test réalisée. Les résultats de décodage des vidéos lors du codage automatique avec le PoC s'en approchent et sont très prometteurs ; ils sont suffisamment corrects pour valider la méthodologie proposée.

	# Mots	Correct	Substitution	Supression	Insertion
contrôle	356	77,8 %	15,4 %	6,7 %	3,9 %
PoC	828	67,5 %	23,3 %	9,2 %	6,3 %

TABLE 1 – Taux d'erreur de décodage des mots

4.3 Résultats qualitatifs

Nous avons analysé les commentaires des participants sur les différentes vidéos. Dans quelques cas, des erreurs de clés ont été soulevées ; elles sont dues, soit à une erreur de conversion graphème-phonème qui a amenée à un mauvais choix de clé, soit à l'accent. D'autres commentaires indiquent que la clé est trop "rapide", ce qui sous-entend que le PoC a sur-estimé le temps de transition et donc sous-estimé le temps d'exposition. Les autres commentaires portent sur l'aspect de la main dans la vidéo : trop transparente et trop floue. En revanche, aucun commentaire n'a porté sur le côté non-naturel de la trajectoire de la main et son angle constant. Toutes ces informations permettront d'établir des priorités sur les actions à mener lors de l'élaboration du système. Ci-après, se trouvent quelques uns des commentaires :

- difficulté à savoir si c'est "à six" ou "assis"
- dans la vidéo, le "è" final de "sorbet", est représenté par la clé du "é"
- dernière clé trop rapide
- main mal positionnée sur la pommette
- la main est un peu trop transparente
- main pas assez claire : trop de flou entre deux clés

Enfin, nous avons recueilli les impressions des participants après leur passage de l'expérience et avons obtenu un retour très positif. Il semble que le système, lorsqu'il sera en phase finale, serait susceptible de trouver sa place dans la communauté. Enfin, le manque de ressources numériques pour la LfPC (vidéos codées notamment) a été mentionné par presque tous les participants.

5 Conclusion et perspectives

En combinaison avec les mouvements labiaux, le « Cued Speech » rend les phonèmes d'une langue parlée visuellement différents les uns des autres. Les clés du codage sont positionnées autour du visage, près des lèvres, ce qui facilite le suivi simultané des mouvements des lèvres et des mouvements de la main codeuse. Cet article a décrit une preuve de concept de génération automatique du codage en Langue française Parlée Complétée. La preuve de concept proposée est en mesure de prédire 1/ la clé à coder (position et forme de la main), 2/ les moments pour l'afficher, en changer et la déplacer, 3/ l'emplacement de la main, et 4/ d'augmenter la vidéo avec une main artificielle à partir de ces informations. Ce PoC a été bien accueilli par la communauté concernée, et les évaluations quantitatives ont révélé son fort potentiel, ce qui permet de le considérer comme étant approuvé.

Dans un futur proche, nous développerons un système basé sur l'analyse des annotations du corpus CLeLfPC. Pour la question du "quoi", le système restera à base de règles de productions. Pour les questions de "quand" et "où" placer la clé, nous pensons utiliser des techniques d'apprentissage automatique avec des modèles prédictifs appris à partir des données annotées, afin de rendre le système plus efficace dans les différentes prédictions. Quant à la question du "comment", seule une collaboration avec les personnes concernées permettra de l'améliorer. Avec un tel système de codage automatique, toutes sortes de vidéos codées pourront être élaborées et diffusées pour tous les types d'utilisations. Dans le contexte du présent projet, des textes sélectionnés sur différents thèmes (Gala *et al.*, 2024) seront lus par un acteur, afin de collecter des vidéos. Une fois codées automatiquement, évaluées et sélectionnées, ces vidéos seront assemblées pour créer des capsules pédagogiques destinées au grand public, aux débutants apprenant le code et aux enfants sourds.

6 Reproductibilité

Toutes les données et tous les codes sources mentionnés dans ce document respectent les principes de la science ouverte. Le code source de la preuve de concept est déposé sous les termes de la licence libre GNU Affero General Public License v3. Il fait partie du logiciel SPPAS. Les codes sources pour réaliser les expérimentations décrites dans cet article sont sous la licence GNU AGPL v3 et les données utilisées sont soumis aux termes des licences ODbL (Open Database License 1.0) et CC-By-NC-4.0. Nous avons utilisé SPPAS 4.11 https://sppas.org/, et SCTK 2.4.12 https://github.com/usnistgov/SCTK.

7 Remerciements

Nous tenons à remercier tous les participants à l'expérience qui ont accepté de donner de leur temps pour nous aider, lors du stage annuel organisé par l'ALPC https://alpc.asso.fr. Nous remercions également l'ALPC de nous avoir offert l'opportunité d'assister au stage et de présenter notre travail.

Les recherches présentées dans cet article ont été réalisées dans le cadre d'un projet financé par la FIRAH sous la référence APa2022_022 https://auto-cuedspeech.org/, en collaboration avec les associations Datha https://datha.io/ et AISAC https://www.academieinternationale.org/.

Références

ABOUTABIT N. (2007). *Reconnaissance de la Langue Française Parlée Complété (LPC) : décodage phonétique des gestes main-lèvres.* Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

ATTINA V. (2005). *La Langue Française Parlée Complétée : Production et Perception*. Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

BIGI B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, **111–112**, 54–69.

BIGI B., ZIMMERMANN M. & ANDRÉ C. (2022). Clelfpc : a large open multi-speaker corpus of french cued speech. In *Proceedings of The 13th Language Resources and Evaluation Conference*, p. 987–994, Marseille, France : European Language Resources Association. https ://hal.archives-ouvertes.fr/hal-03794830.

BRATAKOS M. S. (1995). *The effect of imperfect cues on the reception of cued speech*. Thèse de doctorat, Massachusetts Institute of Technology.

BRATAKOS M. S., DUCHNOWSKI P. & BRAIDA L. D. (1998). Toward the automatic generation of cued speech. *Cued Speech Journal*, **6**, 1–37.

CATHIARD M.-A., ATTINA V. & ALLOATTI D. (2003). Labial anticipation behavior during speech with and without cued speech. In *Proceedings of the 15th International Congress of Phonetic Sciences*, p. 1935–1938, Barcelona, Spain.

CORNETT R. O. (1967). Cued speech. *American annals of the deaf*, p. 3–13.

CORNETT R. O., BEADLES R. & WILSON B. (1977). Automatic cued speech. In *Research Conference on Speech Processing Aids for the Deaf*, p. 224–239, Gallaudet College (USA).

DUCHNOWSKI P., BRAIDA L. D., LUM D., SEXTON M., KRAUSE J. & BANTHIA S. (1998). Automatic generation of cued speech for the deaf : status and outlook. In *International Conference on Auditory-Visual Speech Processing*, Sydney, Australia.

DUCHNOWSKI P., LUM D. S., KRAUSE J. C., SEXTON M. G., BRATAKOS M. S. & BRAIDA L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE transactions on biomedical engineering*, **47**(4), 487–496.

GALA N., BIGI B. & BAUER M. (2024). Automatically estimating textual and phonemic complexity for cued speech : How to see the sounds from french texts. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Turin, Italy.

KAPLAN H. (1975). *The effects of cued speech on the speech-reading ability of the deaf.* Thèse de doctorat, ProQuest Information & Learning.

LEYBAERT J., COLIN C. & HAGE C. (2010). *Cued speech and cochlear implants*, p. 107–125.

NEEF N. A. & IWATA B. A. (1985). The development of generative lipreading skills in deaf persons using cued speech training. *Analysis and intervention in developmental disabilities*, **5**(4), 289–305.

SANKAR S., BEAUTEMPS D., ELISEI F., PERROTIN O. & HUEBER T. (2023). Investigating the dynamics of hand and lips in French Cued Speech using attention mechanisms and CTC-based decoding. In *Interspeech 2023 - 24th Annual Conference of the International Speech Communication Association*, Dublin, Ireland.

SEXTON M. G. (1997). *A video display system for an automatic cue generator*. Thèse de doctorat, Massachusetts Institute of Technology.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Rôle de l'activité laryngale dans la production des consonnes d'arrière en arabe levantin

Jalal Al-Tamimi[1]
(1) Université Paris Cité, CNRS, Laboratoire de linguistique formelle, UMR 7110, 5 Rue Thomas Mann, 75013, Paris, France
jalal.al-tamimi@u-paris.fr

RÉSUMÉ

Cette étude examine le rôle de l'activité laryngale dans la production des consonnes d'arrière en arabe levantin. 26 mesures incluant la hauteur du larynx (HL), le contact de glotte (quotient fermé ; QF) et la pente spectrale (PS) ont été obtenues de données d'éléctroglottographies et d'acoustiques synchronisées. À partir des classifications via des forêts aléatoires (Random Forests), sept mesures ont été identifiées comme les plus importantes pour discriminer entre les six classes. Ensuite, une modélisation via des Régressions Additives à Effets-Mixtes montre que les consonnes pharyngales sont associées à ⇑HL, ⇑QF et ⇑PS, résultant d'une différence de la saillance spectrale causée par une constriction épilaryngale. Les consonnes pharyngalisées induisent des traits ⇓HL, ⇓QF et ⇓PS causés par une fermeture abrupte de la glotte ; les consonnes uvulaires induisent des traits ⇑HL, ⇓QF et ⇓PS. Ces changements sont corrélés avec le trait [+Constricted Glottis] et suivent les prédictions du Laryngeal Articulator Model.

ABSTRACT

Role of the laryngeal activity during the production of back consonants in Levantine Arabic

This study examines the role of the laryngeal activity during the production of back consonants in Levantine Arabic. 26 measures including Larynx Height (LH), glottal contact (Closed Quotient ; CQ) and Spectral Tilt (ST) were obtained from synchronised electroglottographic and acoustic data. Using classifications via Random Forests, seven measures were identified as the most important to discriminate between the six classes. Next, modelling the data via Generalised Additive Mixed-Models revealed that pharyngeal consonants to be associated with an ⇑LH, a ⇑CQ and an ⇑ST, resulting from a difference in spectral saliency due to an epilaryngeal constriction. Pharyngealised consonants induced the features ⇓LH, ⇓CQ and ⇓ST caused by an abrupt closure of the glottis. Uvular consonants show the features ⇑LH, ⇓CQ and ⇓ST. These changes are correlated with the feature [+Constricted Glottis] and follow predictions of the Laryngeal Articulator Model.

MOTS-CLÉS : Epilarynx ; Electroglottographie ; Acoustique ; Qualité de voix ; Consonnes d'arrière ; Arabe levantin.

KEYWORDS: Epilarynx; Electroglottography; Acoustics; Voice quality; Back consonants; Levantine Arabic.

Actes des 35èmes Journées d'Études sur la Parole, pages 521–531.

1 Introduction

Les consonnes d'arrière, ou « gutturales », en arabe sont généralement considérées comme faisant partie d'une classe naturelle à cause d'alternance phonologique particulière et/ou une réalisation dans une zone oro-sensorielle commune tout au long de la cavité pharyngale (McCarthy, 1994; Sylak-Glassman, 2014). Cette classe de consonnes d'arrière est traditionnellement composée de consonnes pharyngales /ħ ʕ/ (ou épilaryngales /ʔ ʢ ʜ/) et uvulaires /χ ʁ q/ (McCarthy, 1994). Les consonnes pharyngalisées /tˤ dˤ ðˤ sˤ/ (Sylak-Glassman, 2014) et/ou glottales /h ʔ/ (Zawaydeh, 1999) sont également considérées comme en faisant partie, à cause d'une similarité du lieu (et non pas de degré) de constriction entre les consonnes pharyngalisées et pharyngales (Laufer & Baer, 1988) ou à cause de la similarité d'augmentation du premier formant (F1) associée aux consonnes d'arrière en comparaison avec les consonnes alvéolaires (Zawaydeh, 1999). En suivant les prédictions du Laryngeal Articulator Model (LAM, Esling *et al.*, 2019), les consonnes épilaryngales sont produites dans la partie inférieure de l'épilarynx, les consonnes uvulaires dans la partie supérieure et les consonnes pharyngales/pharyngalisées entre les deux. L'élévation du larynx a pour rôle de faciliter la constriction de l'épilarynx et est associée à une réalisation épilaryngale (Esling *et al.*, 2019).

Utilisant l'échographie de la langue, Al-Tamimi & Palo (2023) ont montré qu'il existe une similarité articulatoire entre les consonnes d'arrière concernant les parties de la langue impactées : une dépression de la partie antérieure de la langue est observée avec une élévation et rétraction variable du dos et de la racine de la langue, respectivement. De plus, en utilisant les prédictions des Modèles GAMMs (Régressions Généralisées Additives à Effets-Mixtes) qui s'étendent au-delà des zones observables à partir de l'échographie de la langue, Al-Tamimi & Palo (2023) ont montré que les consonnes d'arrière induisent une élévation variable du larynx. En effet, l'activité laryngale est souvent négligée dans la littérature décrivant la réalisation des consonnes d'arrière. Néanmoins, quelques études ont examiné le rôle du larynx en arabe. Heselwood (2007) a montré l'existence d'une voix craquée variable quantifiée via la mesure $H1$-$H2$ [1] dans les réalisations variables de /ʕ/ dans les dialectes arabes. Al-Tamimi & Heselwood (2011) ont montré une position plus élevée du larynx lors de la production des consonnes pharyngalisées en utilisant la vidéofluoroscopie. Heselwood & Al-Tamimi (2011) ont également montré une élévation du larynx et une voix craquée dans les consonnes pharyngales en utilisant la vidéofluoroscopie et les analyses acoustiques de $H1$-$H2$. Enfin et plus récemment, Al-Tamimi (2015, 2017) a montré une baisse générale de la pente spectrale comme corrélat secondaire de la voix tendue associée aux consonnes pharyngalisées en arabe jordanien et en arabe marocain. Les résultats de ces études montrent que l'activité laryngale peut être utilisée afin de valider les prédictions du LAM. Comme indiqué ci-dessus, le LAM prédit qu'une constriction extrême de l'épilarynx entraîne une élévation du larynx, induisant une voix tendue, craquée et/ou laryngalisée, associée à une rétraction importante du dos et de la racine de la langue (Esling *et al.*, 2019). Ces changements doivent être vue comme combinés et non distincts l'un de l'autre qui d'après Sylak-Glassman (2014) contribuent à la caractérisation de la classe des consonnes d'arrière dans la grande majorité des langues.

Ces observations ont motivé l'étude que nous avons menée. Notre but ici est d'examiner le rôle de l'activité laryngale dans la production des consonnes d'arrière en utilisant une combinaison de mesures articulatoires obtenues à partir de l'éléctroglottographie (EGG) et de mesures acoustiques de la qualité de voix. Cette combinaison nous permettra d'évaluer la complémentarité entre corrélats articulatoires et acoustiques pour quantifier les similarités et différences entre ces consonnes d'arrière.

1. $H1$ = premier harmonique ; $H2$ = deuxième harmonique

2 Corrélats articulatoires et acoustiques de qualité de voix

2.1 Corrélats articulatoires

Kuang & Keating (2014) ont décrit les corrélats articulatoires obtenus de l'EGG de la voix tendue (en comparaison avec une voix relâchée). ⇑QF (Quotient Fermé, % contact) causée par une augmentation du contact dans la glotte. ⇓PIC (amplitude du pic positif de la dérivée du signal EGG) et ⇓PDC (amplitude du pic négatif de la dérivée d'EGG) sont causées par une fermeture abrupte ou rapide de la glotte induisant une augmentation d'énergie dans les hautes fréquences (Michaud, 2004; Kuang & Keating, 2014). Finalement, ⇓SQ (quotient de vitesse d'ouverture/fermeture calculé comme le ratio de la phase de fermeture et la phase d'ouverture) est à observer dans la voix tendue causée par une phase de fermeture plus courte et une phase d'ouverture plus longue (Holmberg *et al.*, 1988; Kuang & Keating, 2014). Enfin, une HL (Hauteur du Larynx) variable est observée en fonction de la catégorie examinée. En suivant les prédictions du LAM, les consonnes pharyngales auront une ⇑HL, avec une position variable pour les consonnes pharyngalisées et uvulaires.

2.2 Corrélats acoustiques

Pour ce qui est des mesures acoustiques, Kuang & Keating (2014) et Al-Tamimi (2017) ont montré qu'une voix tendue, craquée et/ou laryngalisée conduit à un abaissement global de la pente spectrale, en suivant le modèle psychoacoustique de la qualité de voix (Garellek *et al.*, 2016; Kreiman *et al.*, 2021). Ces changements induisent un abaissement des mesures suivantes [2] : ⇓*H*1-*H*2, ⇓*H*1-*A*1, ⇓*H*1-*A*2, ⇓*H*2-*H*4, ⇓*H*4-*H*2kHz, ⇓*H*2kHz-*H*5kHz, ⇓HNR et ⇓SHR (Fulop *et al.*, 1998; Guion *et al.*, 2004; Aralova *et al.*, 2011; Kuang & Keating, 2014; Al-Tamimi, 2015; Garellek *et al.*, 2016; Al-Tamimi, 2017). De plus, un renforcement de la saillance spectrale entre F1 et F2, quantifié via ⇓*A*1-*A*2, est observé. Enfin, et comme rapporté précédemment dans Al-Tamimi (2015, 2017) pour les consonnes pharyngalisées en arabe jordanien et marocain, nous prédisons que ce type de voix induit une augmentation de l'énergie spectrale autour de F3 et au-delà, qui peut être causée par deux facteurs. Le premier peut résulter d'une fermeture abrupte de la glotte avec augmentation de l'énergie spectrale autour de F3 avec comme conséquence ⇓*H*1-*A*3, ⇓*A*1-*A*3 et ⇓*A*2-*A*3 (Hanson & Chuang, 1999; Hanson *et al.*, 2001; Al-Tamimi, 2015, 2017). Le second est dû à une qualité vocale renforcée causée par une constriction épilaryngale extrême, comme dans le chant (ex. opéra), avec un renforcement bien marqué de l'énergie spectrale autour de F3, F4 et F5 comme conséquences du « formant du chanteur ». Ce qui induit une saillance spectrale autour de F3, F4 et F5 associée à un changement de la saillance spectrale autour de F1 et F2, induisant ⇑*H*1-*A*3, ⇓*A*1-*A*3, ⇓*A*2-*A*3 (Titze & Story, 1997; Moisik & Esling, 2010; Story, 2019). Tous ces changements doivent être évalués en comparaison avec une voix modale, où les changements de pentes spectrales sont minimaux.

Les prédictions sont donc qu'un abaissement global de la pente spectrale dans les consonnes d'arrière est à observer bien qu'on s'attendrait à des différences liées au degré de constriction épilaryngale. On propose donc que les consonnes pharyngales soient associées à une baisse globale de la pente spectrale avec une augmentation de l'énergie dans les hautes fréquences causée par une constriction épilaryngale extrême. Les consonnes pharyngalisées causeront quant à elles une baisse globale de la pente spectrale avec une augmentation de l'énergie dans les hautes fréquences moins marquée que dans les consonnes

2. Avec *H*1, 2 ou 4 = Harmonique 1, 2 ou 4 ; *H*2kHz ou 5kHz = Harmonique proche des 2 ou 5 kHz ; *A*1-3 = Harmonique proche de F1, F2 ou F3 ; HNR = Ratio d'énergie dans les harmoniques et le bruit ; SHR = ratio d'énergie subharmonique

pharyngales causé par une constriction laryngale. Enfin, les consonnes uvulaires montreront des similarités avec les deux autres classes avec des conséquences intermédiaires. Cette étude examine la complémentarité entre corrélats articulatoires et acoustiques permettant la discrimination entre les consonnes d'arrière en comparaison avec les consonnes alvéolaires vélaires et glottales. Nous montrerons s'il existe une similarité entre les consonnes d'arrière en quantifiant la gradience de leurs impacts sur les mesures articulatoires et acoustiques.

3 Corpus et méthode

3.1 Participants et enregistrements

Dix locuteurs urbains parlant l'arabe levantin [3] (5 hommes et 5 femmes), âgés entre 25 et 45 ans (moyenne 31,8 ; écart type = 6,9) ont été recrutés. Trois types de données ont été enregistrés simultanément : échographie de la langue (voir Al-Tamimi & Palo, 2023), éléctroglottographie (système à deux-canaux EG2-PCX2 de Glottal Entreprise avec hauteur et contact) et audio (avec un microphone Roland Pro connecté à une carte son Roland Quad-Capture, taux d'échantillonnage à 44.1 kHz, et quantification 16-bits avec distance d'environ 15 cm de la bouche du participant). Les trois systèmes étaient reliés à un système de synchronisation et d'alignement automatique des signaux (pour plus de détails, voir Wrench & Scobbie, 2008). Après l'installation des électrodes et du casque de stabilisation de la sonde de l'échographie de la langue, nous avons vérifié que les capteurs d'EGG étaient bien positionnés sur le larynx de chaque participant et que le traçage automatique du second canal de l'EGG, qui capte les mouvements verticaux du larynx, était au centre des 15 témoins lumineux.

3.2 Corpus et segmentation

Les participants ont produit une liste de mots réels (près de 75% du corpus) et des pseudo-mots (25% du corpus) répétés trois fois, dans la séquence /ˈʔVːˈCVː/, où Vː = les voyelles /iː aː uː/ symétriques ; C = toutes les consonnes possibles en arabe levantin et d'autres variétés arabes = /b t d m n r f θ ð s z ʃ ʒ l w j k g x ɣ q tˤ dˤ ðˤ sˤ zˤ lˤ ħ ʕ ʔ h/ (avec un nombre maximal théorique de 2790 items : 31 C * 3 V * 3 répétitions * 10 locuteurs). Ensuite, 21 consonnes ont été retenues pour la suite d'analyse et ont été divisées en six classes, avec un total de 1940 items (21 C * 3 V * 3 répétitions * 10 locuteurs + répétitions additionnelles) : **Alvéolaires** ⇒ /t d ð s z l/, **Vélaires** ⇒ /k g x ɣ/, **Uvulaire** ⇒ /q/, **Pharyngalisées** ⇒ /tˤ dˤ ðˤ sˤ zˤ lˤ/, **Pharyngales** ⇒ /ħ ʕ/ et **Glottales** ⇒ /h ʔ/.

Les données ont été translittérées manuellement utilisant la convention ATR (Al-Tamimi *et al.*, 2022) et alignées avec le système `MAUS` (Schiel, 2015) implémenté dans `PraatAlign` (Lubbers & Torreira, 2013). La segmentation automatique a été corrigée manuellement afin de prévenir les erreurs potentielles (suivant les critères définis dans Al-Tamimi, 2017 et Al-Tamimi & Khattab, 2018).

3.3 Extraction des mesures

Les mesures articulatoires et acoustiques ont été obtenues en utilisant `EGGWorks` (Tehrani, 2020) pour l'EGG et `VoiceSauce` (Shue *et al.*, 2011) pour les mesures de la pente spectrale avec les

3. L'arabe levantin regroupe les parlés urbains de la Palestine, Jordanie, Syrie et du Liban. Même si les participants viennent de différentes régions, ils partagent tous des traits spécifiques aux parlers syro-libanais, à savoir la réalisation de /q/ en Arabe Standard Moderne (ASM) comme /ʔ/, de /θ ð ðˤ/ en ASM comme /t d dˤ/ ou /s z zˤ/, de /ʤ/ en ASM comme /ʒ/ et de /χ ʁ/ en ASM comme /x ɣ/ (Embarki, 2008)

paramètres spécifiques à `Praat` (Boersma & Weenink, 2020). Les seuils de $f0$ et de formants ont été adaptés à chaque locuteur (suivant Al-Tamimi, 2017, Al-Tamimi & Khattab, 2018 et Al-Tamimi, 2022) et les mesures de $f0$ et des quatre premiers formants ont été manuellement vérifiées. Toutes les mesures décrites ci-dessous ont été initialement calculées avec un pas de déplacement d'1 ms. Ensuite, 11 intervalles moyennés sur la durée de chaque segment de la séquence VCV ont été extraits.

Pour les mesures articulatoires, un total de sept mesures a été calculé. La HL a été obtenue du second canal de l'EGG sans normalisation. Ensuite, nous avons calculé les mesures de contact telles que le QF en suivant la méthode hybride (Howard, 1995), le PIC (Michaud, 2004), le PDC, la vitesse d'ouverture et de fermeture, ainsi que leur ratio (voir Kuang & Keating, 2014).

Pour les mesures acoustiques, un total de 19 mesures a été obtenu en suivant deux approches. Pour la première, nous avons calculé toutes les mesures décrites dans le modèle psychoacoustique de la qualité de voix (Garellek *et al.*, 2016; Kreiman *et al.*, 2021) comme : la pente spectrale normalisée suivant Iseli *et al.* (2007) sur diverses bandes d'harmoniques (ex. *H*1*-*H*2*, *H*2*-*H*4*, *H*4*-*H*2kHz*, *H*2kHz*-*H*5kHz*, *H*1*-*A*1*, *H*1*-*A*2* et *H*1*-*A*3*) ; les HNR sur plusieurs bandes ; SHR ; CPP ; soe [4]. Pour la seconde, nous avons calculé les trois mesures d'amplitude dans les hautes fréquences (ex. *A*1*-*A*2*, *A*1*-*A*3* et *A*2*-*A*3*) en suivant les prédictions avancées dans Al-Tamimi (2015, 2017).

3.4 Approche statistique

Afin d'évaluer la contribution des mesures articulatoires et acoustiques pour montrer les similarités entre les consonnes d'arrière et les différences avec les autres contextes, nous avons procédé à deux types d'analyse. Premièrement, une classification avec les forêts aléatoires (Random Forests) a permis d'évaluer le taux de succès de discrimination suivi de l'évaluation du poids relatif pour chaque variable. Pour ce faire, nous avons entrainé plusieurs forêts aléatoires sur 66.7% des données, avec les mesures articulatoires, acoustiques et leur combinaison afin d'évaluer la robustesse de l'approche. Ces entrainements ont été effectué sur la partie V1, C ou V2 (voyelle précédente, consonne médiane et voyelle suivante, respectivement). Nous avons utilisé deux types de réponses : les six classes (alvéolaires, vélaires, uvulaire, pharyngalisées, pharyngales et glottales), ou deux classes (gutturale = uvulaire, pharyngalisées et pharyngales vs non-gutturale = alvéolaires, vélaires et glottales). Nous avons utilisé l'approche de `TidyModels` (Kuhn & Wickham, 2020) et la librairie `ranger` (Wright & Ziegler, 2017). Chaque forêt a été entrainée avec 2000 arbres ; le nombre de variables que chaque forêt pouvait utiliser dans l'apprentissage représentait la racine carrée arrondie du nombre total de prédicteurs (3 pour l'EGG ; 4 pour l'acoustique ; 5 pour l'EGG+acoustique). Une validation croisée avec 10 plis a été effectuée afin d'obtenir un apprentissage robuste sur les données d'entrainement. A la fin, des prédictions sur les données de test (33.3%) ont été effectuées afin d'obtenir le taux de classification et le poids relatif de chaque variable via des tests de permutation (Strobl *et al.*, 2009).

Ensuite, à partir des poids relatifs identifiés dans la classification combinant les mesures articulatoires et acoustiques, nous avons choisi les trois prédicteurs les plus importants pour chacun des V1, C et V2. Au total, nous avons identifié six mesures partagées en plus de la HL. Ces sept mesures ont ensuite été modélisées avec des GAMMs via la librairie `mgcv` (Wood, 2017). Le but de ces modélisations est de quantifier la dynamique des trajectoires tout au long de la séquence VCV. Pour ce faire, nous avons utilisé le prédicteur comme variable réponse, et l'interaction entre contexte et voyelle (variables ordonnées) et le sexe du locuteur comme variables fixes. Deux trajectoires représentant les 11 intervalles et la position dans la syllabe (1, 2 et 3 pour V1, C, et V2) ont été lissées

4. HNR = Ratio d'énergie dans les harmoniques et dans le bruit ; CPP = amplitude du pic cepstral ; soe = Force d'excitation

avec une pénalisation en régression cubique, utilisant une interaction entre les deux trajectoires et avec ajustements par les variables fixes. Comme variables aléatoires, nous avons utilisé le locuteur et le mot ajustés par les deux trajectoires, ainsi que par l'interaction entre contextes et voyelles (pour le locuteur), et le sexe du participant (pour le mot). Nos modèles ont suivi l'approche maximale en régression à effets mixtes (Barr *et al.*, 2013) ; permettant une normalisation entre locuteurs, avec une approche similaire à celle utilisée dans Al-Tamimi & Palo (2023). Nous avons vérifié la structure de nos modèles avec la fonction `gam.check`. Ensuite, les prédictions de chaque modèle ont été obtenues via la fonction `predict.bam`, qui ont été utilisées pour générées les visualisations de la Figure 2 avec `ggplot2` (Wickham, 2009) de la suite `tidyverse` (Wickham *et al.*, 2019).

4 Résultats

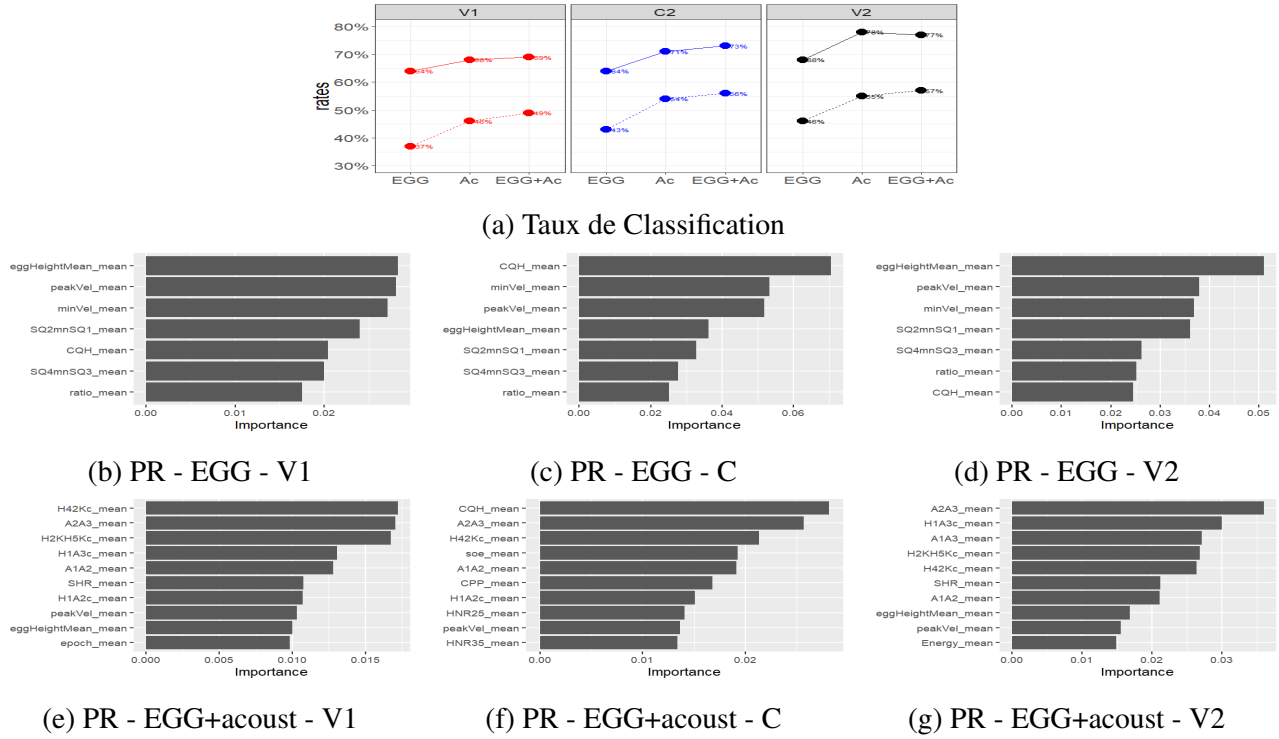
4.1 Classification avec les forêts aléatoires, Random Forests

Lorsque les deux classes étaient utilisées comme réponses (Figure 1a), les taux de classification variaient entre 64-78% avec les taux les plus élevés se situant dans la V2 en utilisant soit les mesures acoustiques soit EGG+acoustique. Ceci indique clairement une forme de coarticulation progressive vers la V2. Ces taux sont assez élevés mais ne permettent pas une discrimination parfaite des deux classes ; la qualité de voix quantifiée via des méthodes articulatoires et acoustiques joue un rôle secondaire dans la discrimination entre classes. Ce résultat est attendu étant donné que l'arabe n'a pas de contraste phonologique marqué par une différence de phonation. En évaluant la classification par classe, les taux baissent d'une façon spectaculaire avec des taux proches des 37-67%. Il est à noter que lorsque les mesures articulatoires étaient utilisées, les taux étaient les plus bas en comparaison avec les mesures acoustiques, indiquant plus de facilité pour les forêts aléatoires à identifier les schémas spécifiques à nos classes en s'appuyant sur ces dernières. Les Figures 1b, 1c et 1d montrent le poids relatif pour les mesures articulatoires [5]. La HL était le prédicteur le plus important dans la classification dans V1 et V2, mais le QF était le prédicteur le plus important pour C. Les Figures 1e, 1f et 1g montrent le poids relatif des dix mesures les plus importantes en combinant toutes les mesures articulatoires et acoustiques. Le QF reste le prédicteur le plus important pour C, tandis que les mesures acoustiques *A*1*-*A*3*, *A*2*-*A*3*, *H*1*-*A*3*, *H*4*-*H*2kHz*, *H*2kHz*-*H*5kHz* étaient parmi les prédicteurs les plus importants dans toutes les séquences pour discriminer les six classes. Toutes ces mesures sont indicatives de variations liées à la pente spectrale, avec des pentes plus raides indicatives d'une augmentation d'énergie dans les hautes fréquences causée soit par une fermeture abrupte de la glotte (Hanson *et al.*, 2001; Al-Tamimi, 2017) soit par une constriction épilaryngale induisant une qualité de voix renforcée (Titze & Story, 1997; Moisik & Esling, 2010; Story, 2019).

4.2 Modélisation avec les GAMMs

Figure 2 montrent les courbes moyennes prédites par nos modèles GAMMs par prédicteur en fonction des six classes par V1, C ou V2. Comme les résultats de *A*1*-*A*3* montrent des patterns comparables à ceux de *A*2*-*A*3*, nous présentons uniquement ceux de cette dernière. Les changements dynamiques de la qualité de voix au niveau articulatoire et acoustique suivent un pattern générique de ⇓ dans V1, ⇑/⇓ dans C et ⇑ dans V2. Plus spécifiquement, une HL variable est observée (Figure 2a) avec les

5. CQH = Quotient Fermé hybride ; minVel = PDC ; peakVel = PIC ; SQ2mnSQ1 = phase fermante ; SQ4mnSQ3 = phase ouvrante ; ratio = Quotient de vitesse ; eggHeightMean = Hauteur du Larynx



(a) Taux de Classification

(b) PR - EGG - V1

(c) PR - EGG - C

(d) PR - EGG - V2

(e) PR - EGG+acoust - V1

(f) PR - EGG+acoust - C

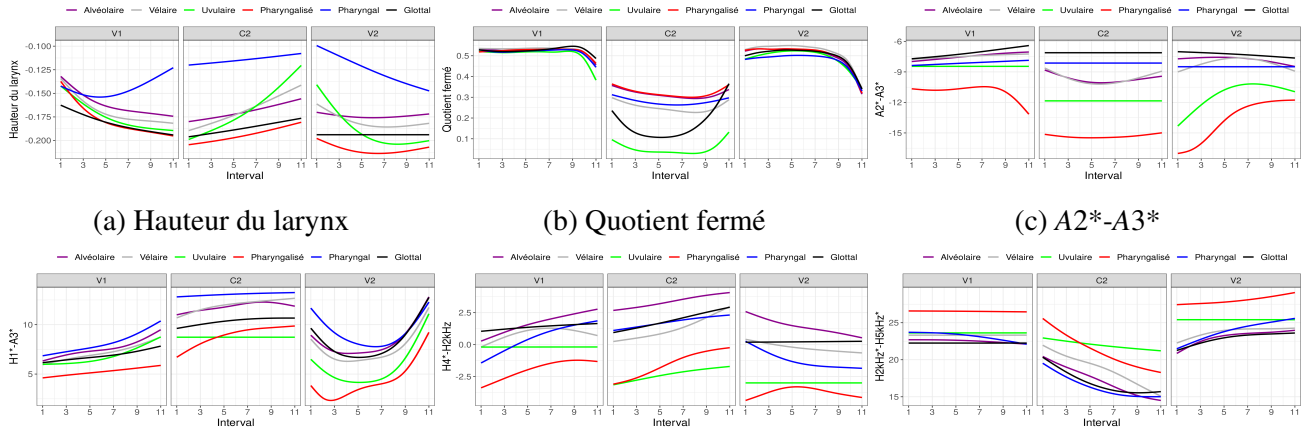
(g) PR - EGG+acoust - V2

FIGURE 1 – Taux de classification par forêt aléatoire sur les données de tests (a, ac = acoustique), le poids relatif (PR) des variables dans la tâche de classification des six classes combinant les mesures articulatoires (EGG) dans V1 (b), C (c) et V2 (d) et les dix mesures les plus importantes en combinant toutes les mesures articulatoires et acoustiques (EGG+acoust) dans V1 (e), C (f) et V2 (g).

consonnes pharyngales montrant un pattern ⇑HL tout au long de la séquence VCV, les consonnes pharyngalisées avec un pattern ⇓ dans V1, ⇑ dans C et ⇓ dans V2 ; la consonne uvulaire suit un pattern similaire mais avec ⇑HL dans la phase de relâchement (intervalle 9). Pour le QF (Figure 2b), les courbes semblent plus variées dans C, avec des changements mineurs dans V1 et V2. La consonne uvulaire montre le QF le plus bas, suivi des consonnes glottales et vélaires, puis pharyngales avec les consonnes alvéolaires et pharyngalisées ayant le QF le plus haut (autour de 0,35). Ces résultats corrèlent bien avec un contact moins élevé pour les consonnes uvulaires et pharyngales en comparaison avec les consonnes pharyngalisées. Figures 2c, 2d et 2e montrent des résultats très similaires pour les consonnes uvulaires et pharyngalisées : une réduction globale de *A*2*-*A*3*, *H*1*-*A*3* et de *H*4*-*H*2kHz* est observée, plus marquée dans les consonnes pharyngalisées. Ceci est indicatif d'un changement abrupt de la pente spectrale autour des 2kHz et 3kHz qui peut être corrélé avec une glotte plus serrée ; pour les consonnes pharyngales, l'augmentation dans *H*1*-*A*3* et la baisse dans les autres mesures indiquent un épilarynx plus serré, suivant les prédictions avancées ci-dessus. Les résultats de *H*2kHz*-*H*5kHz* (Figure 2f) montrent un effet inverse avec un changement abrupt de la pente spectrale dans les consonnes pharyngales causé par une augmentation d'énergie dans les hautes fréquences causée par une constriction épilaryngale (avec ⇓*H*2kHz*-*H*5kHz*) avec une baisse d'énergie dans les consonnes uvulaires et pharyngalisés causée par une constriction glottale.

5 Discussion et conclusions

Cette étude a démontré que les consonnes d'arrière en arabe levantin induisent une élévation et une constriction du larynx variables. Les consonnes pharyngales sont produites avec une élévation



(a) Hauteur du larynx (b) Quotient fermé (c) *A*2*-*A*3*

(d) *H*1*-*A*3* (e) *H*4*-*H*2kHz* (f) *H*2kHz*-*H*5kHz*

FIGURE 2 – Résultats des modélisations avec les GAMMs, avec les courbes prédites par V1, C ou V2 pour la HL (a), le QF (b), *A*2*-*A*3* (c), *H*1*-*A*3* (d), *H*4*-*H*2kHz* (e) et *H*2kHz*-*H*5kHz* (f).

maximale du larynx tout au long de la séquence VCV causée par une constriction épilaryngale, suivant les prédictions du LAM. Les consonnes pharyngalisées sont produites avec un abaissement du larynx, qui augmente vers la période du relâchement de la consonne, mais avec une fermeture abrupte de la glotte. La consonne uvulaire partage un larynx plus élevé avec les consonnes pharyngales mais une fermeture abrupte de la glotte avec les consonnes pharyngalisées. Les résultats articulatoires de HL et de QF et acoustiques liés à l'augmentation de l'énergie spectrale dans les hautes fréquences dans les consonnes pharyngalisées va dans le sens d'un contraste marqué par une différence de voix tendue dans ces dernières en comparaison avec une voix relâchée/modale dans les consonnes alvéolaires comparable aux résultats rapportés dans (Kuang & Keating, 2014) de contraste phonologique entre voix tendues et relâchées. La complémentarité entre corrélats articulatoires et acoustiques dans cette étude confirme que les mesures de la pente spectrale, qui sont validées comme permettant de classifier les qualités de voix (Garellek *et al.*, 2016; Kreiman *et al.*, 2021) jouent un rôle important dans la caractérisation des différences phonatoires entre les consonnes d'arrière de l'arabe. Nos résultats articulatoires et acoustiques fournissent une évidence empirique que les consonnes d'arrière en arabe partagent le trait phonologique [+Constricted Glottis] comme trait secondaire actif avec une variation graduelle bien marquée permettant de les différencier. Ces résultats suivent les prédictions du LAM : les consonnes d'arrière en arabe sont produites avec une constriction graduelle de l'épilarynx qui induit une élévation et une constriction variables de la glotte avec une rétraction variable du dos et de la racine de la langue. Cette complémentarité permet ainsi de conclure que les trois membres de cette classe de consonnes d'arrière partagent des traits phonologiques similaires permettant de les unir.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'IdEx Université Paris Cité (ANR-18-IDEX-0001) au titre du Labex Empirical Foundations of Linguistics - EFL. Il a également bénéficié d'un financement de la British Academy/Leverhulme small research grant, Royaume Uni (SG160181 ; 2017-2019) et d'un financement Leverhulme International Academic Fellowship, Royaume Uni (IAF-2018-016). Ce travail a bénéficié du support pour l'utilisation du High Power Computing (HPC) de l'université de Newcastle, au Royaume Uni, du CNRS/TGIR HUMA-NUM, IN2P3 et du GENCI-IDRIS, France (2022-AD010613733).

Références

AL-TAMIMI F. & HESELWOOD B. (2011). Nasoendoscopic, videofluoroscopic and acoustic study of plain and emphatic coronals in Jordanian Arabic. In B. HESELWOOD & Z. HASSAN, Éds., *Instrumental Studies in Arabic Phonetics*, p. 165–191. John Benjamins. DOI : 10.1075/cilt.319.

AL-TAMIMI J. (2015). Spectral tilt as an acoustic correlate to pharyngealisation in Jordanian and Moroccan Arabic (Article : 0436). In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*.

AL-TAMIMI J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic : Implications for formal representations. *Laboratory Phonology : Journal of the Association for Laboratory Phonology*, **8**(1), 1–40. DOI : 10.5334/labphon.19.

AL-TAMIMI J. (2022). Praat-f0-Accurate-Estimation. https ://jalalal-tamimi.github.io/Praat-f0-Accurate-Estimation/.

AL-TAMIMI J. & KHATTAB G. (2018). Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops. *Journal of Phonetics*, **71**, 306–325. DOI : 10.1016/j.wocn.2018.09.010.

AL-TAMIMI J. & PALO P. (2023). Dynamics of the tongue contour in the production of guttural consonants in Levantine Arabic. In *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, p. 2095–2099.

AL-TAMIMI J., SCHIEL F., KHATTAB G., SOKHEY N., AMAZOUZ D., DALLAK A. & MOUSSA H. (2022). A Romanization System and WebMAUS Aligner for Arabic Varieties. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, p. 7269–7276.

ARALOVA N., GRAWUNDER S. & WINTER B. (2011). The Acoustic Correlates of Tongue Root Vowel Harmony in Even (Tungusic). In *Proceedings of the 17th International Congress of Phonetic Sciences, ICPhS*, p. 240–243.

BARR D. J., LEVY R., SCHEEPERS C. & TILY H. J. (2013). Random effects structure for confirmatory hypothesis testing : Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. DOI : 10.1016/j.jml.2012.11.001.

BOERSMA P. & WEENINK D. (2020). *Praat Software*. University of Amsterdam.

EMBARKI M. (2008). Les dialectes arabes modernes : État et nouvelles perspectives pour la classification géo-sociologique. *Arabica*, **55**(5/6), 583–604. DOI : 10.1163/157005808X364616.

ESLING J., MOISIK S., BENNER A. & CREVIER-BUCHMAN L. (2019). *Voice Quality : The Laryngeal Articulator Model*. Cambridge University Press. DOI : 10.1017/9781108696555.

FULOP S. A., KARI E. & LADEFOGED P. (1998). An Acoustic Study of the Tongue Root Contrast in Degema Vowels. *Phonetica*, **55**(1-2), 80–98. DOI : 10.1159/000028425.

GARELLEK M., SAMLAN R., GERRATT B. R. & KREIMAN J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, **139**(3), 1404–1410. DOI : 10.1121/1.4944474.

GUION S. G., POST M. W. & PAYNE D. L. (2004). Phonetic correlates of tongue root vowel contrasts in Maa. *Journal of Phonetics*, **32**(4), 517–542. DOI : 10.1016/j.wocn.2004.04.002.

HANSON H. M. & CHUANG E. S. (1999). Glottal characteristics of male speakers : Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, **106**(2), 1064–1077. DOI : 10.1121/1.427116.

HANSON H. M., STEVENS K. N., KUO H.-K. J., CHEN M. Y. & SLIFKA J. (2001). Towards models of phonation. *Journal of Phonetics*, **29**(4), 451–480. DOI : 10.1006/jpho.2001.0146.

HESELWOOD B. (2007). The 'tight approximant' variant of the Arabic 'ayn. *Journal of the International Phonetic Association*, **37**(1), 1. DOI : 10.1017/S0025100306002787.

HESELWOOD B. & AL-TAMIMI F. (2011). A study of the laryngeal and pharyngeal consonants in Jordanian Arabic using nasoendoscopy, videofluoroscopy and spectrography. In B. HESELWOOD & Z. HASSAN, Éds., *Instrumental Studies in Arabic Phonetics*, p. 101–128. John Benjamins. DOI : 10.1075/cilt.319.

HOLMBERG E. B., HILLMAN R. E. & PERKELL J. S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the Acoustical Society of America*, **84**(2), 511–529. DOI : 10.1121/1.396829.

HOWARD D. M. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. *Journal of Voice*, **9**(2), 163–172. DOI : 10.1016/S0892-1997(05)80250-4.

ISELI M., SHUE Y.-L. & ALWAN A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America*, **121**(4), 2283–2295. DOI : 10.1121/1.2697522.

KREIMAN J., LEE Y., GARELLEK M., SAMLAN R. & GERRATT B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, **149**(1), 457–465. DOI : 10.1121/10.0003331.

KUANG J. & KEATING P. (2014). Vocal fold vibratory patterns in tense versus lax phonation contrasts. *The Journal of the Acoustical Society of America*, **136**(5), 2784–2797. DOI : 10.1121/1.4896462.

KUHN M. & WICKHAM H. (2020). *tidymodels : Easily Install and Load the 'Tidymodels' Packages*.

LAUFER A. & BAER T. (1988). The emphatic and pharyngeal sounds in Hebrew and in Arabic. *Language and Speech*, **31**, 181–205. DOI : 10.1177/002383098803100205.

LUBBERS M. & TORREIRA F. (2013). Praatalign : An interactive Praat plug-in for performing phonetic forced alignment : https ://github.com/dopefishh/praatalign. Version 2.0a.

MCCARTHY J. (1994). The phonetics and phonology of Semitic pharyngeals. In P. KEATING, Éd., *Phonological Structure and Phonetic Form*, p. 191–233. Cambridge University Press. DOI : 10.1017/CBO9780511659461.012.

MICHAUD A. (2004). A measurement from electroglottography : DECPA, and its application in prosody. In *Speech Prosody 2004*, p. 633–636.

MOISIK S. & ESLING J. (2010). Examining the Acoustic Contributions of the Epilaryngeal Tube to the Voice Source and Vocal Tract Resonance. *Canadian Acoustics*, **38**(3), 138–139.

SCHIEL F. (2015). A statistical model for predicting pronunciation (Article : 0195). In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*.

SHUE Y.-L., KEATING P., VICENIK C., YU K. & YEN-LIANG K. P. V. C. Y. K. S. (2011). VoiceSauce : A program for voice analysis. In *Proceedings of the 17th International Congress of Phonetic Sciences, ICPhS*, p. 1846–1849.

STORY B. (2019). The Vocal Tract in Singing. In *The Oxford Handbook of Singing*, p. 144–166. Oxford University Press. DOI : 10.1093/oxfordhb/9780199660773.013.012.

STROBL C., MALLEY J. & TUTZ G. (2009). An introduction to recursive partitioning : Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, **14**(4), 323–348. DOI : 10.1037/a0016973.

SYLAK-GLASSMAN J. (2014). *Deriving Natural Classes : The Phonology and Typology of Post-Velar Consonants*. Thèse de doctorat, University of California, Berkeley.

TEHRANI H. (2020). http ://www.appsobabble.com/functions/eggworks.aspx (accessible 18/02/2020).

TITZE I. R. & STORY B. H. (1997). Acoustic interactions of the voice source with the lower vocal tract. *The Journal of the Acoustical Society of America*, **101**(4), 2234–2243. DOI : 10.1121/1.418246.

WICKHAM H. (2009). *Ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York.

WICKHAM H., AVERICK M., BRYAN J., CHANG W., MCGOWAN L. D., FRANÇOIS R., GROLEMUND G., HAYES A., HENRY L., HESTER J., KUHN M., PEDERSEN T. L., MILLER E., BACHE S. M., MÜLLER K., OOMS J., ROBINSON D., SEIDEL D. P., SPINU V., TAKAHASHI K., VAUGHAN D., WILKE C., WOO K. & YUTANI H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686. DOI : 10.21105/joss.01686.

WOOD S. N. (2017). *Generalized Additive Models : An Introduction with R*. CRC Press/Taylor & Francis Group. DOI : 10.1201/9781315370279.

WRENCH A. A. & SCOBBIE J. M. (2008). High-speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging : Comparison of Front and Back Lingual Gesture Location and Relative Timing. In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, p. 57–60.

WRIGHT M. N. & ZIEGLER A. (2017). Ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**(1). DOI : 10.18637/jss.v077.i01.

ZAWAYDEH B. (1999). *The Phonetics and Phonology of Gutturals in Arabic*. Thèse de doctorat, Bloomington, IN : Indiana University.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Sandhi tonal en shanghaïen : une étude acoustique des contours dissyllabiques chez des locuteurs jeunes

Yu Chen[1], Nathalie Vallée[1], Thi-Thuy-Hien Tran[1], Silvain Gerber [1]

(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
yu.chen1@etu.univ-grenoble-alpes.fr, nathalie.vallee@gipsa-lab.grenoble-inp.fr,
thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr, silvain.gerber@gipsa-lab.grenoble-inp.fr

RESUME

Le shanghaïen possède deux types de sandhi tonal : *Left Dominant Sandhi* (LDS) dans les composés sémantiques de type syntagme nominal (SN) et *Right Dominant Sandhi* (RDS) dans des phrases prosodiques de type syntagme verbal (SV). Cette étude examine les caractéristiques acoustiques du contour tonal dans des SN et SV dissyllabiques chez trois locutrices jeunes. Nos résultats montrent que les tons des SN subissent des changements phonologiques relevant du LDS, alors que les SV sont plutôt soumis aux effets phonétiques de la coarticulation tonale plutôt qu'au RDS. L'absence de différences significatives entre les SN et les SV ne permet pas de généraliser une distinction entre eux uniquement sur la base des réalisations tonales. Cette étude exploratoire ouvre des perspectives pour de futurs travaux intergénérationnels sur les productions tonales et la perception du sandhi tonal, en étendant le corpus à différentes positions au sein de la phrase et différentes classes d'âge.

ABSTRACT

Shanghainese has two types of tonal sandhi: Left Dominant Sandhi (LDS) in semantic compounds such as nominal phrase (NP) and Right Dominant Sandhi (RDS) in prosodic phrase such as verbal phrase (VP). This study examines acoustic characteristics of tonal contours in dissyllabic NPs and VPs in Shanghainese in three young female speakers. Our results show that the tones of NPs undergo phonological changes related to LDS, while VPs are rather subject to the phonetic effects of tonal coarticulation rather than RDS. The absence of significant differences between the contours of NPs and those of VPs does not allow for a generalization of a distinction between them solely based on tonal realizations. This exploratory study opens perspectives for future intergenerational work on tonal productions and on the perception of tonal sandhi, by extending the corpus to different positions within the sentence and to different age groups.

MOTS-CLÉS : langues wu ; coarticulation tonale ; propagation ; directionalité ; contraste
KEYWORDS : Wu languages; tone coarticulation; spreading; directionality; contrast

1 Introduction

Comme la plupart des langues chinoises, les langues wu (wuu, code ISO 639-3, groupe linguistique sino-tibétain) présentent des phénomènes complexes de sandhi tonal, probablement en raison de la complexité de leur système tonal, où les unités de base sont définies par des combinaisons complexes

de contours, de hauteurs et de types phonatoires, mais aussi en raison de la variabilité des formes de surface. Cela se traduit par le fait que, très souvent, les tons de base (tons absolus ou tons de citation) subissent des modifications lorsqu'ils sont inclus dans des enchainements composés de deux syllabes ou plus (Wee, 2018 ; Zhang, 2022). Par surcroît, les tons relevés dans le domaine de réalisation du sandhi peuvent ne pas exister dans le système des tons de base (Yan et al., 2020). La directionalité de l'influence tonale est aussi un facteur important à considérer. En effet, par exemple, les langues wu présentent un système de type *Left Dominant Sandhi* (LDS) très productif où le ton précédent influence le ton suivant, tandis que les autres langues chinoises ont essentiellement le système inverse *Right Dominant Sandhi* (RDS) (Xu B. et al., 1981 ; Yan, 2018 ; Yan et al., 2020 ; Zee & Maddieson, 1980 ; Zhang & Meng, 2016). Ainsi, le sandhi tonal en wu comporte les deux directionalités – bien que RDS soit bien plus restreint que LDS – lesquelles auraient un rôle contrastif en permettant aux auditeurs de discriminer deux séquences dissyllabiques homophones (au niveau segmental) relevant l'une d'une fonction plutôt morpholexicale, l'autre d'une fonction plutôt morphosyntaxique. Plus précisément, LDS se manifeste au sein des composés de type Modificateur+Nom et de certains composés de type Verbe+Nom, Verbe+Modificateur, Sujet+Prédicat et composés par coordination (Qian, 1992 ; Xu B. et al., 1981, 1988; Yan et al., 2020). Cela permet à certains linguistes de soutenir que le domaine LDS shanghaïen correspondrait à un mot prosodique ou phonologique (Ling & Liang, 2017, 2019 ; Zhang et al., 2011 ; Zhang & Meng, 2016). RDS est quant à lui restreint à des structures syntaxiques spécifiques, telles que Verbe-Objet, Sujet-Prédicat, Verbe-Complément, Adverbe-Verbe, et s'étend également à des constructions coordonnées et aux structures adjectivales endocentriques (Feng, 2009 ; Xu B. et al., 1981, Xu B. et al., 1988). En conséquence, certains linguistes tels Ling & Liang (2019), Yan (2018) ou encore Zhang & Meng (2016) proposent que le domaine d'application du RDS correspondrait à une phrase phonologique/prosodique située à un niveau de construction supérieur par rapport à LDS lequel s'appliquerait plutôt à un mot prosodique/phonologique.

Dans de telles langues, les règles de formation du sandhi tonal impliqué dans la discrimination et la catégorisation de fonctions linguistiques font partie intégrante du fonctionnement tonal. Cependant, bien que des règles aient été formulées à partir de plusieurs études pour le shanghaïen (p. ex. Qian, 1992 ; Xu B. et al., 1981 ; Yan, 2018 ; *cf*. Table 1), elles sont exposées, dans leurs réalisations, à des variations considérables dues à des facteurs sociolinguistiques qui ont principalement pour origine l'influence dominante du chinois mandarin. De plus en plus de jeunes parlent les langues locales avec une influence forte du mandarin et, parmi les plus jeunes, beaucoup ne les parlent pas au quotidien et continuent à utiliser fréquemment le mandarin à la maison, même si le wu reste généralement pratiqué par les parents et grands-parents (Gao, 2016 ; Xu B. & Tao, 1997 ; Yan et al., 2020 ; Zhu Y. & Jiao, 2021).

σ_1 \ σ_2	T1 (52) T2 (34) T3 (23)	T4(<u>55</u>) T5(<u>13</u>)
T1 (52)	53 + 31	53 + 31
T2 (34)	33 + 44	33 + 44
T3 (23)	22 + 44	22 + 44
T4 (<u>55</u>)	<u>33</u> + 44	<u>33</u> + <u>44</u>
T5 (<u>13</u>)	<u>11</u> + 13	<u>11</u> + <u>13</u>

σ_2 \ σ_1	T1 (52) T2 (34)	T3(23)	T4(<u>55</u>)	T5(<u>13</u>)
T1 (52)	44 + 52	33 + 52	<u>44</u> + 52	<u>22</u> + 52
T2 (34)	44 + 34	33 + 34	<u>44</u> + 34	<u>22</u> + 34
T3 (23)	44 + 23	33 + 23	<u>44</u> + 23	<u>22</u> + 23
T4 (<u>55</u>)	44 + <u>55</u>	33 + <u>55</u>	<u>44</u> + <u>55</u>	<u>22</u> + <u>55</u>
T5 (<u>13</u>)	44 + <u>13</u>	33 + <u>13</u>	<u>44</u> + <u>13</u>	<u>22</u> + <u>13</u>

TABLE 1: Règles de sandhi tonal du shanghaïen pour *Left Dominant Sandhi* (à gauche) et *Right Dominant Sandhi* (à droite). Le soulignement indique une syllabe courte (coda glottale /ʔ/).

2 Objectifs et questions de recherche

L'étude que nous présentons ici est préliminaire à un projet de recherche plus ambitieux qui démarre. Celui-ci propose d'examiner la variabilité acoustique et la stabilité du système tonal wu et, en particulier, celle du sandhi tonal afin de répondre aux questions suivantes : quelles sont les règles de formation du sandhi tonal en wu ? Sont-elles homogènes ou sujettes à la variation selon l'âge, le genre du locuteur, le NSE, la variété dialectale, la situation géographique ? Nous pensons que les jeunes générations pourraient présenter des variations différentes de celles des générations plus âgées, et que le sandhi pourrait être influencé par d'autres langues, en particulier le mandarin, langue de scolarisation et de communication entre les jeunes, impulsant une dynamique de changement avec peut-être des répercussions sur l'ensemble du système tonal. De plus, les langues wu peuvent s'influencer mutuellement, en particulier depuis Shanghai. L'objectif de ce projet sera aussi de considérer l'impact de la variation sur le traitement perceptif des contrastes basés sur le sandhi tonal.

Mais revenons à cette étude initiale pour laquelle nous avons examiné les caractéristiques acoustiques des contours tonals des deux types de sandhi, LDS et RDS, dans les productions de trois jeunes adultes shanghaïens, en choisissant comme cibles des séquences dissyllabiques d'acceptions morphologique et syntaxique bien définies et couramment utilisées dans la langue. En puisant dans des recherches antérieures (Ling & Liang, 2019 ; Takahashi, 2011 ; Zhang & Meng, 2016 ; Yan et al., 2020 entre autres), nous nous interrogeons sur la réalisation des règles de sandhi lorsqu'elles sont produites par des jeunes adultes et cherchons pour cela à examiner et comparer la variation des réalisations tonales dans des syntagmes nominaux (censés soumis à LDS) ainsi que dans des syntagmes verbaux (soumis à RDS), présentant une structure syllabique identique.

3 Méthode

3.1 Corpus enregistré

Afin de convenir à l'ensemble des combinaisons tonales décrites dans les règles du sandhi tonal (*cf.* Table 1), quatre séquences dissyllabiques (deux pour la condition LDS et 2 pour la condition RDS) ont été sélectionnées en consultant principalement le *Grand dictionnaire du shanghaïen* de Qian et al. (2007) ainsi que le *Dictionnaire électronique du shanghaïen* créé par l'*École de langues wu* (吳語學堂, 2023), laquelle est animée par une communauté en ligne de locuteurs du shanghaïen et des autres langues wu. Nous nous sommes aussi référés aux listes de mots de Zhang & Meng (2016), Xu B. et al. (1981) et Takahashi (2011). Les quatre séquences ont aussi été sélectionnées de manière à pouvoir apparier les syllabes à attaque nasale entre les deux types de sandhi pour leur effet sur la fréquence fondamentale (diminution de hauteur) (*cf.* Xu Y., 1999). À noter aussi qu'un intérêt particulier a été porté aux tons T1 et T2 qui permettent d'obtenir des séquences dissyllabiques identiques au niveau segmental (et de même structure syllabique) dans les deux conditions de concaténation tonale de manière à ce que seul le sandhi tonal porte le contraste entre les paires. Enfin, le phénomène de sandhi générant des valeurs tonales qui ne sont pas forcément présentes dans les tons de base (p. ex. Yan et al., 2020), nous avons aussi retenu les monosyllabes correspondant aux deux éléments des dissyllabes cibles de manière à ce que leur prononciation isolée serve de référence (*baseline*) dans les analyses des effets de la contextualisation tonale.

Notre étude est ainsi basée sur l'examen des contours tonals d'une liste de 100 dissyllabes (5 tons de citation × 5 tons de citation × 4 stimuli = 100 stimuli) qui correspondent tous à des objets et des aliments couramment rencontrés dans la vie quotidienne. Le corpus a été soumis au préalable à

l'expertise d'une locutrice native du shanghaïen née en 1997 et a ensuite été mis en ordre aléatoire en raison d'un ordre par participant et chaque liste de 100 stimuli répétée 3 fois.

Les conditions de production LDS et RDS ont été contrôlées en utilisant deux phrases porteuses de structures morphosyntaxiques distinctes (TABLE 2). Ces deux phrases assurent un contrôle strict de l'environnement des mots-cibles, tout en permettant un environnement sémantiquement et syntaxiquement correct, mettant ainsi en évidence la relation entre la morphosyntaxe et les unités suprasegmentales.

荷蘭賣 + SN (p. ex. 炒飯)	後天我 + SV (p. ex. 炒飯)
/ɦu^{22} lɛ^{33} ma^{23}/ + /ts^{h}ɔ^{33} vɛ^{44}/	/ɦɤ^{22} t^{h}i^{33} ŋu^{23}/ + /ts^{h}ɔ^{44} vɛ^{13}/
Les Pays-Bas vendent *le riz frit*	Le jour après demain, je vais faire *frire du riz*

TABLE 2: Phrases porteuses pour *LDS* (à gauche) et *RDS* (à droite) avec exemples de mot cible.

La consigne était donnée aux participants de lire à voix haute et à une vitesse d'élocution normale les phrases présentées une par une sur un écran de 23". Celles-ci étaient transcrites en chinois simplifié. La liste des monosyllabes était acquise en suivant le même protocole à l'issue du recueil de la liste de phrases. Les enregistrements ont été réalisés dans la chambre anéchoïque du Gipsa-lab à l'Université Grenoble Alpes pendant l'été 2023. Les équipements utilisés sont un enregistreur numérique Marantz PMD 670 et un microphone AKG C1000 S. Les signaux ont été échantillonnés à 44.1 kHz.

3.2 Participants

Trois locutrices natives du shanghaïen, nées en 2000 à Shanghai et ayant grandi dans cette ville, ont accepté de participer à l'étude. Elles sont chacune originaires du quartier qu'elles ont toujours habité : Minhang, Qingpu et Putuo. Ces locutrices étaient étudiantes à la Faculté de la Langue Française de l'Université des Études Internationales de Shanghai et, au moment de l'enregistrement, étaient en programme d'échange en France depuis environ six mois. Leur profil linguistique est remarquablement similaire, maîtrisant toutes le shanghaïen, le mandarin, l'anglais et le français. Selon leurs auto-évaluations, elles ont un niveau natif en shanghaïen et en mandarin, tandis que leur maîtrise de l'anglais et du français se situe autour du niveau B2-C1. Quant à l'utilisation du shanghaïen, d'après leurs réponses, elles l'utilisent principalement avec leur famille, alors que le mandarin prédomine dans les autres circonstances. Ce profil linguistique est représentatif de la jeune génération shanghaienne, où le shanghaïen de la famille et le mandarin de l'école sont les deux langues premières, et au moins une langue étrangère (principalement l'anglais) est pratiquée. Le profil linguistique de ces trois participantes devrait nous permettre de cibler des tendances actuelles dans la réalisation du sandhi tonal en shanghaïen.

3.3 Traitements et analyses

Le corpus a été segmenté avec Praat (Boersma & Weenink, 2023) et annoté manuellement au niveau des séquences cibles, de leur décomposition en syllabes et en phones. L'extraction des valeurs de la f_0 pour chaque syllabe a été effectuée automatiquement, après vérification et parfois quelques ajustements manuels, avec le script ProsodyPro, développé par Xu Y. (2013). À partir de l'algorithme de lissage intégré pour la représentation des contours (utilisant une fenêtre triangulaire de Bartlett de 301 échantillons), dix mesures temporellement équidistantes ont été effectuées sur la totalité de la durée de la syllabe si l'attaque est voisée et sur la totalité de la rime dans le cas d'attaque

non voisée, avec un pas de 10. Les valeurs de f_0 ont été converties en demi-tons avec $y = 12\log_2(\frac{Hz}{50})$ puis ces valeurs ont été normalisées selon une transformation logarithmique z-score recommandée par Zhu X. (1995): $z = \frac{y_i - \mu}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}}$, avec $y_i = \log_{10}x$, $\mu = \overline{y_i}$ et x = valeur de f_0 observée.

L'analyse statistique des différences de niveaux de f_0 a été réalisée à l'aide d'un modèle linéaire mixte pour étudier l'impact des effets fixes : COMBINAISON TONALE, TYPE DE SANDHI et POINT DE MESURE (considéré ici comme une variable catégorielle, de 10 jusqu'à 100 par pas de 10), et de leur interaction, sur la variable réponse f_0. Ce modèle permet à la fois de tenir compte de la répétition des mesures (le facteur *LOCUTEUR* a été introduit comme effet aléatoire dans le modèle), de la variance résiduelle et/ou de la variabilité inter-individuelle qui peut changer d'un contexte à l'autre, d'un ton à l'autre ou d'un point de mesure à l'autre, mais aussi de la corrélation entre les valeurs de la variable réponse pour les différents points de mesures. Pour réaliser le modèle, nous avons utilisé la fonction *lme* du package *nlme* du logiciel R. Puis nous avons utilisé la fonction *glht* du package *multcomp* du logiciel R pour réaliser des comparaisons multiples d'où sont issues les p valeurs données ci-après. Les graphiques sont générés à l'aide des fonctions du package *ggplot2*.

4 Résultats

4.1 Tons de citation

Examinons d'abord les réalisations des cinq tons de citation. Les lignes noires tracées FIGURE 1 et FIGURE 2 représentent les contours normalisés des tons de citation prononcés dans la condition monosyllabique (prononciation isolée). Nous les superposons avec leurs contours correspondants dans les deux conditions dissyllabiques afin de faciliter plus loin les comparaisons. De manière générale, nos résultats sont conformes aux résultats d'études antérieures (Qian, 1992 ; B. Xu et al., 1988 ; Zhu X., 1995) : T1 est le seul ton haut descendant, T3 et T5 sont tous deux des tons bas montants, et T4 est un ton haut plat. Il faut pourtant noter que bien que T5 (13) et T3 (23) soient décrits différemment à l'initiation de la réalisation tonale, les contours de T5 et T3 sont en réalité trouvés très similaires, ce qui ne semble pas justifier cette différence de description. De plus, la tendance montante du contour de T2 (34) n'est pas tellement évidente et ne semble pas confirmer les observations des recherches antérieures (Qian, 1992 ; B. Xu et al., 1988 ; Zhu X., 1995).

4.2 Ton de sandhi *vs* ton de citation

Parmi les 25 combinaisons tonales de LDS examinées (FIGURE 1), onze présentent des différences significatives avec les tons de citation. La plupart de ces différences se manifestent au niveau de la deuxième syllabe confirmant pour ces cas l'existence d'une combinatoire vers la droite dans les contextes tonals T1+T1, T1+T4, T1+T5, T3+T1, T4+T1 et T5+T1, ainsi qu'au niveau de la frontière entre les deux syllabes T3+T3, T3+T5, T5+T2, T5+T3 et T5+T5. Seule la combinaison tonale T5+T3 montre une différence significative de la f_0 sur presque toute la durée de la première syllabe (désormais σ_1) – de 20 % à 100 % de la durée totale ($-4.445 \leq z \leq 8.997$, $p \leq 0.09$) – en plus d'un écart significatif au début de la syllabe 2 (désormais σ_2) – de 10 % à 40 % ($4.051 \leq z \leq 9.75$, $p \leq 0.02$). Il convient de remarquer que certaines combinaisons présentent des différences de f_0 importantes entre les tons dissyllabiques et les tons monosyllabiques comme cela est montré FIGURE 1, mais ces différences ne sont pas reflétées dans le modèle statistique en raison d'une forte

variabilité intra- et inter-locuteur. Cela concerne par exemple la combinaison T5+T4 ($1.401 \leq z \leq 2.958$, $p \geq 0.117$). On remarque également que les combinaisons T2+X ne présentent jamais de différences significatives entre tons de citation et tons de sandhi LDS ; idem pour T4+X sauf si X=T1 ($3.26 \leq z \leq 3.974$, $p \leq 0.044$). La raison est soit parce que contours et hauteurs de f_0 sont similaires, soit il existe une forte variabilité des réalisations.

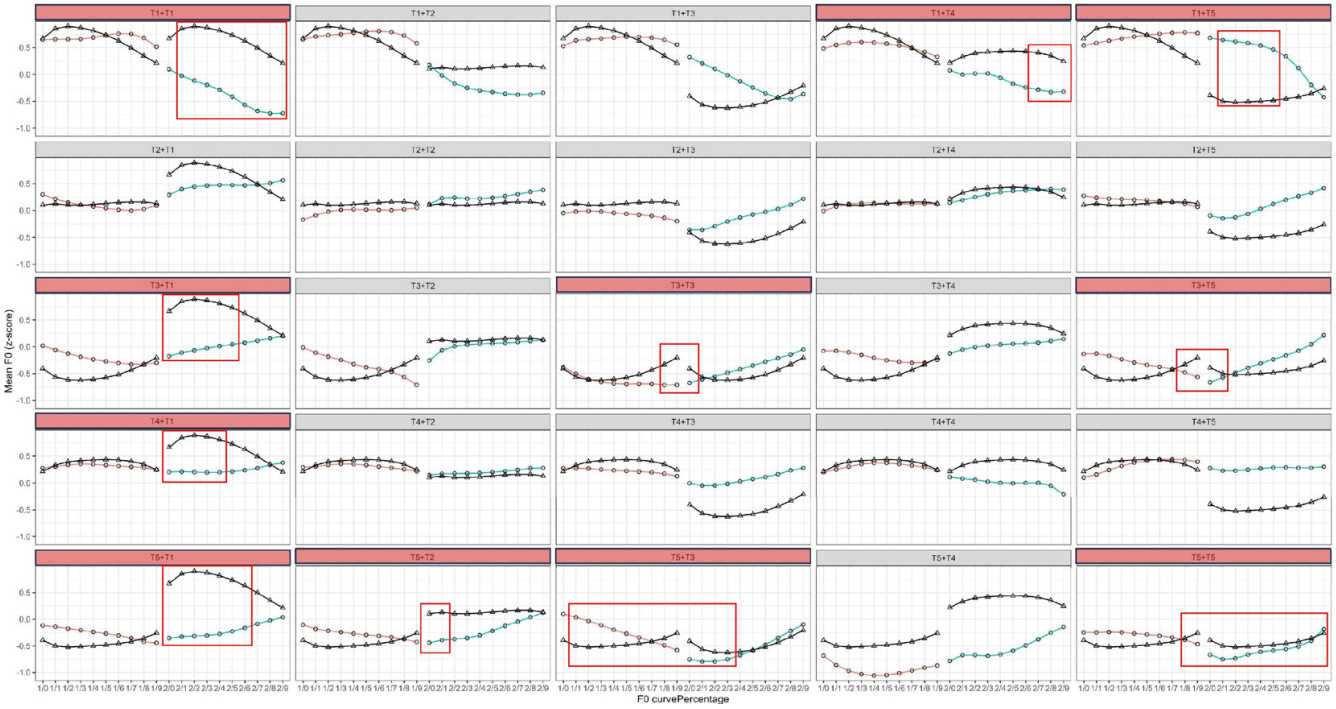


FIGURE 1: Contours moyens normalisés de f_0 pour les monosyllabes (ligne noire, ▲) et les dissyllabes sous l'effet du sandhi LDS (ligne colorée, ○). Les zones rouges indiquent les différences significatives ($p \leq 0{,}05$).

Les contours des tons de sandhi diffèrent des contours des tons de citation, surtout concernant σ_2. Pour T1+X, σ_1 perd la tendance descendante pour devenir un ton haut plat, tandis que σ_2, quel que soit son ton de citation, devient un ton descendant. De plus, les contours tonals T2+X, T3+X et T5+X présentent globalement une hauteur de ton légèrement supérieure pour σ_2 par rapport à σ_1. Cependant, contrairement aux études antérieures (Qian, 1992 ; Xu B. et al., 1988) qui décrivent les tons des séquences dissyllabiques T2/T3/T5+X comme des tons plats (*cf.* TABLE 1), les contours que nous observons présentent souvent des descentes ou des montées de faible amplitude, possiblement en raison de l'influence de la hauteur des tons environnants. Le contour tonal T4+X est quasiment plat et de hauteur similaire sur σ_1 et σ_2, ce qui diffère de la description traditionnelle (33+44).

Pour le cas de RDS, seules quatre des 25 combinaisons, T3+T5, T5+T1, T5+T3 et T5+T5, présentent des différences significatives dans la condition RDS (FIGURE 2). Parmi elles, T5+T3, ainsi qu'une partie de T5+T5 et T3+T5 affichent ces différences à la jonction entre les deux syllabes. Seule la combinaison T3+T5 montre un effet de RDS sur le début de σ_1. Enfin σ_2 est la plus impactée par le sandhi dans la combinaison T5+T1 sur la première moitié de sa durée et sur la totalité dans T5+T5. Ces résultats montrent qu'une combinatoire vers la gauche de la coarticulation tonale est peu présente dans le phénomène de RDS. À l'exception de quelques cas figurant dans les zones rouges, les contours des tons de σ_1 et σ_2 ne diffèrent pas beaucoup des tons de citation correspondants et des variations se produisent principalement aux frontières entre les deux syllabes. De plus, le ton de sandhi conserve souvent la hauteur et la tendance du contour de citation, ce qui souscrit davantage à la description de Takahashi (2011) qu'à celles de Qian (1992) et de Xu B. et al. (1988).

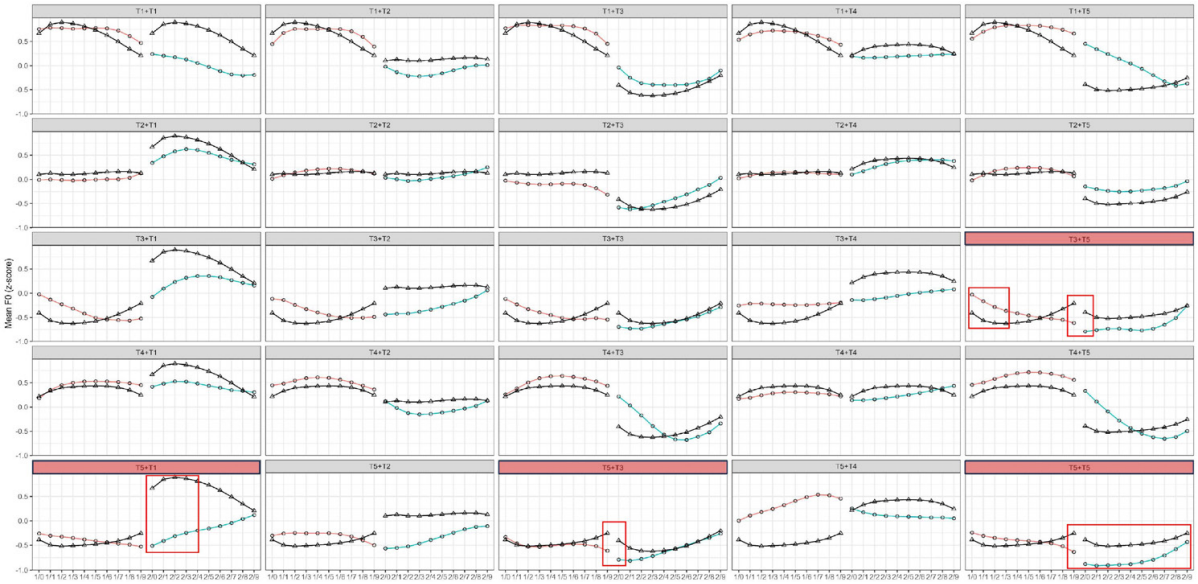


FIGURE 1 : Contours moyens normalisés de f_0 pour les monosyllabes (ligne noire, ▲) et les dissyllabes sous le sandhi RDS (ligne colorée, ○). Les zones rouges indiquent les différences significatives ($p \leq 0{,}05$).

4.3 LDS *vs* RDS

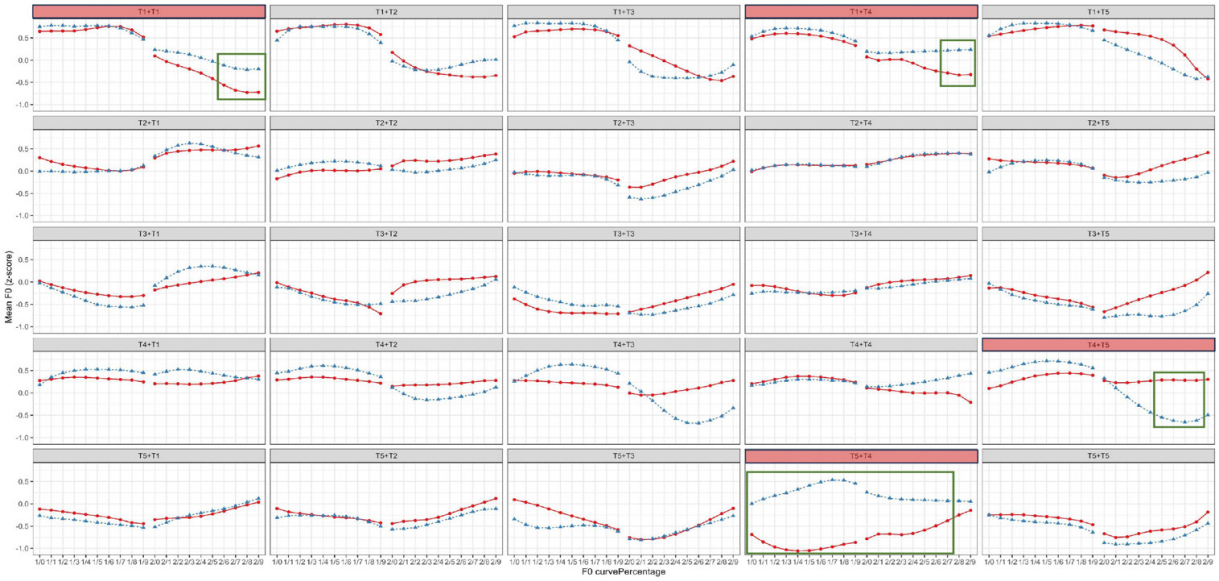


FIGURE 3: Valeurs moyennes normalisées de f_0 et contours tonals moyens des dissyllabes sous le sandhi LDS (ligne rouge, ●) et RDS (ligne bleue, ▲). Les zones vertes indiquent les différences significatives ($p \leq 0{,}05$).

FIGURE 3 sont présentés les contours superposés des deux types de sandhi. À notre grande surprise, les comparaisons entre LDS et RDS pour les mêmes combinaisons tonales montrent peu de différences significatives dans les valeurs de f_0, mise à part T5+T4 qui présente un effet de la directionalité du sandhi sur la quasi-totalité de la durée des deux syllabes (FIGURE 3). Pour trois autres combinaisons, T1+T1, T1+T4 et T4+T5, le seuil de significativité est atteint seulement à partir de la deuxième moitié de σ_2. Bien que nous puissions conclure que les différences significatives entre SN et SV résultent de différentes réalisations du sandhi tonal, nous ne pouvons pas écarter la possibilité de l'influence d'autres phénomènes prosodiques, tels que le patron intonatif de l'assertion

et ses caractéristiques en fin de phrase. En outre, bien que les réalisations tonales T4+T3 montrent des différences entre les deux types de sandhi, l'analyse statistique n'indique qu'une seule différence significative à 70 % de la durée de σ_2 ($p = 0,044$), ce qui est toutefois marginal. Nos observations concordent avec les résultats statistiques : à l'exception des cas signalés par les zones rouges et le cas de T4+T3, il n'existe guère de différence entre les contours de ton entre les types SV (RDS) et SN (LDS). Les contours tonals sont même quasi identiques dans certaines combinaisons telles que T2+T4 et T5+T1.

4.4 Variations tonales en fonction du rang syllabique et du contexte tonal

Pour présenter ces variations, nous les divisons en dix sous-catégories : T1+X, X+T1, T2+X, X+T2, T3+X, X+T3, T4+X, X+T4, T5+X et X+T5 (*cf.* FIGURE 3). Dans les SN (LDS), les seules différences significatives pour σ_1 sont trouvées uniquement dans les comparaisons entre T1+T4 et l'ensemble des autres combinaisons T1+X ($-3.642 \leq z \leq -6.491$, $p \leq 0.027$), ainsi que dans les comparaisons entre T5+T4 et l'ensemble des autres combinaisons T5+X ($-3.499 \leq z \leq -4.39$, $p \leq 0.046$). En revanche, en ce qui concerne les réalisations tonales de σ_2, les différences significatives sont trouvées principalement dans les contextes T3/T5+Tx et T2/T4+Tx, Tx représentant ici le même ton sur σ_2, p. ex. T2/T4+T5 *vs* T3/T5+T5 ($3.745 \leq z \leq 7.508$, $p \leq 0.018$) ; T2/T4+T3 *vs* T3/T5+T3 ($3.587 \leq z \leq 5.744$, $p \leq 0.033$). Dans ces cas, les réalisations jusqu'au milieu de σ_2 dans les contextes T3/T5+Tx présentent systématiquement une hauteur plus basse que dans les contextes T2/T4+Tx. Dans les SV (RDS), les différences tonales significatives sont juste trouvées au niveau de la finalisation de σ_1 et dans T5+T4 sur toute la durée de σ_1. Les variations tonales sur σ_2 sont toujours observées au niveau de la première moitié de sa durée et liées à la hauteur tonale de la deuxième partie de σ_1. Ainsi, les effets de la concaténation tonale sont majoritairement à l'initiation de σ_2.

5 Conclusion

Notre étude indique que généralement les valeurs de f_0 au niveau de σ_1 dans les SNs et SVs observés ne diffèrent pas significativement de celles mesurées pour les monosyllabes. Juste quelques différences ont été relevées à la frontière avec σ_2 qui peuvent être attribuées à la coarticulation tonale. Ce résultat est révélateur de l'effet du LDS dans les SNs, où le ton de citation est maintenu sur σ_1, tandis que σ_2 reçoit une valeur tonale en fonction du ton de σ_1. En revanche, les tons de citation semblent conservés sur les deux syllabes des SVs, avec quelques différences interprétables comme des effets de coarticulation tonale influençant de manière très brève les tons juste au niveau de la frontière syllabique. Par conséquent, nos résultats ne montrent pas l'existence d'une neutralisation tonale dans les SVs confirmant ainsi Takahashi (2011), tout comme ils ne montrent pas la présence de RDS chez nos trois jeunes locutrices contrairement aux descriptions « traditionnelles » citées plus haut. La présence d'un effet persévératif du ton de σ_1 sur le tout début de σ_2 ne relève pas du sandhi tonal phonologique ; cependant son rôle dans la perception et la catégorisation des syntagmes est certainement un domaine à explorer, tout comme d'ailleurs l'absence de RDS.

Nous prenons la mesure des limites de notre étude (3 participantes jeunes adultes, pas de contraste de position syntaxique, possibilité de variation par quartier urbain). L'extension du corpus est en cours en conservant les mêmes listes de mots et en recueillant des données cette fois sur le terrain wu. La prolongation de cette étude est d'autant plus nécessaire que les descriptions des langues wu n'ont pas encore réussi à proposer des règles de sandhi tonal claires. Nous testerons si cette situation est la conséquence d'une variation diatopique et diastratique importante avec pour principaux facteurs le développement de la mégapole de Shanghai, l'influence du mandarin, ainsi que le contact avec les langues de populations immigrantes (42 % de la population de la ville).

Références

Boersma, P., & Weenink, D. (2023). *Praat, a system for doing phonetics by computer.* (Version 6.3.09) [Computer software]. https://www.fon.hum.uva.nl/praat/

Feng, L. (2009). *Parlons shanghaïen*. l'Harmattan.

Gao, J. (2016). Sociolinguistic motivations in sound change: On-going loss of low tone breathy voice in Shanghai Chinese. *Papers in Historical Phonology*, *1*, 166.

Ling, B., & Liang, J. (2017). Focus encoding and prosodic structure in Shanghai Chinese. *The Journal of the Acoustical Society of America*, *141*(6), EL610–EL616. https://doi.org/10.1121/1.4989739

Ling, B., & Liang, J. (2019). The nature of left- and right-dominant sandhi in Shanghai Chinese—Evidence from the effects of speech rate and focus conditions. *Lingua*, *218*, 38–53. https://doi.org/10.1016/j.lingua.2018.02.004

Qian, N. (1992). *當代吳語研究[Studies in the contemporary Wu-dialects]* (1st ed.). Shanghai Educational Publishing House.

Qian, N., Tang, Z., & Xu, B. (2007). *上海話大詞典[Grand dictionnaire du shanghaïen]*. Shanghai Lexicographical Publishing House.

Takahashi, Y. (2011). 上海語声調音韻論における窄用式変調の地位. 中国語学, *258*, 99–114.

Wee, L.-H. (2018). *Phonological tone*. Cambridge University Press.

Xu, B., Tang, Z., & Qian, N. (1981). 新派上海方言的連讀變調[Sandhi tonal in modern Shanghainese]. *方言 [Dialects]*, *2*, 145–155.

Xu, B., Tang, Z., You, R., & Qian, N. (Eds.). (1988). *上海市區方言志[Dialect documents of Shanghai Municipal]*. Shanghai Educational Publishing House.

Xu, B., & Tao, H. (1997). *上海方言詞典[Dictionary of Shanghai dialect]* (1st ed.). Jiangsu Education Publishing, Ltd.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of *f0* contours. *Journal of Phonetics*, *27*(1), 55–105. https://doi.org/10.1006/jpho.1999.0086

Yan, H. (2018). *The Nature of Variation in Tone Sandhi Patterns of Shanghai and Wuxi Wu* (Vol. 4). Springer Singapore. https://doi.org/10.1007/978-981-10-6181-3

Yan, H., Chien, Y.-F., & Zhang, J. (2020). Priming the Representation of Left-Dominant Sandhi Words: A Shanghai Dialect Case Study. *Language and Speech*, *63*(2), 362–380. https://doi.org/10.1177/0023830919849081

Zee, E., & Maddieson, I. (1980). Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis. *Glossa*, *14*(1), 45–88.

Zhang, J. (2022). Tonal Processes Defined as Tone Sandhi. In C.-R. Huang, Y.-H. Lin, & I.-H. Chen (Eds.), *The Cambridge Handbook of Chinese Linguistics* (1st ed., pp. 291–312). Cambridge University Press. https://doi.org/10.1017/9781108329019.017

Zhang, J., Lai, Y., & Sailor, C. (2011). Modeling Taiwanese speakers' knowledge of tone sandhi in reduplication. *Lingua*, *121*(2), 181–206. https://doi.org/10.1016/j.lingua.2010.06.010

Zhang, J., & Meng, Y. (2016). Structure-dependent tone sandhi in real and nonce disyllables in Shanghai Wu. *Journal of Phonetics*, *54*, 169–201. https://doi.org/10.1016/j.wocn.2015.10.004

Zhu, X. (1995). *Shanghai Tonetics* [Doctoral dissertation]. The Australian National University.

Zhu, Y., & Jiao, Z. (2021). A Case Study of Intergenerational Inheritance of Shanghai Dialect in the Environment of Mandarin Popularization. *Journal of Tianjin Foreign Studies University*, *28*(2), 98–108.

吳語學堂. (2023). https://www.wugniu.com/

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Synthèse de syllabes avec un modèle de Maeda piloté par une représentation complexe

Frédéric Berthommier
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
frederic.berthommier@gipsa-lab.grenoble-inp.fr

RÉSUMÉ

Un modèle mathématique est construit sur une notion de coordination des articulateurs à partir d'une représentation bidimensionnelle complexe. Les voyelles sont représentées par des positions en bordure du cercle unité, et pour le modèle de Maeda, les paramètres articulatoires sont générés avec une fonction de coordination facile à configurer. Les consonnes plosives /bdg/ sont encodées de la même manière, mais pour produire des syllabes, le graphe reliant les positions phonétiques distingue les arcs vocaliques et les arcs consonantiques. Un flux de paramètres articulatoires est dérivé par application sélective de la fonction de coordination. Les contributions de deux groupes d'articulateurs sont ainsi superposées et synchronisées pour piloter le modèle de Maeda et obtenir la synthèse de trajectoires formantiques. Ce modèle possède un schéma déterministe similaire à celui de la phonologie articulatoire, mais de nombreuses simplifications sont opérées.

ABSTRACT

Syllable synthesis with a Maeda model driven by a complex representation

A mathematical model is built on a notion of articulator coordination based on a complex two-dimensional representation. Vowels are represented by positions at the edge of the unit circle, and for the Maeda model, articulatory parameters are generated with an easy-to-configure coordination function. The plosive consonants /bdg/ are encoded in the same way, but to produce syllables, the graph linking phonetic positions distinguishes between vowel arcs and consonant arcs. A stream of articulatory parameters is derived by selective application of the coordination function. The contributions of two groups of articulators are thus superimposed and synchronised to drive a Maeda model and obtain the synthesis of formantic trajectories. This model has a deterministic scheme similar to that of articulatory phonology, but many simplifications are made.

MOTS-CLÉS : synthèse articulatoire, relation articulatori-acoustique, voyelles, consonnes, syllabes, coordination des articulateurs, phonologie articulatoire.

KEYWORDS: articulatory synthesis, articulatory-acoustic relationship, vowels, consonants, syllables, coordination of articulators, articulatory phonology.

1 Introduction

Lors de la production de la parole, les mouvements de la langue, de la mâchoire et des lèvres sont pseudo-périodiques. Prosaïquement, le geste associé à /aiua/ est une rotation de la langue, d'abord vers le haut et l'avant pour /ai/, puis vers l'arrière avec /iu/ et enfin vers l'arrière et vers le bas pour clore le cycle avec /ua/. Ici, la cyclicité de la position des formants est observable avec un simple

spectrogramme et on en déduit une dépendance entre gestes de la langue et position des formants. Des *nomogrammes* synthétisant la relation articulatori-acoustique ont été construits avec le tube de Fant (Badin *et al.*, 1990) en faisant varier la position et l'ouverture d'une constriction représentant la langue ainsi que l'ouverture à une extrémité. Malheureusement, cette relation est très complexe et elle ne peut pas être modélisée mathématiquement. Dans un spectrogramme ordinaire, les modulations formantiques sont bien présentes, mais on ne peut pas en déduire facilement les mouvements du tractus vocal à cause de la non-linéarité de la relation et surtout de la multiplicité des configurations articulatoires produisant un ensemble de formants donné. De fait, ceci limite considérablement l'étude de la relation entre perception et production de la parole.

Les consonnes sont engendrées par des constrictions du conduit vocal disposées en demi-cercle depuis les lèvres pour le /b/, en position coronale pour le /d/, palatale ou vélaire pour /g/ jusqu'aux lieux d'articulation pharyngal et épiglottal (Schwartz *et al.*, 2012). La Task Dynamics (TD) avec (Nam *et al.*, 2004) encode explicitement les lieux d'articulation des consonnes comme des angles et il est suggéré que des positions angulaires discrètes émergent aussi pour les voyelles avec le modèle de Maeda (Gaines *et al.*, 2021). Pour réaliser les constrictions de concert avec la production des voyelles, la langue effectue des mouvements qui sont potentiellement planifiables dans une représentation complexe par le biais d'un codage angulaire. L'intérêt de celle-ci semble secondaire si on ne coordonne pas tous les articulateurs (la langue et les lèvres) au cours du temps. Coordination et synchronisation sont les clefs de la planification syllabique selon (Xu, 2017, 2020) et nous introduisons ces deux mécanismes en faisant appel à une représentation complexe.

C'est à rebours des approches fondées sur les réseaux neuronaux que nous proposons une méthode de synthèse des syllabes basée sur des régularités du mécanisme de production exprimables dans des espaces de très faible dimension. Le modèle de Maeda (Maeda, 1979, 1990) est construit à partir de coupes sagittales et la factorisation de ces données en termes de commandes articulatoires favorise cette approche. Les aspects cognitifs et la notion d'apprentissage sont relégués au second plan pour privilégier l'identification de ces régularités. En établissant une continuité entre l'espace vocalique et la structure syllabique de la parole avec l'appui de quelques travaux, le but n'est pas de construire des applications, mais d'offrir un éclairage sur la structure profonde de la parole, complémentaire de celui apporté par les réseaux neuronaux (Dupoux, 2018).

2 La construction de l'espace vocalique

Préalablement, nous avons montré que le DRM, constitué de huit tubes (Mrayati *et al.*, 1988), est très approprié pour construire une *bijection* entre le cercle unité et l'espace vocalique F1-F2 en exploitant les relations décrites entre variations formantiques et variations du diamètre de chacun des tubes (Berthommier, 2021). Ce modèle n'est pas anatomique et les diamètres des tubes représentent grossièrement, par régions distinctives, la fonction d'aire obtenue à partir d'une coupe sagittale du conduit vocal. Nous avons remplacé la méthode itérative de couverture de l'espace vocalique partant du tube neutre par une détermination directe de la périphérie, où sont situées les voyelles cardinales. Une fonction de coordination du diamètre des tubes est définie à partir des diamètres fixés pour les trois voyelles cardinales /aiu/ (Berthommier, 2021). Cette fonction relie une position dans le cercle unité avec les diamètres des tubes en les corrélant. Pour aller vers les modèles articulatoires, une comparaison a été réalisée entre le DRM et le modèle de Fant pour la construction du triangle vocalique. Dans les deux cas, la détermination de la fonction de coordination n'est pas triviale, car

on associe une représentation trigonométrique à la géométrie linéaire d'un système de tubes. En revanche, avec un modèle de Maeda, ici représenté par VLAM (Boë & Maeda, 1998), nous constatons que cette fonction est déductible sans calcul. Ceci atteste la compatibilité entre un modèle dérivé d'une statistique de configurations articulatoires réelles projetées sur une grille semi-polaire et la fonction de coordination qui *génère* des configurations à partir d'une représentation complexe.

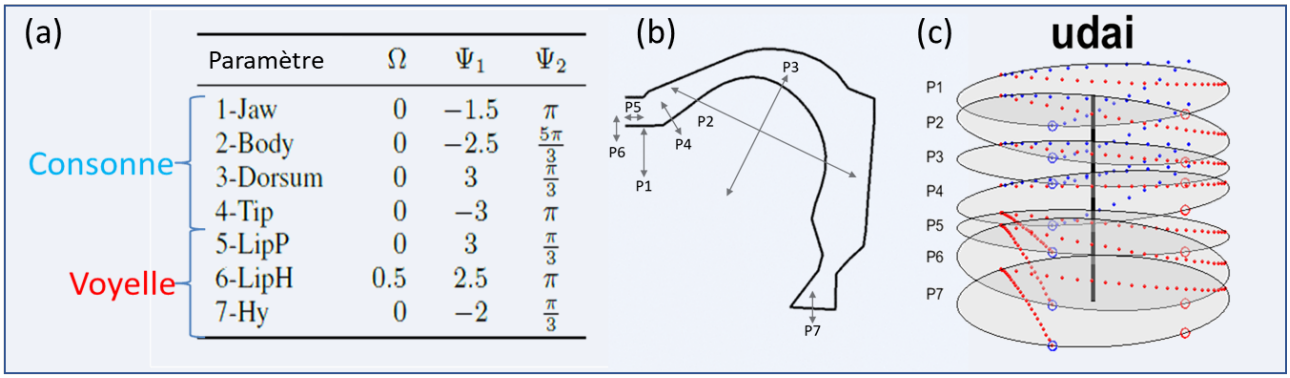


FIGURE 1 – (a) Configuration du modèle de Maeda. (b) Action des 7 paramètres articulatoires. (c) Représentation des trajectoires paramétriques en pile pour la syllabe /udai/ où les valeurs apparaissent verticalement. Les trajectoires vocaliques /ua/ et /ai/ sont en rouge et en bleu pour la consonne /d/.

La configuration de la fonction de coordination Fig 1a est basée sur la moyenne Ω et l'étendue Ψ_1 de chaque paramètre (données a priori de VLAM) plus un angle à déterminer Ψ_2. La valeur de chaque paramètre articulatoire $P_i, i = 1..7$ est calculée indépendamment pour un point donné (ρ_V, θ_V) du domaine complexe. La coordination entre P_i est assurée par le produit du même complexe conjugué $\rho_V e^{-j\theta_V}$ avec chaque valeur complexe Ψ_i du modèle :

$$P_i - \Omega_i = Re\left[\Psi_i \rho_V e^{-j\theta_V}\right]$$
$$\boldsymbol{P - \Omega} = Re\left[\boldsymbol{\Psi}\rho_V e^{-j\theta_V}\right] = \rho_V \boldsymbol{\Psi_1}\, cos(\boldsymbol{\Psi_2} - \theta_V) \quad (1)$$

Remarquons que c'est une simple cosinusoïde. Les angles inconnus $\boldsymbol{\Psi_2}$ sont fixés afin que les angles des voyelles /iau/ sur le cercle unité ($\rho_V = 1$) soient $\theta_V = \{\frac{5\pi}{3}, \pi, \frac{\pi}{3}\}$. Pour cela, on assigne l'une des voyelles /iau/ à chaque articulateur de telle sorte que $P_i - \Omega_i = \Psi_{1i}$ pour cet articulateur lorsque $\theta = \Psi_{2i}$. Le paramètre corps de la langue (Body) est aligné sur la réalisation du /i/, l'ouverture des lèvres (LipH), l'abaissement de la pointe (Tip) et de la mâchoire (Jaw) sur /a/ et l'allongement du conduit vocal (LipP, Hy) ainsi que la flexion de la langue (Dorsum) avec /u/ (Fig. 1b pour l'action des paramètres). Ces spécifications établissent une bijection entre le domaine du cercle unité Fig. 2a et une surface dans l'espace des formants F1-F2-F3 (dite surface vocalique Fig. 2c). La répartition angulaire d'autres voyelles cardinales est déductible par antisymétrie centrale. Par exemple, l'angle de /ɛ/ est égal à $\frac{\pi}{3} + \pi = \frac{4\pi}{3}$ avec comme conséquence un renversement de la valeur des paramètres articulatoires du /u/ par rapport à la valeur du neutre car $cos(x + \pi) = -cos(x)$. Les angles $\theta_V \in \{0, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \pi, \frac{4\pi}{3}, \frac{3\pi}{4}, \frac{5\pi}{3}\}$ sont associés respectivement aux voyelles /ɨ, u, o, ɔ, a, ɛ, e, i/. Celles-ci résultent de l'application de la fonction de coordination et elles ne nécessitent pas d'ajustement particulier. En revanche, nous avons retrouvé à l'écoute l'angle $\frac{11\pi}{6}$ pour la voyelle /y/ moins fréquente (Moran & McCloy, 2019). Les paramètres articulatoires du modèle de Maeda reflètent les commandes musculaires du tractus vocal essentiellement basées sur des relations *agonistes-antagonistes* (Kröger & Bekolay, 2022). L'antisymétrie centrale est une conséquence directe de la

structure anatomique associant forme semi-circulaire du tractus vocal et commandes musculaires. Cet ensemble est physiquement relié à la structure de l'espace vocalique (Schroeder, 1967; Mrayati *et al.*, 1988; Berthommier, 2021). Ces régularités sont prises en compte par la fonction de coordination, expression mathématique qui ancre l'espace vocalique dans l'anatomie du tractus vocal en établissant une continuité structurelle forte (Fig. 2).

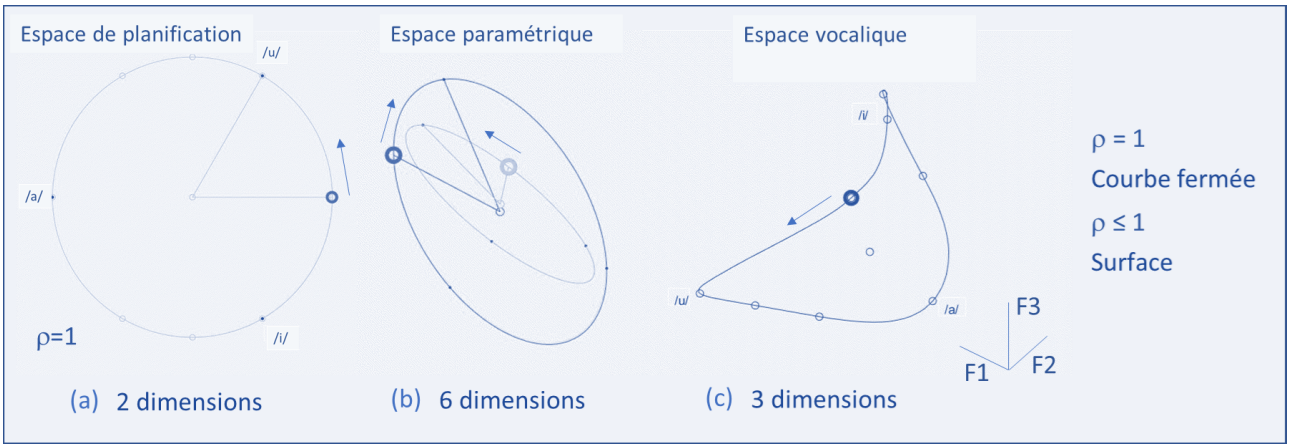


FIGURE 2 – Illustration de l'effet de la fonction de coordination. Un parcours entre 0 et 2π dans le sens trigonométrique dans (a) l'espace de planification où les voyelles cardinales sont codées par des angles se traduit par des parcours *elliptiques* dans (b) l'espace paramétrique. Le point de départ $\odot$ et le secteur $\{0, \frac{\pi}{3}\}$ sont figurés. Ici, les paramètres articulatoires sont séparés en 2 groupes visibles {Jaw, LipP, LipH} et non visibles {Body, Dorsum, Apex} pour figurer les corrélations induites par la fonction de coordination. La représentation du paramètre Hy est omise bien qu'il soit pris en compte. (c) Ces mêmes paramètres entraînent un parcours de la périphérie d'une surface vocalique exprimée par la position des formants F1-F2-F3. On retrouve dans l'ordre les voyelles cardinales /u, o, ɔ, a, ɛ, e, i/ notées ∘ en périphérie de cette surface. La fonction de coordination établit une bijection qui associe tout point du domaine complexe situé à l'intérieur du cercle unité avec un vecteur de paramètres inclus dans les deux ellipses et un point situé sur la surface vocalique, calculé avec (Badin & Fant, 1984) à partir de ce même vecteur.

La relation qu'entretiennent les deux surfaces Fig. 2a-c avec l'espace paramétrique de dimension 6 est visualisée Fig. 2b à l'aide de deux ellipses placées dans un espace 3D. Chacune d'elles représente 3 paramètres visibles et non visibles avec deux points qui sont en rotation lorsque l'on parcourt le cercle unité Fig. 2a. À droite Fig. 2c, la périphérie de la surface vocalique est décrite de façon concomitante. Si l'on se dirige vers le centre du cercle unité, les paramètres se rapprocheront de ceux de la voyelle neutre, et les formants des valeurs (f1n, f2n, f3n) en suivant cette surface. Cela réduit considérablement le calcul de la trajectoire d'une diphtongue dans l'espace paramétrique à sept dimensions. Sans une telle simplification et dans un espace plus grand, la synthèse des diphtongues avec VocalTractLab se montre laborieuse (Xu *et al.*, 2023). (Story *et al.*, 2018) construisent une bijection entre une paramétrisation 2D des formes du conduit vocal et F1-F2. Il s'agit du résultat le plus proche du nôtre, mais il est obtenu avec un modèle à tubes et l'espace vocalique résultant présente des défauts notables. Ici, le *pointage* exercé depuis le domaine complexe vers la surface F1-F2-F3 est très régulier et les trajectoires dessinées dans l'un apparaissent peu déformées sur l'autre.

3 La représentation des consonnes et leur coproduction

Nous proposons d'étendre l'usage de la fonction de coordination aux consonnes en les plaçant dans le même référentiel. Selon (Öhman, 1966) avec des syllabes V1CV2, les consonnes plosives perturbent la trajectoire de F2 marquant une transition entre les voyelles V1 et V2. La modélisation la coproduction voyelles/consonne (Öhman, 1967) est basée sur une pondération qui ne *sépare pas* les parties du conduit vocal pour les affecter à la constriction ou à cette transition vocalique. Par contre, la structure du DRM est très appropriée pour effectuer une telle séparation. L'effet de perturbation de F2 est obtenu en sélectionnant le tube du lieu d'articulation pour effectuer la constriction et en laissant les autres tubes contribuer à la transition vocalique (Carré & Chennoukh, 1995). Fondé sur un principe équivalent, TubeTalker (Story, 2009, 2013) utilise des formes anatomiques issues d'IRMs, mais le pincement des tubes dans la région orale n'est pas anatomique. On obtient les plosives /bdg/ sur le plan acoustique avec une trace correcte de la coarticulation (Story & Bunton, 2021) mais cela ne renseigne pas bien sur la relation articulatori-acoustique et son contrôle. Cependant, la propriété de séparabilité des paramètres mise en évidence avec ces modèles est intéressante car elle simplifie la coproduction envisagée par (Öhman, 1967). Comme amélioration aboutissant à un résultat audible, (Birkholz, 2013) dispose pour chaque consonne de 3 cibles articulatoires pour /aiu/ et il réalise une moyenne pondérée pour les autres voyelles. Pour planifier la synthèse de syllabes, VocalTractLab possède un niveau vocalique spécifique dans son panel de planification gestuelle. Mais n'y a pas de séparabilité comme avec les modèles à tubes. La TD adopte une solution flexible en coordonnant et en pondérant la participation des articulateurs à une tâche de constriction. La synthèse d'un sous-ensemble de VCVs avec un modèle composite suivant ses principes et à l'aide du modèle de Maeda a été décrite récemment (Alexander *et al.*, 2019). La propriété de séparabilité des articulateurs n'est pas ou peu appliquée en dehors des modèles à tubes.

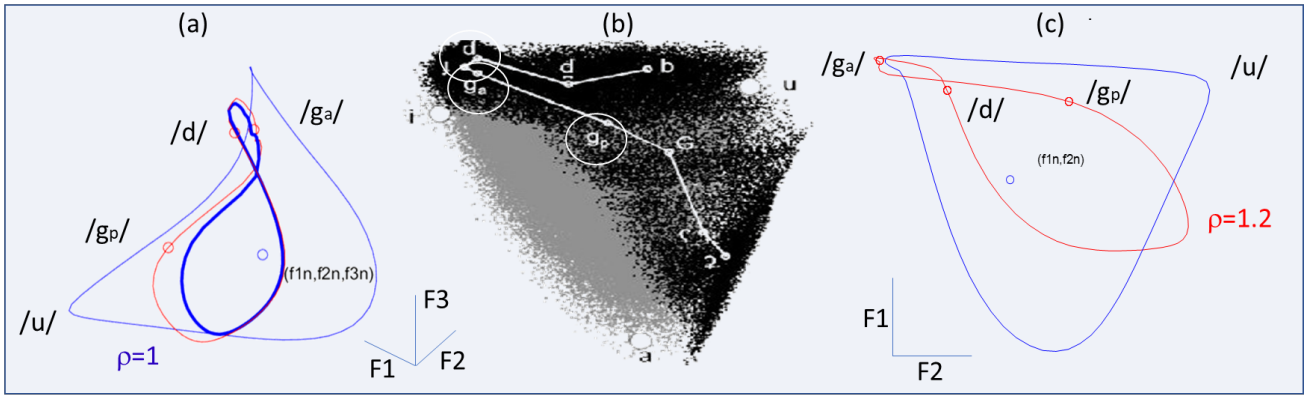


FIGURE 3 – Répartition des consonnes plosives relative à la surface vocalique. La marge périphérique du cercle unité est parcourue tout en bloquant les 3 paramètres {LipH, LipP, Hy} en position neutre. (a) Dans l'espace F1-F2-F3 (b) Référence F1-F2 (Schwartz *et al.*, 2012) (c) Dans l'espace F1-F2.

Les consonnes plosives /bdg/ ont des caractéristiques dynamiques complexes et elles sont présentes en grande proportion dans les langues du Monde, ce qui motive leur choix. En phonologie articulatoire (Saltzman & Munhall, 1989), la notion de tâche pour la réalisation de constrictions *privilégie* implicitement les consonnes par rapport aux voyelles associées à des formes globales du tractus vocal. Cependant, une analogie de production entre consonnes et voyelles est inhérente au lieu de

constriction (Gaines *et al.*, 2021). Une statistique exhaustive des lieux de constriction a été réalisée avec VLAM par (Schwartz *et al.*, 2012). Par le biais d'un nomogramme, elle met en relation les lieux d'articulation des consonnes avec les formants F1-F2-F3 produits au relâchement de la constriction.

Avec la fonction de coordination, les résultats de (Schwartz *et al.*, 2012) sont reproduits sans recourir à des simulations de Monte Carlo Fig. 3b. Avec le même principe que pour les voyelles et en délimitant $1 \leq \rho_C \leq 1.2$ tout en parcourant le cercle unité, on ne sélectionne que les 4 paramètres contrôlant la langue et la mâchoire Fig. 1a-b, en fixant en position neutre les trois paramètres restants Fig. 3a-c. D'une part, cette sélection laisse libres trois paramètres qui sont séparables pour effectuer une tâche de coproduction des voyelles. D'autre part, en augmentant les valeurs de ρ, une fermeture du conduit vocal, limitée par une rectification douce, est engendrée sur une partie du cercle unité. Consonnes et voyelles se retrouvent plongées dans une représentation commune. La procédure de découverte proposée comme analogue au babillage (Schwartz *et al.*, 2012) est également simplifiée puisque les angles attribuables aux consonnes /dg/ (respectivement $\theta_C \cong \{\frac{3\pi}{2}, -\frac{\pi}{12}, \frac{\pi}{3}\}$ avec deux positions, antérieure et postérieure, pour /g/) sont très proches de ceux des voyelles cardinales /e, i, u/. Avec un tel a priori, leur découverte ne demande que peu de tâtonnements et nous restons dans le cadre de l'émergence structurale mise en évidence pour les voyelles cardinales. Les phonèmes sont replacés dans un cadre intrinsèquement structuré et le conflit inhérent au choix de contrôle des lieux de constriction plutôt que celui de la forme globale du tractus vocal est résolu naturellement.

4 La synthèse des syllabes

Pour définir une coproduction fondée sur la séparabilité des articulateurs, nous proposons que la structure syllabique soit représentée par un graphe constitué de noeuds pour les phonèmes et d'arcs représentant des trajectoires entre chaque noeuds, planifiées dans le plan complexe. On distingue les trajectoires vocaliques et consonantiques entre deux points de rendez-vous avec une *synchronisation* (Xu, 2017, 2020). Les articulateurs associés à chacune des branches sont définis par un processus de sélection. La synchronisation est quant à elle obtenue en associant les arcs à des multiples entiers d'une période de référence T. La structure temporelle de la production est liée à l'enchaînement des gestes articulatoires qui forment des unités syllabiques de taille intermédiaire puis des mots par concaténation. Sur le plan temporel, nous faisons l'hypothèse que les segments ont une durée relativement constante à court terme mais que la structure superficielle reste pseudo-périodique.

La synthèse des syllabes CV découle de cette représentation des consonnes. En effet, les articulateurs qui ne sont pas affectés sont *libres* pour les transitions vocaliques. S'il est couramment admis que C et V ont un début synchrone, en phase selon la phonologie articulatoire (Browman & Goldstein, 1992), il faut spécifier l'anticipation de la forme des lèvres. Exemple classique, l'arrondissement des lèvres précède l'onset consonantique pour /Cu/ (Daniloff & Moll, 1968). Pour réaliser la coproduction CV et assurer la synchronisation des articulateurs, nous introduisons comme point d'ancrage une voyelle réduite (centralisée). Ici, on assigne à la trajectoire /u_r/ (réduit) vers /u/ les paramètres qui ne sont pas nécessaires à l'articulation de la consonne. Pour /Cu/, l'anticipation de la protrusion des lèvres apparaît avec le paramètre LipP.

Les trajectoires des voyelles et des consonnes sont planifiées dans le plan complexe Fig. 4a en formant des arcs entre 2 points (ρ_1, θ_1) et (ρ_2, θ_2) :

$$z(t) = (1 - \rho(t))\, \rho_1\, e^{j\theta_1} + \rho(t)\, \rho_2\, e^{j(\theta_2 + \frac{\nu}{K}\theta(t))} \tag{2}$$

où t varie entre 0 et nT, $\rho(t) = cos(\frac{\theta(t)}{2})$ détermine le profil de vitesse et $\nu = \pm 1$ et K sont des paramètres de forme de la trajectoire (voir aussi (Berthommier, 2023)). Lorsque K est grand ($K = 30$ pour les arcs de voyelles et $K = 10$ pour les arcs de consonnes), les trajectoires deviennent plus rectilignes dans le plan complexe (voir Fig. 4a).

À chaque instant t, l'ensemble des paramètres est coordonné avec l'équation 1 et la trajectoire paramétrique est obtenue par enchaînement de périodes de temps de durée nT. Au cours de chaque période, les trajectoires des paramètres sont la partie réelle du produit du vecteur colonne complexe $\mathbf{\Psi}$ qui représente le modèle articulatoire, et du vecteur ligne complexe $\bar{z}(t)$ issu de la planification. Il en résulte des matrices de dimension $7 * nT$ qui sont concaténées :

$$\begin{aligned} \boldsymbol{P}(t) - \mathbf{\Omega} = Re\left[\mathbf{\Psi}\bar{z}(t)\right] = (1-\rho(t))\,\rho_1\,\mathbf{\Psi_1}\,cos(\mathbf{\Psi_2} - \theta_1) \\ + \rho(t)\,\rho_2\,\mathbf{\Psi_1}\,cos(\mathbf{\Psi_2} - \theta_2 - \frac{\nu}{K}\theta(t)) \end{aligned} \tag{3}$$

Il reste à instancier le processus de sélection reposant sur la séparabilité des articulateurs. Les trajectoires des segments vocaliques et des pauses sont définies par l'équation précédente seule, tandis que la *superposition* des trajectoires vocaliques et consonantiques nécessite une coordination séparée. Notons que la fonction de coordination s'applique alors en chaque t sur les deux trajectoires $\bar{z}_v$ et $\bar{z}_c$. La coarticulation entre voyelles et consonnes est produite par la superposition de ces 2 branches ayant la même durée nT ainsi que les mêmes points de départ et d'arrivée qui sont des voyelles éventuellement réduites (i.e. VCVr pour VC ou VrCV pour CV) :

$$\boldsymbol{P}(t) - \mathbf{\Omega} = Re\left[S_v.\mathbf{\Psi}\bar{z}_v(t) + S_c.\mathbf{\Psi}\bar{z}_c(t)\right] \tag{4}$$

où S_v et S_c sont les deux vecteurs de sélection exclusifs composés de zéros et de uns pour les articulateurs sélectionnés avec $S_v + S_c = \mathbf{1}$ (un vecteur colonne de 7*1). Avec un produit de Hadamard, les composantes non sélectionnées du vecteur complexe $\mathbf{\Psi}$ sont annulées. La composition de ces vecteurs dépend de la (ou des) consonnes. Nous avons vu que pour /dg/, la mâchoire et les 3 paramètres de la langue sont nécessaires tandis que pour /b/ la sélection de l'ouverture des lèvres est complémentée par la mâchoire et le corps de la langue. D'autres détails concernant la sélection des articulateurs et la planification des clusters consonantiques sont disponibles (Berthommier, 2023).

Dans l'exemple /udai/ Fig. 4a, la trajectoire /ua/ est perturbée par /d/. Durant cette coproduction, deux ensembles d'articulateurs notés Fig. 1a sont coordonnés séparément. Tandis que le premier ensemble (consonne) suit la trajectoire de /u/ vers /d/ puis vers /a/ en deux périodes T, le second (voyelle) va de /u/ à /a/ de façon synchronisée. Les trajectoires visibles dans la pile paramétrique Fig. 1c sont différentes et leurs effets sur l'unique trajectoire formantique se superposent. Mathématiquement, la conséquence de la superposition est d'augmenter la dimension de la planification de 2 (1 nombre complexe) à 4 (2 nombres complexes). Cette augmentation est observable dans l'espace des formants Fig. 4c : alors que la trajectoire de la diphtongue /ai/ reste sur la surface vocalique, la trajectoire des segments superposés /uda/ Fig. 4c en sort. La coarticulation du /d/ est réalisée par une attraction du locus situé au point de rebroussement de la trajectoire, vers la branche /ua/ portée par les lèvres et le larynx. Notons enfin que si la méthode est adaptée pour planifier Fig. 4a les trajectoires des formants Fig. 4c-d et offrir des animations du modèle de Maeda, ce qui concerne le contrôle effectif des articulateurs reste très hypothétique. Comme nous l'avons indiqué, le champ d'application est celui de la phonologie articulatoire où il peut apporter une série de simplifications, en particulier sur la planification des syllabes. Pour la reproduction de gestes articulatoires, la TD s'appuie sur des enregistrements de données EMA correspondant aux variables articulatoires du modèle, tandis qu'ici une telle référence n'existe pas. En revanche, les paramètres de Maeda issus d'une factorisation par

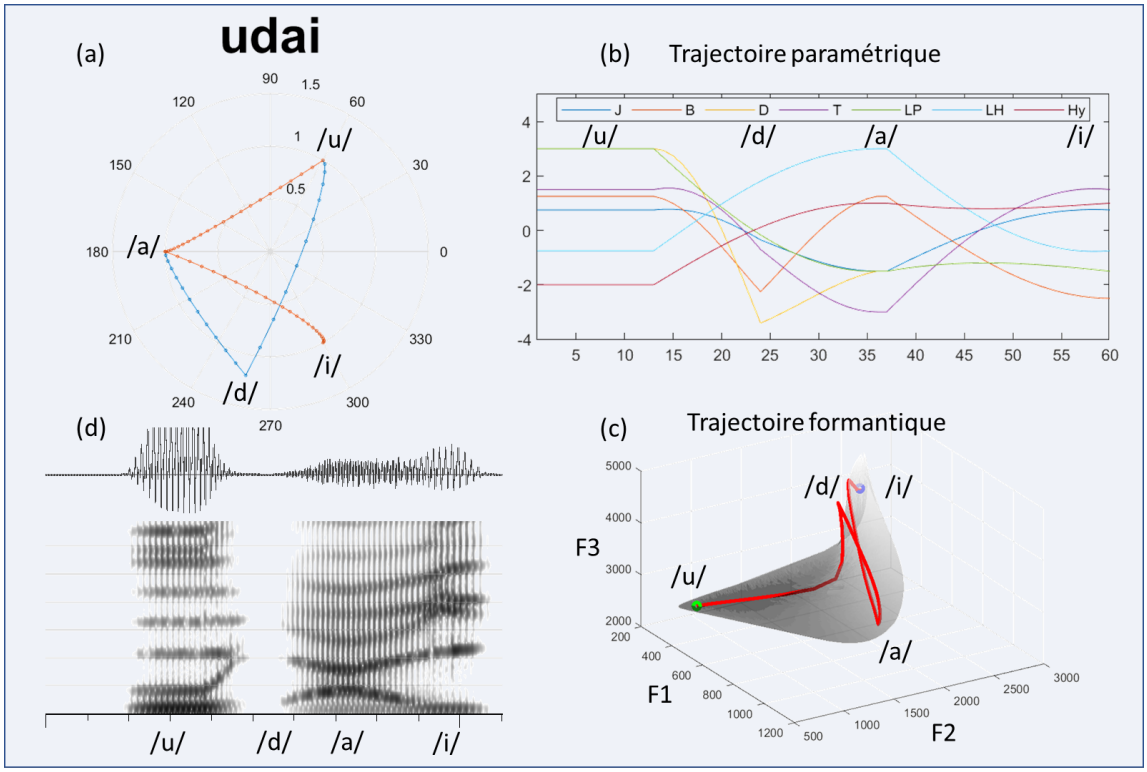


FIGURE 4 – Synthèse de la syllabe /udai/. (a) Graphe de planification syllabique avec en rouge les trajectoires vocaliques et en bleu les consonantiques. Elles relient les phonèmes situés en position périphérique du cercle unité. (b) Flux de 7 paramètres articulatoires évalués par périodes T=120 ms concaténées sans lissage (ici 5T). (c) Trajectoire des 3 premiers formants mise en rapport avec la surface vocalique. (d) Enveloppe et spectrogramme de sortie. Une modulation construite à partir de la structure syllabique et pour les consonnes plosives /bdg/ est appliquée.

PCA guidée reflètent des groupes musculaires dont le contrôle serait séparable (Maeda & Honda, 1994; Kröger & Bekolay, 2022). Très hypothétiquement, on obtiendrait ici Fig. 4b une image de leurs modulations d'activité.

5 Conclusion

Avec un dessin équivalent, combinant une planification à l'échelle syllabique et une inférence des commandes articulatoires d'un modèle, les simplifications exercées par rapport l'AP/TD (Saltzman & Munhall, 1989; Browman & Goldstein, 1992) sont nombreuses. Les relations temporelles entre gestes articulatoires étant fixées au moment de la planification syllabique, le recours à une évaluation dynamique à l'aide d'oscillateurs couplés n'est pas nécessaire (Xu, 2020). Ceci entraîne des différences de description de la structure syllabique. En ce qui concerne la TD, nous avons vu qu'il n'est pas nécessaire de réaliser une transformation pour déduire des paramètres de commande d'un modèle articulatoire à partir de variables articulatoires liées aux tâches de constriction. En effet, ces paramètres sont obtenus directement, et les trajectoires vocaliques sont pointées spatialement depuis la représentation complexe, reliant production et formes acoustiques résultantes de façon cohérente.

Références

ALEXANDER R., SORENSEN T., TOUTIOS A. & NARAYANAN S. (2019). A modular architecture for articulatory synthesis from gestural specification. *The Journal of the Acoustical Society of America*, **146**(6), 4458–4471. DOI : 10.1121/1.5139413.

BADIN P. & FANT G. (1984). *Notes on vocal tract computations*. Stl- qpsr 2-3/1984, Royal Institute of Technology, Stockholm, Sweden.

BADIN P., PERRIER P., BOË L. & ABRY C. (1990). Vocalic nomograms : Acoustic and articulatory considerations upon formant convergences. *The Journal of the Acoustical Society of America*, **87**(3), 1290–1300. DOI : 10.1121/1.398804.

BERTHOMMIER F. (2021). A mathematical model of the vowel space. DOI : 10.48550/arXiv.2111.00868.

BERTHOMMIER F. (2023). Why can big.bi be changed to bi.gbi ? a mathematical model of syllabification and articulatory synthesis. DOI : 10.48550/arXiv.2307.02299.

BIRKHOLZ P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS ONE*, **8**(4), 1–17. DOI : 10.1371/journal.pone.0060603.

BOË L.-J. & MAEDA S. (1998). Modélisation de la croissance du conduit vocal. journées d'Études linguistiques. In *La voyelle dans tous ses états*, p. 98–105.

BROWMAN C. P. & GOLDSTEIN L. M. (1992). Articulatory phonology : An overview. *Phonetica*, **49**, 155–180. DOI : 10.1159/000261913.

CARRÉ R. & CHENNOUKH S. (1995). Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. *Journal of Phonetics*, **23**(1), 231–241. DOI : 10.1016/S0095-4470(95)80045-X.

DANILOFF R. & MOLL K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, **11**(4), 707–721. DOI : 10.1044/jshr.1104.707.

DUPOUX E. (2018). Cognitive science in the era of artificial intelligence : A roadmap for reverse-engineering the infant language-learner. *Cognition*, **173**, 43–59. DOI : 10.1016/j.cognition.2017.11.008.

GAINES J. L., KIM K. S., PARRELL B., RAMANARAYANAN V., NAGARAJAN S. S. & HOUDE J. F. (2021). Discrete constriction locations describe a comprehensive range of vocal tract shapes in the Maeda model. *JASA Express Letters*, **1**(12), 124402. DOI : 10.1121/10.0009058.

KRÖGER B. J. & BEKOLAY T. (2022). Producing syllables : motor planning, motor programming and execution. In O. NIEBUHR, M. S. LUNDMARK & H. WESTON, Éds., *Studientexte zur Sprachkommunikation : Elektronische Sprachsignalverarbeitung 2022*, p. 1–8 : TUDpress, Dresden.

MAEDA S. (1979). Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10 èmes Journées d'Etude sur la Parole*, p. 152–162.

MAEDA S. (1990). Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In D. J. HARDCASTLE & A. MARCHAL, Éds., *Speech Production and Speech Modelling*, NATO ASI Series, p. 131–149. Springer Netherlands, Dordrecht. DOI : 10.1007/978-94-009-2037-8_6.

MAEDA S. & HONDA K. (1994). From emg to formant patterns of vowels : The implication of vowel spaces. *Phonetica*, **51**(1-3), 17–29. DOI : 10.1159/000261955.

MORAN S. & MCCLOY D., Éds. (2019). *PHOIBLE 2.0*. Jena : Max Planck Institute for the Science of Human History.

MRAYATI M., CARRÉ R. & GUÉRIN B. (1988). Distinctive regions and modes : A new theory of speech production. *Speech Commun.*, **7**(3), 257–286. DOI : 10.1016/0167-6393(88)90073-8.

NAM H., GOLDSTEIN L., SALTZMAN E. & BYRD D. (2004). Tada : An enhanced, portable task dynamics model in matlab. *The Journal of the Acoustical Society of America*, **115**(5), 2430–2430. DOI : 10.1121/1.4781490.

SALTZMAN E. L. & MUNHALL K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**(4), 333–382. DOI : 10.1207/s15326969eco0104_2.

SCHROEDER M. R. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *The Journal of the Acoustical Society of America*, **41**(4B), 1002–1010. DOI : 10.1121/1.1910429.

SCHWARTZ J.-L., BOË L.-J., BADIN P. & SAWALLIS T. R. (2012). Grounding stop place systems in the perceptuo-motor substance of speech : On the universality of the labial–coronal–velar stop series. *Journal of Phonetics*, **40**(1), 20–36. DOI : 10.1016/j.wocn.2011.10.004.

STORY B. H. (2009). Vowel and consonant contributions to vocal tract shape. *The Journal of the Acoustical Society of America*, **126**(2), 825–836. DOI : 10.1121/1.3158816.

STORY B. H. (2013). Phrase-level speech simulation with an airway modulation model of speech production. *Computer Speech and Language*, **27**(4), 989–1010. DOI : 10.1016/j.csl.2012.10.005.

STORY B. H. & BUNTON K. (2021). Identification of voiced stop consonants produced by acoustically driven vocal tract modulations. *JASA Express Letters*, **1**(8), 085203. DOI : 10.1121/10.0005917.

STORY B. H., VORPERIAN H. K., BUNTON K. & DURTSCHI R. B. (2018). An age-dependent vocal tract model for males and females based on anatomic measurements. *The Journal of the Acoustical Society of America*, **143**(5), 3079–3102. DOI : 10.1121/1.5038264.

XU A., VAN NIEKERK D., KRUG P., PROM-ON S., BIRKHOLZ P. & XU Y. (2023). Computational models for articulatory learning of english diphthongs : One dynamic target vs. two static targets. In *Proc. of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, p. 4140–4144.

XU Y. (2017). Syllable as a synchronization mechanism. In *Proceedings of 8th Tutorial and Research Workshop on Experimental Linguistics*, p. 9–12. DOI : 10.36505/ExLing-2017/08/0003/000305.

XU Y. (2020). Syllable is a synchronization mechanism that makes human speech possible. DOI : 10.31234/osf.io/9v4hr.

ÖHMAN S. E. G. (1966). Coarticulation in vcv utterances : Spectrographic measurements. *The Journal of the Acoustical Society of America*, **39**(1), 151–168. DOI : 10.1121/1.1909864.

ÖHMAN S. E. G. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, **41**(2), 310–320. DOI : 10.1121/1.1910340.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Traitement incrémental de la prosodie en L2

Giuseppina Turco[1] Chie Nakamura[2,] Hiyon Yoo[1]
(1) Université Paris-Cité, CNRS, Laboratoire de Linguistique Formelle,
Rue Albert Einstein, F-75013 Paris,
(2) Global Center for Science and Engineering, Waseda University, Tokyo, Japon
giuseppina.turco@cnrs.fr, cnakamura@aoni.waseda.jp, hi-yon.yoo@u-paris.fr

RESUME

Les auditeurs natifs s'appuient sur des indices prosodiques pour résoudre les ambiguïtés syntaxiques à un stade très précoce du traitement en ligne des phrases. Nous proposons de tester si un mécanisme similaire est utilisé par les auditeurs de langue seconde. En utilisant un paradigme du monde visuel, nous avons testé des phrases en anglais avec une ambiguïté d'attachement du syntagme prépositionnel avec des apprenants français d'anglais L2. L'impact de la frontière prosodique placée avant ou après le Syntagme Nominal objet a été examiné (p. ex. *The boy will write to % the panda with the crayon*, ou *The boy will write to the panda % with the crayon*). Nos résultats préliminaires montrent que les apprenants français sont capables d'intégrer l'information des frontières prosodiques pour résoudre l'ambiguïté syntaxique plus rapidement que les populations d'apprenants testées précédemment (c.-à-d. L1 japonais et L2 anglais). Cela suggère que les apprenants exploitent des indices prosodiques fins dans les décisions d'analyse syntaxique pour localiser l'information de frontière.

ABSTRACT

Native listeners rely on prosodic cues for the resolution of syntactic ambiguity at very early stage of online sentence processing. In the current study we test whether a similar mechanism is shared by second language (L2) listeners. In a visual word paradigm experiment, we used sentences with PP attachment ambiguity such as the boy will write to the panda with the crayon and tested French learners of L2 English. We examined the impact of the prosodic boundary that was placed either before or after the patient NP (e.g., *The boy will write to % the panda with the crayon*, or *The boy will write to the panda % with the crayon*). Our preliminary results show that French learners can integrate prosodic boundary information for the resolution of syntactic ambiguity faster than previously tested learner populations (i.e. L1 Japanese-to-L2 English). This suggests that learners exploit fine-grained prosodic cues in parsing decisions to locate boundary information.

MOTS-CLES : prosodie L2, frontière prosodique, traitement prosodique, oculométrie
KEYWORDS : L2 prosody, prosodic boundary, processing, eye-tracking

1. Introduction

La frontière prosodique peut être utilisée comme indice prosodique pour signaler la structure syntaxique (voir Pierrehumbert 1980 ; Shafer et al. 2000 ; pour le français, Michelas et D'Imperio 2015 et références y citées). Il a été précédemment montré que les auditeurs intègrent rapidement les informations de la frontière prosodique (telles que le ton de frontière, l'allongement de la fin de phrase, etc.) pour résoudre certaines ambiguïtés syntaxiques (p. ex. Nakamura et al. 2012). L'exemple dans (1) illustre un cas où l'ambiguïté existe à cause de la portée du syntagme

prépositionnel (SP), qui peut donner lieu à une interprétation du SP comme *instrument* ou *modificateur* du syntagme nominal (SN) objet.

The boy will write to the panda ***with the crayon***. (1)
Le garçon écrira au panda [avec le crayon]
(Interprétation de l'instrument : le garçon écrira au panda *en utilisant le* crayon)
(Interprétation du modificateur : Le garçon écrira au panda *qui a* le crayon)

Dans la présente étude, nous cherchons à savoir si les apprenants de L2 utilisent des indices prosodiques pour construire des prédictions structurelles. Certaines études sur le traitement des phrases en L2 suggèrent que les apprenants ne sont pas capables d'anticiper les informations à venir et qu'ils sont donc moins performants dans le traitement prédictif que les natifs (Grüter & Rohde 2013, Ito et al. 2018). Le paradigme expérimental de modalité visuelle a été utilisé pour tester le traitement prédictif notamment quand la prosodie entre en jeu (Ito et al. (2018) ; Nakamura et al. (2019). Contrairement à Ito et al. (2018), Nakamura et al. (2019) ont montré que les apprenants japonais L2 d'anglais étaient en effet capables de désambiguïser les structures avec un SP potentiellement ambigu en utilisant la prosodie, mais qu'ils le faisaient avec un décalage temporel par rapport aux natifs. Selon eux, un tel retard peut s'expliquer par : i) un effort cognitif accru causé par plusieurs sources d'information que les apprenants doivent traiter pendant le traitement en ligne (Hale 2006 ; Levy 2008), et ii) et un accès plus faible des apprenants aux régularités distributionnelles statistiques en raison de leur exposition limitée à la L2 (voir Farmer 2013).

L'objectif de la présente étude est donc de tester ces possibilités auprès d'une population d'apprenants dans laquelle les L1 et L2 utilisent des indices prosodiques plus similaires pour signaler l'attachement du SP. À cette fin, nous avons testé des apprenants français d'anglais L2 en adoptant le paradigme du monde visuel et les stimuli utilisés dans Nakamura et al. (2019). Bien que le français et le japonais partagent les mêmes principes phrastiques (c.-à-d., la phrase accentuelle, Jun & Fougeron 2002 pour le français, Venditti 2005 pour le japonais), ils diffèrent dans la pondération des indices prosodiques utilisés pour localiser les frontières. Le français est plus proche de l'anglais en ce sens qu'ils utilisent tous deux des modèles d'allongement final et de montée de la F0 (Beckman 1986 ; Féry 2016) bien que les deux langues présentent des différences au niveau du nombre de constituants de frontières et « scaling tonal » au sein de ces niveaux, entre autres (voir Michelas & D'Imperio 2015). En revanche, le japonais utilise un ton de frontière bas à la fin des groupes prosodiques (Pierrehumbert & Beckman 1988).

En outre, des travaux antérieurs sur le traitement des phrases par des auditeurs natifs ont montré que les attentes en matière de traitement peuvent changer en cas de variation de l'input linguistique. Les régularités statistiques disponibles dès l'input sont fortement exploitées par les auditeurs L1 afin de générer des prédictions fortes sur les informations à venir, et elles sont continuellement mises à jour lorsque l'input diffère de ce qui était attendu (cf. *adaptation linguistique,* Norris *et al.* 1995). Comme pour les apprenants japonais d'anglais L2, nous testons ici si les apprenants français d'anglais L2 sont capables de faire de telles adaptations lorsqu'ils sont confrontés à la variabilité de l'input linguistique.

1 Protocole expérimental

L'expérience 1 cherche à établir si les apprenants français d'anglais L2 utilisent les frontières prosodiques de manière incrémentale pour résoudre l'ambiguïté syntaxique liée à l'attachement du SP. L'expérience 2 cherche à savoir si les apprenants sont capables de faire des ajustements lorsque les frontières prosodiques ne s'alignent pas sur la syntaxe, c'est-à-dire si le degré auquel les

apprenants utilisent les indices prosodiques dans l'analyse structurelle est modulé par la fiabilité de la prosodie.

En utilisant un design 2x2, nous avons manipulé i) l'emplacement de la frontière prosodique (étiquetée "LH%" ci-dessous) qui a été placée *avant* (2a-3a) ou *après* (2b-3b) le modificateur (par exemple, *le panda*) et ii) la *plausibilité* (2a-b) ou *non plausibilité* (3a-b) de l'instrument (p. ex., *la gelée*). Ces deux manipulations ont donné lieu à 4 conditions, présentées dans le Tableau 1. Si les apprenants de L2 utilisent des informations sur les frontières prosodiques pour l'analyse structurelle en ligne, nous nous attendons à **i)** plus de regards vers l'objet modificateur (p. ex. *panda tenant un crayon, panda tenant de la gelée*) en entendant la prosodie du modificateur (2a-3a dans le Tableau 1), et **ii)** plus de regards vers l'objet de l'instrument (p. ex., le *crayon*) en entendant la prosodie de l'instrument (2b). En outre, nous nous attendons à **iii)** des effets « garden-path » en entendant "*gelée*" dans la condition de non-concordance (c.-à-d. 3b).

	Instrument plausible	**Instrument non plausible**
Modificateur	2a) The boy$_{L\text{-}H\%}$ will write to$_{L\text{-}H\%}$ the panda with the crayon$_{L\text{-}L\%}$.	3a) The boy$_{L\text{-}H\%}$ will write to$_{L\text{-}H\%}$ the panda with the jelly$_{L\text{-}L\%}$.
Instrument	2b) The boy$_{L\text{-}H\%}$ will write to the panda$_{L\text{-}H\%}$ with the crayon$_{L\text{-}L\%}$.	3b) The boy$_{L\text{-}H\%}$ will write to the panda$_{L\text{-}H\%}$ with the jelly$_{L\text{-}L\%}$.

Tableau 1: *Exemple d'élément utilisé dans quatre conditions. La frontière prosodique avant et après le modificateur (c.-à-d. "panda") est étiquetée avec les tons de frontière L-H%. Le ton de frontière L-L% signale la fin de la phrase.*

Les deux expériences étaient toutes les deux composées du même ensemble de 24 éléments expérimentaux, à l'exception des *fillers*. La scène visuelle contenait cinq images correspondant à des objets représentés sous forme de dessins animés. Leur position sur l'écran était contrebalancée.

Chaque énoncé a été enregistré par un locuteur natif anglais. Les items expérimentaux ont été intercalés avec 48 items de *fillers* et pseudo-randomisés sur quatre listes suivant une procédure de carré latin. Afin de tester si les apprenants de L2 sont capables de s'adapter à la fiabilité de l'information prosodique, l'Expérience 2 contenait des *fillers* produits avec une frontière prosodique placée dans des positions inattendues de la phrase (« fillers avec frontière atypique »). En d'autres termes, la frontière L-L%, habituellement utilisée en fin de phrase, a été produite entre le déterminant et le syntagme nomina (SN) final de la phrase (phrase (2) ci-dessous).

The boy$_{L\text{-}H\%}$ will touch the tie and the razor$_{L\text{-}L\%}$. (2)
(Au lieu de : *Le garçon$_{L\text{-}H\%}$ touchera $_{L\text{-}H\%}$ la cravate et le rasoir $_{L\text{-}L\%}$* utilisés dans l'Exp. 1)

FIGURE 1(A)-(B) : *Scènes visuelles - image plausible de rattachement à l'instrument (a : à gauche) et image non plausible de l'instrument (b : à droite).*

Vingt-et-un locuteurs natifs français autopositionnés de niveau B2/C1 selon le Cadre européen commun de référence pour les langues (CECR, 2021) ont participé à l'expérience 1, et 14 locuteurs natifs français à l'expérience 2. Tous les participants ont déclaré n'avoir aucune déficience visuelle ou auditive. Les expériences se sont déroulées dans la salle d'expérience au Laboratoire de Linguistique Formelle de l'Université Paris-Cité.

La tâche était d'écouter les énoncés tout en prêtant attention à la scène visuelle affichée sur l'écran de l'ordinateur. Après les instructions, l'œil dominant de chaque participant a été calibré à l'aide d'une échelle de 9 points. Les mouvements oculaires des participants ont été enregistrés à l'aide de l'Eye-link II. Pour chaque session, les éléments ont été présentés comme suit : la scène visuelle est apparue en premier et a été suivie (après 2500 ms) par l'item auditif (présentée par des haut-parleurs). Chaque session était suivie d'une question afin de s'assurer que les participants restent concentrés pendant la durée de l'expérience (30 minutes). Une correction de la dérive a été mise en place au début de chaque essai.

2 Analyse des données

Le *logit* des regards a été calculé pour chaque objet par rapport aux regards portés sur tous les objets de la scène, y compris l'arrière-plan (Barr 2008). Les analyses statistiques ont été réalisées à l'aide de modèles linéaires à effets mixtes (Baayen et al. 2008).

Nous avons analysé les regards portés sur les deux objets cibles (modificateur et instrument) dans deux fenêtres temporelles. Tout d'abord, dans la fenêtre de ce que nous appelons « fenêtre d'anticipation », nous avons analysé les regards portés sur les deux objets cibles pendant une durée de 341ms depuis l'apparition de la préposition (« with » *avec*) jusqu'à l'apparition du mot final (« crayon/jelly », *crayon/gelée*). Il s'agit de déterminer si les participants ont adopté une analyse structurelle du SP, basée sur l'emplacement des frontières prosodiques *avant de* rencontrer la cible. Si les apprenants utilisent l'emplacement des frontières pour anticiper la résolution structurelle de l'ambiguïté d'attachement du SP, des mouvements oculaires prédictifs vers l'image la plus compatible avec la structure attendue devront être observés à l'intérieur de cette fenêtre.

Ensuite, nous avons analysé les regards portés sur les deux objets cibles pendant une durée de 618 ms, depuis l'apparition du mot final de la phrase jusqu'au décalage minimal de la phrase (« fenêtre mot final de la phrase »). Dans le cas critique, les auditeurs entendent d'abord la prosodie instrumentale suivie d'un nom qui est incompatible avec une interprétation instrumentale (i.e. *crayon*), comme dans (3b). Si les apprenants anticipent une interprétation instrumentale du SP, la rencontre d'un nom d'instrument non plausible (i.e. *gelée*) dans le SP devrait générer un décalage entre l'audio et leurs attentes, retardant les regards vers le nom d'instrument non plausible, car c'est la mention de l'objet d'instrument plausible (i.e. *crayon*) qui est attendue. Autrement dit, les regards portés sur l'objet instrument dans la fenêtre temporelle du mot final de la phrase seront également analysés afin d'examiner si les apprenants rencontrent des difficultés de traitement lorsque l'attente de la structure à venir ne correspond pas à l'objet dans le SP.

Pour chaque analyse, nous avons exploré la possibilité que les participants adoptent différentes stratégies au cours de l'expérience en analysant les changements des mouvements oculaires vers chaque objet dans la première et la seconde moitié de l'expérience. Le modèle LME utilisé pour chaque analyse comprenait des effets fixes de PROSODIE (prosodie du modificateur, dorénavant, « PM » ou prosodie d'instrument « PI »), de TYPE D'IMAGE (image d'instrument plausible ou non plausible) et de BLOC (premier ou deuxième bloc de l'expérience). L'interaction des trois effets a été prise en compte. Les mêmes analyses statistiques ont été effectuées pour tester les résultats de l'expérience 2.

2.1 Résultats de l'expérience 1 (construite avec des *fillers* sans frontière atypique)

Ci-dessous, nous ne présentons que les résultats significatifs du modèle en ce qui concerne les regards vers les objets cibles (modificateur, instrument) dans les deux différentes fenêtres temporelles. Les résultats du modèle dans la fenêtre temporelle d'anticipation pour les regards vers l'objet modificateur ont montré un effet principal du BLOC (β=-2.09, SE=0.49, t=-4.29, p<0.001)

ansi qu'une interaction entre la PROSODIE et le BLOC (β=-1.29, SE=0.65, t=-1.98, p<0.05). Des analyses supplémentaires sur l'effet de la PROSODIE dans chaque bloc de l'expérience ont montré que l'effet de la PROSODIE était significatif uniquement dans la seconde moitié de l'expérience (p=0.46 dans le premier bloc, p<0.05 dans le second bloc). Cette interaction suggère que les apprenants ont regardé significativement plus l'objet modificateur (*panda tenant un crayon/une gelée* dans la Figure 1) lorsqu'on leur présentait la PM que la PI au fur et à mesure qu'ils faisaient plus d'essais dans l'expérience (Figure 2(A)).

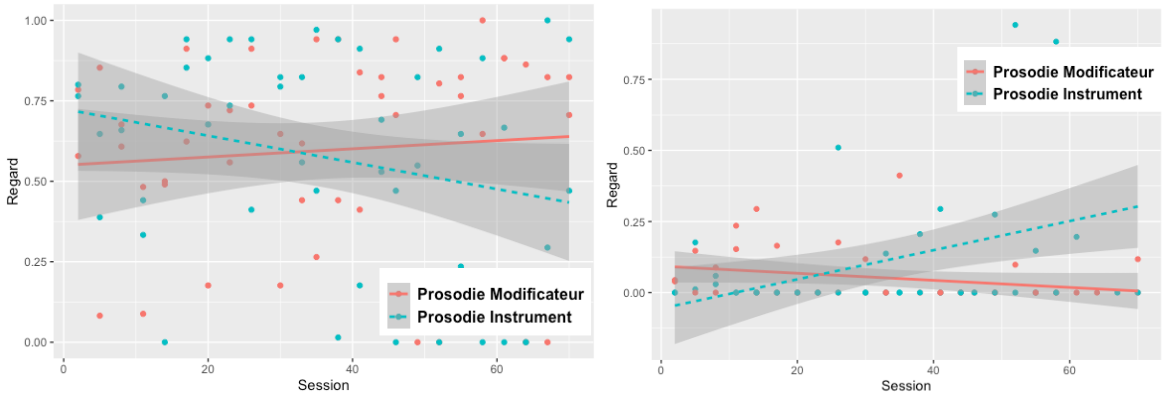


FIGURE 2(A)-(B) : Changement des regards d'une session à l'autre de l'expérience portés sur l'objet modificateur (figure de gauche) et ceux portés sur l'objet instrument (figure de droite) dans deux types de prosodie dans la fenêtre temporelle d'anticipation.

Les résultats pour les regards vers l'objet instrument (i.e. le *crayon* dans la Figure 1) ont révélé une interaction entre la PROSODIE et le BLOC (β=0.96, SE=0.41, t=2.32, p<0.05). Des analyses supplémentaires sur l'effet de la PROSODIE dans chaque bloc de l'expérience ont montré que l'effet de la PROSODIE n'était significatif que dans la seconde moitié de l'expérience (p=0.22 dans le premier bloc, p<0.05 dans le second bloc, voir Figure 2(B)). Comme pour les résultats de la Figure 2(A), cela suggère que plus ils font d'essais, plus les apprenants regardent significativement l'objet instrument lorsqu'on leur présente la PI que la PM.

Les résultats du modèle dans la fenêtre temporelle mot final de la phrase pour les regards vers l'objet modificateur ont montré un effet principal de la PROSODIE (β=-1.18, SE=0.34, t=-3.46, p<0.001) et du BLOC (β=-1.47, SE=0.41, t=-3.56, p<0.01). L'effet principal de la PROSODIE indique que les apprenants ont regardé significativement plus l'objet modificateur avec la PM qu'avec la PI (Figure 3(A)). L'effet principal du BLOC indique qu'ils regardent moins l'objet modificateur à mesure qu'ils se familiarisent avec l'expérience quel que soit le type de prosodie ou le type d'image. Les résultats du modèle pour les regards vers l'objet instrument ont montré un effet principal de la PROSODIE (β=0.53, SE=0.25, t=2.13, p =<0.05) et du TYPE D'IMAGE (β=-1.03, SE=0.25, t=-4.05, p<0.001). Comme dans la Figure 3(B), l'effet principal de la PROSODIE démontre que les apprenants ont regardé l'objet instrument plus souvent lorsqu'on leur a présenté la PI que la PM.

Comme l'illustre également la Figure 3(B), l'effet principal du TYPE D'IMAGE montre que les apprenants regardent significativement plus l'instrument avec l'image plausible de l'instrument (Figure 1(A)) qu'avec l'image non plausible de l'instrument (Figure 1(B)), quel que soit le type de prosodie. Cela indique qu'en entendant le mot final de la phrase, les participants ont regardé l'objet instrument plus souvent lorsque la phrase se terminait par un nom d'objet instrumental et que la scène visuelle contenait deux objets instrumentaux du même type (p. ex., deux *crayons de couleur* dans la Figure 1(A) par rapport à un *crayon de couleur* dans la Figure 1(B)).

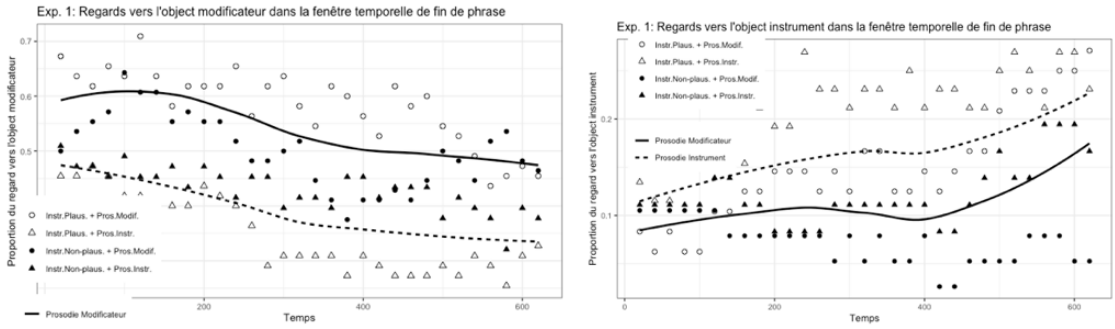


FIGURE 3(A)-(B) : Proportion de regards sur l'objet modificateur (figure de gauche) et sur l'objet instrument (figure de droite) dans la fenêtre temporelle du mot final de la phrase (620 ms à partir du début du mot final de la phrase).

2.2 Expérience 2 (construite avec des fillers avec frontière atypique)

Comme dans l'expérience 1, nous rapportons les résultats des analyses pour les regards vers les deux objets cibles dans la fenêtre d'anticipation et la fenêtre mot final de la phrase. Aucun effet n'a été constaté dans les résultats du modèle dans la fenêtre d'anticipation pour les regards vers l'objet modificateur. Les résultats pour les regards vers l'objet instrument ont montré qu'il y avait un effet principal du BLOC (β=-0.82, SE=0.26, t=-3.14, p<0.01), suggérant que les apprenants regardaient moins souvent l'objet instrument à mesure qu'ils se familiarisent avec, indépendamment du type de prosodie ou du type d'image.

Les résultats du modèle dans la fenêtre mot final de la phrase pour les regards vers l'objet modificateur ont montré un effet principal du BLOC (β=-1.12, SE=0.48, t=-2.33, p<0.05). Cela indique qu'en entendant le mot final de la phrase, les apprenants ont regardé l'objet modificateur moins souvent au cours de l'expérience, indépendamment du type de prosodie ou du type d'image. Les résultats pour les regards vers l'objet instrument ont révélé un effet principal de la PROSODIE (β=-0.58, SE=0.24, t=-2.46, p<0.05) et du BLOC (β=-0.54, SE=0.25, t=-2.18, p<0.05). L'effet principal de la PROSODIE démontre que les apprenants regardent plus l'objet instrument lorsqu'on leur présente la PM que la PI. L'effet principal du BLOC montre également que les apprenants regardent significativement moins l'objet instrument dans la seconde moitié de l'expérience par rapport à la première moitié, quel que soit le type de prosodie ou le type d'image. Cela suggère que les apprenants français L2 d'anglais ont adopté l'analyse du modificateur davantage lorsqu'ils ont entendu des phrases avec une prosodie non informative.

Afin d'explorer davantage les différents modèles des résultats observés entre les expériences 1 et 2, nous avons utilisé une analyse de permutation non paramétrique (p. ex. Dautriche et al. 2015) permettant de mieux saisir le point de divergence entre les deux types de prosodie (voir Maris & Oostenveld 2007). Nous avons effectué deux tests de permutation, l'un pour les regards vers l'objet modificateur et l'autre pour les regards vers l'objet instrument. La Figure 4(A) montre la proportion de regards vers l'objet modificateur dans les expériences 1 et 2, synchronisée avec l'apparition du SP ambigu de chaque item (c-à-d le début de la fenêtre temporelle d'anticipation). Les fenêtres temporelles où les deux types de prosodie diffèrent significativement l'un de l'autre sont indiqués par des blocs gris en haut de chaque panneau d'expérience.

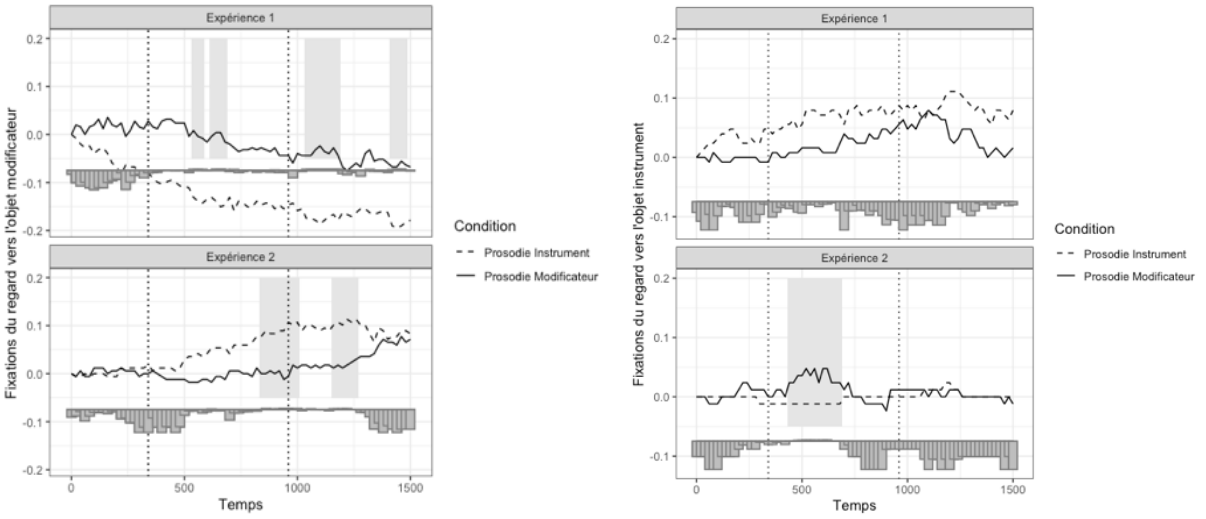


FIGURE 4(A)-(B) : Proportion moyenne de regards vers l'objet modificateur (a) dans l'exp. 1 (en haut à gauche) et l'exp. 2 (en bas à gauche), synchronisée avec l'apparition du SP ambigu (a). Proportion moyenne de regards vers l'objet instrument (b) dans les conditions instrument plausible dans l'exp. 1 (en haut à droite) et l'exp. 2 (en bas à droite). Les grands blocs gris dans chaque panneau représentent les fenêtres temporelles où les deux conditions de prosodie sont significativement différentes par l'analyse de permutation.

L'analyse de permutation a révélé que le moment des regards vers l'objet modificateur différait entre les expériences 1 et 2. Dans l'expérience 1, il y avait 4 fenêtres temporelles dans lesquels les deux types de prosodie s'écartaient significativement l'un de l'autre. Les participants à l'expérience 1 ont regardé significativement plus l'objet modificateur avec la PM qu'avec la PI dans ces fenêtres, et le point le plus précoce où ces fenêtres ont été observés était à 540 ms. Étant donné qu'il faut environ 200 ms pour que le traitement cérébral se reflète dans les mouvements oculaires, il est probable que les participants à l'expérience 1 ont commencé des mouvements oculaires anticipatifs vers l'objet modificateur dans les conditions de PM. Dans l'expérience 2, des résultats opposés ont été observés dans deux groupes de temps à des moments ultérieurs de la phrase. Dans ces fenêtres de temps, les apprenants ont regardé significativement plus l'objet modificateur avec la PI qu'avec la PM.

La Figure 4(B) montre la proportion de regards vers l'objet instrument dans les expériences 1 et 2, synchronisés avec l'apparition du SP ambigu de chaque item. On peut voir que les résultats de l'expérience 1 n'ont révélé aucune fenêtre temporelle présentant une différence significative, bien qu'il y ait une tendance générale des apprenants à regarder davantage l'objet instrument avec la PI dès le début du SP ambigu. Dans l'expérience 2, des résultats opposés ont à nouveau été observés dans une fenêtre temporelle, montrant que les apprenants regardaient davantage l'objet instrument lorsque la phrase comportait la PM. Dans l'ensemble, les analyses de permutation confirment les résultats rapportés dans l'analyse LME.

3 Discussion

L'expérience 1 révèle deux résultats principaux. D'abord, les apprenants français s'appuient sur l'information de frontière prosodique pour traiter les phrases, ce qui est conforme à nos attentes (cf. Intro). Deuxièmement, et de manière plus cruciale, l'interaction entre la prosodie (objet modificateur/instrument) et le bloc (première/deuxième moitié) dans la fenêtre temporelle d'anticipation suggère que les locuteurs français ont appris à faire des prédictions sur la bonne structure syntaxique en utilisant des indices de frontières prosodiques au fur et à mesure qu'ils

devenaient plus familiers avec l'expérience. Cette hypothèse est également confirmée par certains résultats de l'expérience 2. Le fait qu'il n'y ait aucun effet de la prosodie dans l'expérience 2 suggère que les apprenants sont sensibles à l'alignement prosodie-syntaxe ; ils ont moins utilisé l'information prosodique dans l'analyse structurelle prédictive dans les fillers où la prosodie est non informative. En revanche, dans la fenêtre temporelle mot final de phrase de l'expérience 2, il a été montré que les apprenants regardaient davantage l'objet modificateur que l'objet instrument lorsqu'ils entendaient la prosodie instrumentale. Ce résultat est contraire à ce qui a été observé dans l'expérience 1, dans laquelle les participants regardaient l'objet intentionnel lorsqu'ils entendaient les deux types de prosodie.

Dans l'ensemble, nos résultats sont conformes à ceux de l'étude de Nakamura et al. (2019) qui ont testé des apprenants japonais d'anglais L2. Les deux populations d'apprenants utilisent les informations sur les frontières prosodiques dans le traitement en ligne de l'anglais L2, ce qui suggère que des mécanismes généraux guident la résolution de l'analyse syntaxique en L2. Cependant, malgré le même niveau de compétence et d'apprentissage, les apprenants japonais ont traité les phrases avec un décalage, alors que nos participants français ont appris à prédire la structure d'attachement SP prévue dès qu'ils entendent le mot « avec » pendant l'expérience. Contrairement à d'autres travaux (voir Zhang & Ding 2022), nos résultats montrent un effet d'adaptation (p. ex. Norris et al. 1995) dans l'exploitation des connaissances préalables disponibles dans la L1 pour l'analyse structurelle prédictive. Comme mentionné précédemment, les deux paires L1-L2 (français et anglais) de nos apprenants partagent des contours ascendants et un allongement final (Tyler & Cutler 2009) signalant la présence d'une frontière à venir. Ce n'est pas le cas pour le japonais, où les phrases accentuelles peuvent être délimitées par des contours descendants (Venditti 2005). Le ton de frontière montant (L-H%) avant ou après le modificateur (par exemple *panda*) peut donc avoir aidé les apprenants français à accéder rapidement à la structure attendue au cours de l'expérience. En ce qui concerne l'expérience 2, comme pour Nakamura et al. (2019), les apprenants français ont préféré regarder davantage l'objet modificateur lorsqu'ils entendaient des phrases dont la prosodie ne correspondait pas (instrument). Ceci est peut-être dû à la manipulation d'une prosodie non informative dans les *fillers* ; les apprenants peuvent avoir adopté une stratégie où ils se fient davantage aux informations visuelles plutôt qu'aux informations prosodiques. En d'autres termes, lorsque la prosodie n'était pas alignée sur la syntaxe dans certains des items de l'expérience 2, les apprenants ont choisi de regarder l'image compatible avec deux interprétations possibles[1] plutôt que de regarder un objet qui n'est plausible que pour l'interprétation de l'instrument. Cependant, le nombre de participants étant limité à 14 dans l'expérience 2, ces résultats doivent être validés avec un échantillon plus large. Pour conclure, les résultats préliminaires de cette étude suggèrent que la connaissance préalable des indices prosodiques en L1 peut aider les apprenants à accélérer la résolution de l'ambiguïté syntaxique en L2. Nos résultats, qui montrent que les apprenants anticipent la structure à venir à l'aide de la prosodie, vont à l'encontre des recherches antérieures en L2 qui suggéraient que les apprenants en L2 ne s'engageaient pas dans un traitement *prédictif* (Ito et al. 2018). Des recherches supplémentaires permettront de montrer comment les apprenants de L2 adoptent différentes stratégies de traitement basées sur différents types d'indices prosodiques.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'IdEx Université Paris Cité (ANR-18-IDEX-0001) au titre du Labex Empirical Foundations of Linguistics - EFL

[1] Ainsi, l'image du panda tenant un crayon peut être interprétée à la fois comme « le panda qui a le crayon » et « le panda a utilisé le crayon pour écrire à quelqu'un ».

Références bibliographiques

BAAYEN, R. HARALD, DAVIDSON, D. J., & BATES, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.

BARR, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457-474.

BECKMAN, M. E. (1986). Stress and non-stress accent. Dordrecht, The Netherlands: Foris

CECRL Conseil de l'Europe, Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer – Volume complémentaire, Éditions du Conseil de l'Europe, Strasbourg, 2021.

DAUTRICHE, I., SWINGLEY, D., CHRISTOPHE A. Learning novel phonological neighbors: Syntactic category matters. *Cognition*. 2015 Oct; 143:77-86

GRÜTER, Th. & ROHDE, H. (2013). L2 processing is affected by RAGE: Evidence from reference resolution. Paper presented at the *12th conference on Generative Approaches to Second Language Acquisition (GASLA)*. University of Florida, FL.

FARMER, T. A., BROWN, M., & TANENHAUS, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences*, *36*, 211–212.

FÉRY, C. (2016). Intonation and prosodic structure. Cambridge: Cambridge University Press.

HALE, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 609– 642.

ITO, A, PICKERING M. J., & CORLEY, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1-11.

JUN, S.-A., & FOUGERON, C. (2002). Realizations of accentual phrase in French intonation. Probus 14, 147–172.

LEVY, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126– 1177.

MARIS E., OOSTENVELD R. (2007) Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neurosciences Methods*, 164, 177-190.

MICHELAS, A., D'IMPERIO, M. (2015) "Prosodic boundary strength guides syntactic parsing of French utterances" *Laboratory Phonology*, vol. 6, no. 1, pp. 119-146.

NAKAMURA, C., ARAI, M., & MAZUKA, R. (2012). Immediate use of prosody and context in predicting a syntactic structure. *Cognition*, 125, 317-323.

NAKAMURA, C., HARRIS. J., A.H., JUN, S.-., HIROSE, Y. (2019). L2 adaptation to unreliable prosody during structural analysis: A Visual World Study. Proceedings of the 43rd annual Boston University Conference on language Development (eds. Brown, M.M., Dailey, Br.). Cascadilla Press.

NORRIS, D., & J. L., MCQUEEN, J. M. & A. CUTLER (1995). Competition and segmentation in spoken-word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(5), 1209–1228.

PIERREHUMBERT, J. (1980). The phonology and phonetics of English intonation. Doctoral dissertation, MIT.

PIERREHUMBERT, J. B., & BECKMAN, M. E. (1988). Japanese tone structure. Cambridge, MA: MIT Press.

SCHAFER, A. J., SPEER, S. R., WARREN, P., & WHITE, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, 29, 169-182.

VENDITTI, J. (2005). The J-ToBI model of Japanese Intonation. In Jun, S.-A. (Ed.), Prosodic typology: The phonology and intonation of phrasing (pp. 172–200). New York, NY: Oxford University Press.

TYLER, M. D., & CUTLER, A. (2009). Cross-language differences in cue use for speech segmentation. Journal of Acoustical Society of America 126, 367–376.

ZHANG, Y., H., DING, (2022). Asymmetry in L1 and L2 listeners' use of prosody for PP-attachment disambiguation. Proceedings of Speech Prosody 2022, 659-663

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une étude exploratoire de la parole sifflée en tant que signal modulé

Liem Landri[1] Benjamin O'Brien [2] Anna Marczyk [1]
(1) Laboratoire de NeuroPsychoLinguistique, Toulouse, France
(2) Laboratoire Informatique d'Avignon, EA 4128, Université d'Avignon, Avignon, FR
liem.landri@etu.univ-tlse2.fr, anna.marczyk-buklaha@univ-tlse2.fr, benjamin.o-brien@univ-avignon.fr

RESUME

La présente étude propose une analyse comparative exploratoire entre l'espagnol parlé et sifflé (le silbo gomero) en termes du signal modulé à l'aide du MPS (spectre de puissance de modulation). Le résultat met en évidence des similarités entre ces deux modalités de la langue dans la plage des modulations spectrotemporelles lentes (1-8 Hz), associées à la compréhensibilité, tandis que des dissemblances sont observées dans la plage au-delà de 8 Hz sur l'axe temporel et 1 cyc/octave sur l'axe spectral, liées à l'intelligibilité. Ce résultat suggère que la modalité sifflée pourrait optimiser cette niche acoustique spécifique facilitant le décodage du message.

ABSTRACT

An exploratory study of whistled speech as the modulated signal. The present study provides an exploratory comparative analysis between spoken and whistled Spanish (silbo gomero) in terms of modulated signal using the MPS (modulation power spectrum). The result reveals similarities between these two language modalities in the range of slow spectrotemporal modulations (1-8 Hz), an area associated with comprehensibility, and dissimilarities in the range above 8 Hz on the temporal axis and 1 cyc/octave on the spectral axis, associated with intelligibility. This result suggests that the whistled modality may optimize this acoustic niche to facilitate message decoding.

MOTS-CLES : parole sifflée, silbo gomero, modulations spectrotemporelles.
KEYWORDS : whistled speech, silbo gomero, spectrotemporal modulations.

Actes des 35èmes Journées d'Études sur la Parole, pages 560–560.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une étude intra et inter-dialectale des voyelles du korebaju

Jenifer Vega Rodriguez[1,2], Nathalie Vallée[1], Thiago Chacón[2], Christophe Savariaux[1]
Silvain Gerber[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
(2) Univ. of Brasilia, Department of Linguistics, Portuguese and Classical Languages, Institute of Languages

jenifer-andrea.vega-rodriguez@gipsa-lab.fr, nathalie.vallee@gipsa-lab.fr, thiago_chacon@hotmail.com, christophe.savariaux@gipsa-lab.fr, silvain.gerber@gipsa-lab.fr

RÉSUMÉ

Cette étude a pour but la description des qualités vocaliques présentes dans deux variétés de korebaju, une langue tucanoane parlée dans le piémont de l'Amazonie colombienne. Les analyses acoustiques et statistiques révèlent l'absence de différences significatives entre les voyelles des deux variétés. Néanmoins, des variations liées à la génération et au genre au sein d'une même variété ont été constatées. Les résultats suggèrent que la perception d'une glottalisation plus prononcée dans la variété tama pourrait être associée à une distinction morphologique, une hypothèse actuellement en cours d'examen en prenant en considération le système tonal, la nasalisation et la morphologie. Cet article est une version améliorée et en français de celui figurant dans les actes de la conférence *2nd Annual Meeting of the Special Interest Group on Under-resourced Languages, SIGUL-ISCA* en 2023.

ABSTRACT

The aim of this study is to describe the vowel qualities of two varieties of Korebaju, a Tucanoan language spoken in the foothills of the Colombian Amazon. Acoustic and statistical analyses revealed no significant differences between the vowels of the two varieties. Nevertheless, generation- and gender-related variations within the same variety were observed. The results suggest that the perception of a more pronounced glottalization in the Tama variety could be associated with a morphological distinction, a hypothesis currently being examined by taking into consideration the tonal system, nasalization and morphology. This article is an improved version in French of the one published in the conference proceedings *2nd Annual Meeting of the Special Interest Group on Under-resourced Languages, SIGUL-ISCA in 2023.*

MOTS-CLÉS : koreguaje, tama, phonétique acoustique, variation dialectale, langue tukano.
KEYWORDS: Koreguaje, Tama, acoustic phonetics, dialectal variation, Tukanoan language.

1 Introduction

Le korebaju est une langue tonale (Gralow, 1985) parlée dans le piémont de l'Amazonie colombienne, faisant partie de la branche occidentale de la famille tucanoane (Chacón, 2016). La communauté actuelle est le résultat de l'union historique de quatre populations différentes : Korebaju, Tama, Macaguaje et Carijona, qui ont adopté le korebaju comme leur propre langue

après l'extinction de leur langue d'origine. Cependant, ces communautés cherchent à préserver leur culture d'origine à travers des diasporas culturelles. Une variation interdialectale semble être présente en korebaju car chacune de ces communautés s'est établie sur un territoire différent, bien que relevant d'une même région géographique, et se distingue des autres encore aujourd'hui par son appartenance clanique (Communauté Korebaju, 2011). De plus, selon les locuteurs de ces communautés, des variations interdialectales existent et contribuent aux facteurs d'identité liés à l'appartenance à un clan. Avec une population d'environ 2 000 locuteurs natifs (Communauté Korebaju, 2011), le korebaju est une langue en grand danger d'extinction d'après l'Atlas mondial des langues de l'UNESCO (Moseley, 2010). Les communautés Tama (Lat 1,5945, Long -75,41448) et Korebaju (Lat 1,01744, Long -75,2914) se trouvent à une heure l'une de l'autre en canoë moteur de 15 chevaux, mais elles partagent des événements culturels et des réunions organisationnelles auxquelles participent également toutes les communautés Korebaju.

La présente recherche s'inscrit dans le cadre d'une étude comparative entre les deux variétés dialectales tama et korebaju qui, selon les locuteurs, présentent des distinctions non seulement au niveau prosodique mais aussi dans la production et la distribution de la glottalisation.

1.1 Travaux antérieurs sans prise en compte de la variation dialectale

Avant nos travaux, peu d'études ont décrit les voyelles du korebaju. Dupont (s.d.) a proposé un inventaire de 6 voyelles de base /i, e, a, o, u, ɯ[1]/, une nasalisation suprasegmentale basée sur l'harmonie nasale et une glottalisation suprasegmentale en tant que conséquence de l'élision d'une voyelle longue. Herrera Casimilas (1990) a déterminé un système de 12 voyelles /i, e, a, o, u, ɯ, ĩ, ẽ, ã, õ, ũ, ɯ̃/. En suivant Dupont (1988), Cook et Criswell (2013) ont suggéré un système avec plutôt 6 voyelles de base /i, e, a, o, u, ɨ/ et une nasalisation suprasegmentale mais en reléguant la glottalisation à une occlusion glottale faisant partie de l'inventaire consonantique. Cette dernière description a souligné la présence de deux variantes dialectales (tama et korebaju) mais n'a pas fourni de données différenciées entre ces deux variantes.

1.2 Une récente investigation de la variété korebaju

Vega Rodriguez et al (2022), Vega Rodriguez & Vallée (2021) et Vega Rodriguez (2019) ont décrit la variante korebaju (koreguaje) avec un inventaire de six voyelles orales comprenant une voyelle centrale haute non arrondie /i, e, a, o, u, ɨ/, six voyelles nasales /ĩ, ẽ, ã, õ, ũ, ɨ̃/ et trois voyelles glottales /aˀ, eˀ, oˀ/ faisant partie d'un système mixte de glottalisation, segmental et suprasegmental, dépendant de la structure syllabique de la langue ainsi que du contour tonal de la voyelle précédente dans une syllabe de structure CVV.

1.3 Travaux antérieurs sur la variété tama

À ce jour, une seule enquête a permis de décrire la variante dialectale tama (TAM). Mora Cortés (2019) a proposé un inventaire phonémique de 11 voyelles, et 9 allophones correspondants : /i/ [j], /ĩ/, /e/ [ɛ], /ẽ/ [ɛ̃], /a/, /ʉ/ [ɨ] [ɤ], /ʉ̃/ [ɨ̃], /u/ [w], /ũ/, /o/ [ɔ], /õ/ [ɔ̃]. L'auteur a indiqué que : (1) les allophones des voyelles fermées non arrondies sont observés lorsqu'une consonne palatale les précède, comme indiqué par Vega Rodriguez (2019) pour les voyelles fermées postérieures et centrales dans la variante korebaju ; (2) les allophones des voyelles moyennes /e, o/ apparaissent

[1] L'auteur utilise le symbole correspondant à une voyelle centrale arrondie [ʉ] pour la description de la voyelle fermée postérieure non arrondie [ɯ]. Nous incluons ici le symbole de la charte IPA correspondant à la description articulatoire qu'en a faite l'auteur.

lorsqu'elles précèdent ou suivent la rhotique apicale [r], ou dans les syllabes accentuées ; (3) les allophones nasals sont présents dans les contextes de consonnes nasales. Cependant, cette étude n'a pas fourni d'analyse acoustique pour étayer les assimilations phonétiques observées.

Dans la présente investigation, nous proposons une description acoustique des systèmes vocaliques korebaju et tama, augmentée d'une comparaison diastratique, considérant le genre et deux générations de locuteurs.

2 Méthode

Deux terrains ont été réalisés pour collecter les données de cette étude, le premier de décembre 2021 à mars 2022, et le second de décembre 2022 à février 2023.

2.1 Participants

Vingt-quatre locuteurs natifs (12 femmes et 12 hommes), répartis à part égale dans les deux variétés TAM et COE, issus de deux générations différentes (G1 de 18 à 31 ans et G2 de 42 à 70 ans), ont participé à l'étude. Tous étaient locuteurs natifs et de descendance soit Korebaju, soit Tama. Ils avaient l'espagnol comme deuxième langue, apprise à l'école secondaire de la région et utilisée pour des échanges en dehors de la communauté Korabaju. Au moment de l'enregistrement, aucun locuteur n'avait quitté la communauté pendant plus de deux semaines.

2.2 Matériels

Un électroglottographe (EGG) D800 de Laryngograph Ltd. a été utilisé pour recueillir des données sonores (acoustiques), électrophysiologiques (EGG) et aérodynamiques (débits d'air, oral et nasal) synchronisées. L'EGG était connecté directement à un PC portable via un port USB. Le logiciel VoiceSuite 10.4.0 a été utilisé pour l'enregistrement des productions ; Praat pour la segmentation, la transcription et l'analyse des données collectées ; R pour les analyses statistiques. Les enregistrements ont été réalisés avec un microphone omnidirectionnel placé à l'intérieur du masque Oronasal Teen-Adult de Glottal Enterprise et connecté à l'EGG D800. La fréquence d'échantillonnage était de 24 kHz pour chacun des quatre canaux d'entrées (wav, EGG, flux nasal et oral). Les enregistrements ont été effectués dans un espace clos et à certaines heures de la journée pour éviter les bruits de fond et les sons atmosphériques de la forêt amazonienne.

2.3 Corpus et traitement des données

Deux listes de 118 et 145 mots, placés dans une phrase porteuse, ont été enregistrées entre 2021 et 2023, respectivement. La première liste de 118 mots a été collectée auprès de l'ensemble des 24 locuteurs. Cette liste a été conçue pour obtenir la production de paires minimales et quasi-minimales dans autant de contextes de mots que possible. La deuxième liste de 145 mots a été enregistrée auprès de 12 locuteurs (trois locuteurs de chaque variété des deux genres et des deux générations). Cette deuxième liste a été réalisée afin de compléter l'identification de paires minimales et quasi-minimales dans tous les contextes possibles parmi les locuteurs de chaque variété, de vérifier la présence ou non d'harmonie nasale et de relever les contours tonals. La consigne était donnée au locuteur de produire chaque mot à un débit de parole normal en l'insérant dans la phrase suivante :

/cɨkɨ́nà ìkámè ___ kóʔrèbàhɨ́ cɨ́òpí/
{cɨkɨ́nà ìká-mè ___ kóʔrèbàhɨ́ cɨ́òpí}
nous disons -PL ___ korebaju langue
< Nous disons____en korebaju >

Les signaux ont été segmentés et annotés manuellement au niveau des mots cibles, découpés en syllabes et en phonèmes, en utilisant, pour chaque délimitation de voyelle, le début et la fin de la partie stable du deuxième formant. Une extraction automatique de la fréquence fondamentale (f_0) et des trois premiers formants (F_1-F_3) à 30 %, 50 % et 70 % de la durée des voyelles orales a ensuite été effectuée en utilisant le logiciel Praat (Boersma, 2001).

Les analyses statistiques ont été réalisées en utilisant un modèle linéaire mixte généralisé pour chacune des variables réponses (F_1, F_2, F_3). Nous avons étudié l'impact des facteurs fixes : DIALECTE (COE et TAM), GENRE (F et H), GÉNÉRATION (G1 et G2), VOYELLE (a, e, i, o, u, ɨ), MESURE (30, 50, 70), et de leurs interactions. Le facteur *PARTICIPANT* a été introduit comme effet aléatoire. Ces modèles ont permis simultanément de tenir compte de la répétition des mesures, des variances résiduelles qui peuvent varier entre les modalités du même facteur, ainsi que des corrélations des valeurs des variables réponse entre les mesures. Nous les avons réalisés en utilisant la fonction *lme* du package *nlme* du logiciel statistique R (R Core Team, 2021).

Afin de déterminer si la glottalisation est une partie de la voyelle ou si elle est plutôt un segment à part entière et contigu, nous avons mesuré et comparé les durées des voyelles avec et sans glottalisation. Nous avons écarté la variation de la durée liée au débit de parole en divisant la durée de la voyelle par la durée du mot (variable RATIO, correspondant à la durée proportionnelle – ou relative – de la voyelle). Nous avons testé statistiquement l'impact des facteurs fixes : DIALECTE, GENRE, GÉNÉRATION, VOYELLE, et de leurs interactions, sur la variable réponse RATIO. Nous avons réalisé une régression bêta avec effet aléatoire (Cribari-Neto & Zeileis, 2010). Ce modèle nous a permis de tenir compte de la répétition de la mesure (le facteur *PARTICIPANT* a été introduit comme effet aléatoire dans le modèle) et aussi du fait que les valeurs de la variable réponse RATIO étaient, par définition, incluses dans l'intervalle [0, 1]. À cet effet, nous avons utilisé la fonction *glmmTMB* du package *glmmTMB* du logiciel statistique R.

Une fois les modèles établis, nous avons réalisé des analyses de contraste avec la fonction *glht* du package *multcomp* selon la méthode présentée par Bretz et al. (2008), en utilisant le package *emmeans* pour construire les matrices de contraste. Les différentes figures du signal acoustique, de l'EGG et du spectrogramme ont été extraites à l'aide du script Praatfig (Nguyen, 2017) et du logiciel Visible Vowels (Heeringa & Van de Velde, 2018).

3 Résultats et discussion

3.1 Voyelles orales

Notre enquête des paires minimales et quasi-minimales corrobore le statut phonologique de six voyelles orales pour les deux variétés du korebaju /i, e, a, o, u, ɨ/ décrites par Vega Rodriguez (2019).

La Figure 1 présente la distribution de l'ensemble des voyelles orales (avec normalisation de Lobanov) dans l'espace acoustique des deux formants F_1-F_2 pour les locutrices et locuteurs des deux variétés COE et TAM. Chez les femmes des deux variétés, une centralisation de la voyelle mi-fermée antérieure non arrondie [e] est évidente, donnant un allophone mi-central [ə] qui semble être en variation libre. Un phénomène de centralisation est également observé dans les productions

de la voyelle haute antérieure non arrondie /i/ chez les locutrices COE. Chez les hommes, les réalisations centralisées de la voyelle antérieure mi-ouverte /e/ apparaissent de même que chez les locutrices. L'allophone antérieur de la voyelle /ɨ/ est moins évident chez les hommes de la variété COE que chez ceux de la variété TAM. Une grande variabilité acoustique est observée dans la production de la voyelle ouverte /a/ pour l'ensemble des 24 locuteurs. Une tendance à la rétraction de la voyelle /a/ est observée chez tous les participants.

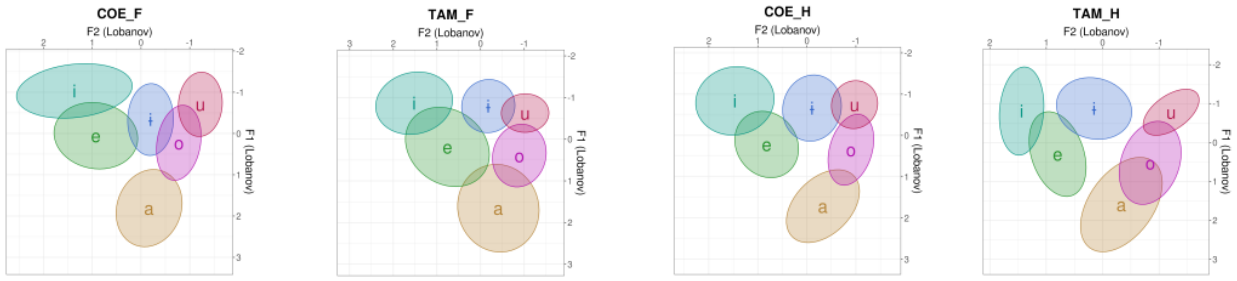


FIGURE 1: Espaces acoustiques avec normalisation de Lobanov pour les six voyelles orales produites par 6 femmes (deux premières figures à gauche) et 6 hommes (deux dernières figures à droite) de chaque variété (COE et TAM).

Devant les consonnes palatales, nous avons observé une antériorisation de la voyelle fermée centrale non arrondie /ɨ/ qui se présente comme un allophone proche de la voyelle haute antérieure non arrondie [ɪ]. Le même phénomène est observé pour la voyelle fermée postérieure arrondie /u/ réalisée comme une voyelle fermée postérieure arrondie [ʊ] (TABLE 1). Ce deuxième allophone n'avait pas été repéré par Vega Rodriguez (2019).

	F_1 (SE)	**F_2 (SE)**	**F_3 (SE)**
COE-F-ɪ	522 (121)	1819 (124)	2936 (590)
COE-M-ɪ	414 (121)	1772 (122)	2588 (597)
TAM-F-ɪ	610 (117)	1707 (163)	3015 (627)
TAM-M-ɪ	479 (117)	1835 (146)	2948 (628)
COE-F-ʊ	577 (127)	1182 (119)	2416 (652)
COE-M-ʊ	514 (129)	1203 (103)	2819 (698)
TAM-F-ʊ	542 (118)	1137 (165)	2661 (622)
TAM-M-ʊ	471 (138)	1200 (90)	2398 (710)

TABLEAU 1 : Valeurs moyennes (et écart-type) de F_1, F_2, F_3 (en Hz) pour les allophones [ɪ] et [ʊ], pour les femmes et les hommes des deux générations TAM et COE.

Dans l'ensemble, aucune différence significative n'a été observée entre les deux variantes linguistiques, ni entre les deux générations de chaque variété. Cependant, des différences significatives ont été relevées entre les hommes et les femmes de la même génération et de la même variété. La FIGURE 2 illustre, pour les deux générations COE, les différences de valeurs de F_1 pour les voyelles [a], [e], [i], [ɨ] [o], et [u], mesurées à 30 %, 50 % et 70 % de la durée de la voyelle.

En outre, une différence significative aux trois points (30 %, 50 % et 70 %) de la durée de la voyelle entre les hommes et les femmes est observée pour le premier formant F_1 des voyelles [a], [e], [i], [ɨ], [o] de COE G2, ainsi que pour la voyelle [a] de COE G1 et TAM G1. Enfin, une différence de genre est évidente pour F_1 de la voyelle [u] de TAM G2 (TABLE 2 et FIGURE 3).

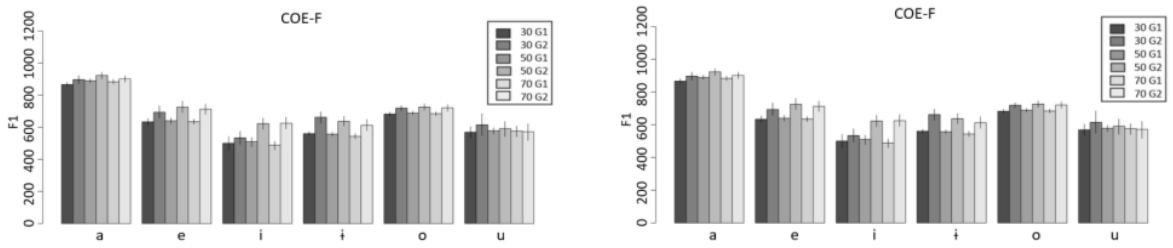


FIGURE 2 : Valeurs moyennes (en Hz, avec intervalles de confiance) de F_1 mesurées à 30 %, 50 % et 70 % de la durée de la voyelle pour les femmes et les hommes COE des deux générations.

	[a]	**[e]**	**[i]**	**[ɨ]**	**[o]**	**[u]**
COE G1	4,70 **(<0,01)**	1,17 (1)	2,88 (0,45)	1,41 (1)	3,87 (0,21)	2,87 (0,47)
TAM G1	4,03 **(<0,01)**	1,63 (1)	1,19 (1)	1,62 (1)	2,21 (0,97)	0,07 (1)
COE G2	7,61 **(<0,01)**	6,69 **(<0,01)**	6,57 **(<0,01)**	5,83 **(<0,01)**	5,53 **(<0,01)**	2,77 (0,56)
TAM G2	3,58 (0,61)	2,51 (0,81)	2,74 (0,60)	1,70 (1)	3,39 (0,12)	4,51 **(<0,001)**

TABLEAU 2 : Valeurs Z (et P) pour F1 mesurées à 50 % de la durée de la voyelle, entre les femmes et les hommes des deux générations et variétés TAM et COE.

Aucune différence significative n'a été trouvée entre les formants F2 et F3 pour aucune voyelle dans aucune variété, genre ou génération.

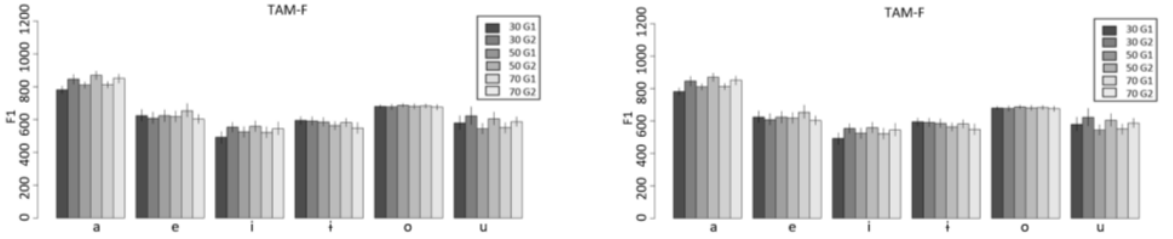


FIGURE 3 : Valeurs moyennes de F_1 (en Hz) et intervalles de confiance mesurés à 30 %, 50 % et 70 % de la durée de la voyelle pour les femmes et les hommes de TAM des deux générations

Dans l'ensemble, la principale différence trouvée pour F_1 entre les hommes et les femmes COE G2 est de + 200 Hz pour l'ensemble des voyelles orales /i, e, a, a, o, u, ɨ/. Pour les locuteurs TAM G1, les valeurs moyennes de F_1 présentent quelques différences de + 120 Hz à + 130 Hz, mais seulement pour les réalisations de la voyelle /a/ puisque les voyelles de TAM G1 ont des valeurs de formants approximativement plus homogènes.

Notre analyse n'a pas trouvé d'allophone pour les voyelles mi-fermées /e, o/ en contexte rhotique tel que décrit par Mora Cortés (2019) pour la variété TAM. Cependant, notre étude atteste d'un allophone en variation libre pour la voyelle antérieure non arrondie /e/ correspondant à la voyelle mi-centrale [ə] (TABLE 2).

	F_1 (SE)	F_2 (SE)	F_3 (SE)
COE-F-ə	655 (102)	1416 (324)	2407 (457)
COE-M-ə	530 (102)	1591 (101)	2312 (422)
TAM-F-ə	596 (101)	1429 (179)	2473 (358)
TAM-M-ə	561 (101)	1614 (179)	2305 (358)

TABLEAU 3 : Valeurs moyennes (et écart-type) de F_1, F_2, F_3 (en Hz) pour la voyelle [ə], pour les femmes et les hommes TAM et COE des deux générations.

3.2 Voyelles glottales

Les analyses des durées des voyelles n'indiquent aucune différence significative entre les voyelles non glottalisées et les voyelles glottalisées de la variété COE, corroborant les résultats de Vega Rodriguez (2019). La FIGURE 4 présente les durées relatives des voyelles modales et de leurs correspondantes glottalisées de la variété TAM. Bien qu'il existe une tendance des voyelles glottalisées à être plus longues, cette différence n'est pas significative pour l'ensemble des voyelles. Ceci, de la même manière que les paires minimales identifiées, suggère de considérer la glottalisation comme une caractéristique de la voyelle dans la description de cette variété. À noter également l'absence de réalisations glottalisées de /u/ chez les femmes TAM, comme mentionné plus haut.

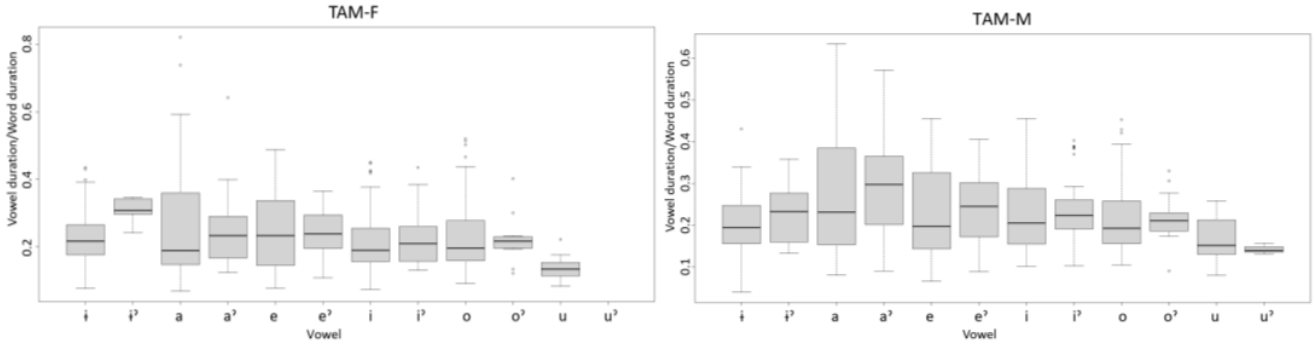


FIGURE 4 : Médiane, quartiles inférieur et supérieur de la variable RATIO pour les femmes TAM (à gauche) et les hommes TAM (à droite).

Notre enquête a fourni un ensemble de 5 voyelles glottalisées phonologiques pour les deux variantes linguistiques /iˀ, eˀ, aˀ, oˀ, ɨˀ/. Une voyelle arrondie postérieure glottalisée [uˀ] a été produite de manière aléatoire par les locuteurs des deux générations et a étonnamment été omise par toutes les femmes TAM. Ce phonème peut survenir dans n'importe quelle position, mais son statut phonologique n'a pas encore été confirmé. Notre enquête soutient un contraste phonologique entre les voyelles périphériques modales et glottalisées, comme suggéré par Vega Rodriguez et al. (2022), Vega Rodriguez & Vallée (2021) et Vega Rodriguez (2019), à partir de l'analyse de paires minimales sur la première ou deuxième syllabe de racines lexicales ainsi que dans certains affixes. Cependant, elle fournit également des preuves d'un statut phonologique pour les voyelles glottalisées hautes /ɨˀ/ (a) et /iˀ/ (b) qui n'avaient pas pu être démontrées dans des études antérieures sur la variante korebaju.

a) /sɨsɨ́/ /sɨˀsɨ́/
{sɨsɨ} {sɨˀsɨ́}
< Sanguinus Mistax > < Opossum Commun >

b) /sìsíà/ /sìˀsí-á/
{sìsí-à} {sìˀsí-á}
apophyse mastoid-CL dirty-CL
< apophyse mastoid > < sale Bactris Gasipaes >

3.3 Voyelles nasales

Une expertise visuelle de la structure spectrale des voyelles a montré que les fréquences des formants nasals appartiennent à la syllabe contenant la voyelle nasale et non à tout le mot. Par conséquent, notre enquête n'a pas trouvé d'harmonie nasale affectant l'ensemble du mot. À ce stade, la nasalité semble limitée, au mieux, au domaine de la syllabe.

De même, cette étude rapporte 6 voyelles nasales [ĩ], [ẽ], [ã], [õ], [ũ], [ɨ̃]. Cependant, aucune paire minimale n'a été trouvée pour démontrer leurs propriétés contrastives. Les paires minimales données dans des études antérieures montrent des changements au niveau du ton ou de la glottalisation qui apparaissent dans certains contextes, comme le montrent les exemples c, d, e et f.

c)	[pɪ́ɪ̃̀ã]	[pɪ̃̀ʔa]	d)	[mɪ̃̂:]	[màɪ̯́á]	f)	[cɪ́ɪ̃̀õ]	[cɪ́ʔɪ̃̀õ]	g) [cáɪ̃́]	[càʔí]
	< piment >	< oiseau>		< perroquet >	< chemin >		< fille >	< culture >	< jaguar >	< liane_yare >

Considérant que le korebaju est une langue tonale et que la glottale intervocalique est encore en cours d'investigation car il n'y a pas de consensus sur son statut segmental ou suprasegmental, ni dans les descriptions précédentes du korebaju ni dans les descriptions d'autres langues de la famille tucanoane (Sorensen, 1969; Klumpp et Klumpp, 1973; Miller, 1999 ; Vallejos, 2013 ; Bruil, 2014 et Stenzel, 2007), de tels paires de mots ne peuvent pas être catégorisées comme des paires minimales qui pourraient distinguer les voyelles orales et nasales phonémiques en korebaju dans l'une ou l'autre variété.

4 Conclusion

Notre analyse n'a trouvé aucune différence interdialectale entre les variétés COE et TAM. Des différences intradialectales ont pu être observées au niveau du genre pour certaines générations et certaines voyelles.

Nos résultats confirment que la glottalisation semble faire partie de la voyelle, comme proposé par Vega Rodriguez (2019) et Vega Rodriguez et al. (2022), bien que son statut en tant que caractéristique articulatoire segmentale ou suprasegmentale soit encore peu clair. Nous avons observé une tendance à ce que les voyelles glottales soient plus longues, mais cette différence n'est pas significative et ce pour l'ensemble des voyelles. De plus, cette recherche a identifié deux phonèmes correspondant aux voyelles glottales fermées /ɨˀ/ et /iˀ/.

Enfin, cette étude suggère que la possible cause perceptuelle d'une forte glottalisation pour les locuteurs de la variété TAM pourrait être due à un changement morphologique de certains mots comme le mot « étroit », où une possible insertion d'un prédicat copulatif à la deuxième syllabe dans la variété TAM crée la condition d'une resyllabification du mot [má-ɪ̯̃-àˀ-kà-rɨ] {CL-COP-étroit-CL}, tandis que la variété COE produira une modulation tonale sans insertion de la même copule [mâˀ-kà-rɨ] {CL-étroit-CL}.

Ces résultats font toujours l'objet d'étude. Nos recherches en cours examinent la relation entre le ton et la nasalisation, et élaborent une typologie des glottalisations observées dans la langue.

Références

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

BRETZ, F., HOTHORN, T., & WESTFALL, P. (2008). Multiple comparison procedures in linear models. In Proc. Compstat 2008 Computational Statistics (pp. 423-431). Physica-Verlag HD.
CHACÓN, T. (2016). The Reconstruction of Laryngealization in Proto-Tukanoan. In COLER, M., AVELINO, H., & WETZELS, W. L. The Phonetics and Phonology of Laryngeal Features in Native American Languages, 258-284. Leiden: Brill.
COOK, D., & CRISWELL, L. (2013). La langue Koreguaje (Tukano Occidental). SIL. Lomalinda : Éditions Townsend.
COMMUNAUTÉ KOREBAJU. (2011). Proposition du modèle pédagogique korebaju. Caquetá, Colombie. (Manuscrit non publié).
CRIBARI-NETO, F., & ZEILEIS, A. (2010). Beta Regression in R. Journal of Statistical Software, 34(2), 1-24.
DUPONT, C. (s.d.). La Langue Koreguaje (Tukano Occidental). Phonologie et Morphologie. Manuscrit non publié.
DUPONT, C. (1988). Armonía Nasal en la Lengua Koreguaje (Tukano Occidental), dans Cuadernos de Lingüística Hispánica, 2, N. 1: 105-125. Tunja. Universidad Pedagógica y Tecnológica de Colombia.
GRALOW, F. (1985). The coreguaje suprasegmental system: tone, stress and intonation. In BREND, R. (ed), Phonology to discuss: Studies in six Colombian languages. Lenguaje Data, Amerindian series 9, 3-11. Dallas: SIL.
HEERINGA, W., & VAN DE VELDE, H. (2018). Visible Vowels: a Tool for the Visualization of Vowel Variation. In Proc. CLARIN Annual Conference 2018, Pisa, Italy.
HERRERA CASIMILAS, G. E. (1990). Manuel de prononciation espagnole pour locuteurs koreguajes basé sur l'analyse contrastive au niveau phonologique des deux langues [Spanish Pronunciation Manual for Koreguaje Speakers Based on Contrastive Phonological Analysis of the Two Languages]. Mémoire de licence, Universidad Nacional de Colombia, Bogotá, Colombie.
LARYNGOGRAPH LTD. (2021). VoiceSuite 10.4.0. London.
LOBANOV, B. M. (1971). Classification of Russian vowels spoken by different speakers. Journal of the Acoustical Society of America, 49(2), 606–608.
MOSELEY, C. (ed.). 2010. Atlas des langues en danger dans le monde, 3ème édition. Paris: Éditions UNESCO.
MORA CORTÉS, L. E. (2019). Reconocimiento del pueblo Tama. Descripción fonológica de su variante lingüística. Cali: Programa Editorial Universidad del Valle.
NGUYEN, M.C. (2017). Script Praatfig. Github.
https://github.com/MinhChauNGUYEN/praatfig?tab=readme-ov-file.
R CORE TEAM. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
STENZEL, K. (2007). Glottalization and other suprasegmental features in Wanano. International Journal of American Linguistics, 73, 331-366.
VEGA RODRIGUEZ, J. & VALLÉE, N. (2021). Glottal Sounds in Korebaju. INTERSPEECH 2021. (pp. 1011-1014). ISCA and Brno University of Technology. Brno, Czech Republic. ISCA grant awardee. ⟨hal-03337770⟩ [Communication orale]
VEGA RODRIGUEZ, J. (2019a). The Vowel System of Korebaju. INTERSPEECH 2019 (pp. 3975-3979). Graz, Austria. ⟨hal-02420035f⟩ [Communication orale]

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Une nouvelle grammaire de l'intonation de la phrase française

Philippe Martin
LLF, UFRL, Université Paris Cité
Place Paul Ricoeur, 75013 Paris, France
philippe.martin@utoronto.ca

RESUME

On propose une nouvelle grammaire prosodique de l'intonation de la phrase en français. Cette grammaire rassemble des règles de réécriture opérant sur des événements prosodiques alignés sur les voyelles des syllabes accentuées des groupes accentuels, en position finale en français.

Ces règles définissent les séquences bien formées d'évènements prosodiques de la phrase, quelle que soit sa complexité, en s'assurant que les conditions nécessaires et suffisantes pour indiquer sans ambiguïté une structure prosodique donnée soient remplies. Elles remettent en cause, entre autres, l'annotation d'un ton H* ou LH* aligné sur la frontière droite des syntagmes intonatifs intermédiaires ip.

ABSTRACT

A new grammar of French sentence intonation
A new prosodic grammar of sentence intonation in French is proposed. This grammar gathers rewriting rules operating on prosodic events aligned on the vowels of stressed syllables of accent phrases, in final position in French.

These rules define well-formed sequences of prosodic events in the sentence, whatever its complexity, ensuring that the necessary and sufficient conditions for unambiguously indicating a given prosodic structure are met. Among other things, they call into question the annotation of an H* or LH* tone aligned with the right boundary of ip intermediate intonational phrases.

MOTS-CLES : Grammaire prosodique, français, groupe accentuel, structure prosodique
KEYWORDS : Prosodic grammar, French, accent phrase, prosodic structure

1 Introduction

L'intonation de la phrase française, analysée traditionnellement en termes de contours (Coustenoble et Armstrong, 1934 ; Delattre, 1966 ; Vaissière, 1974 ; Ph. Martin, 1975 ; Léon, 1993), a fait depuis plus de 40 ans l'objet d'analyses impliquant l'annotation par cibles tonales ToBI, analyses menées pour la plupart dans le cadre théorique autosegmental-métrique (Hirst et Di Cristo, 1984 ; Mertens, 1987 ; Post, 1999 ; Jun et Fougeron, 2002 ; D'Imperio et al., 2016 ; Michelas, 2011 ; Delais, Post et

Yoo, 2020, parmi d'autres). La majorité de ces recherches considèrent implicitement ou explicitement que les événements prosodiques procèdent d'un mapping réalisé à partir de l'organisation morphosyntaxique de la phrase, et en particulier des frontières et des catégories syntaxiques.

En inversant l'ordre des opérations, c'est-à-dire en considérant que la structure syntaxique résulte d'une insertion des unités syntaxiques dans la structure prosodique et non le contraire, on est conduit à élaborer une grammaire qui rende compte des séquences bien formées des évènements prosodiques advenant à l'endroit des voyelles des syllabes accentuées et indiquant une structure prosodique donnée, indépendamment de la morphosyntaxe et de constructions syntaxiques particulières.

2 Grammaire prosodique

Si la structure prosodique est considérée comme autonome par rapport à la morphosyntaxe et à toute autre structure de la phrase, en plus de générer des séquences tonales bien formées, une grammaire prosodique devra aussi rendre compte du rôle et du fonctionnement de ces séquences tonales dans l'indication des regroupements successifs en plusieurs niveaux des groupes accentuels, possiblement mais non nécessairement étiquetés selon le modèle Autosegmental-Métrique, en AP (*accent phrases*) en syntagmes intonatifs intermédiaires ip (*intermediate intonation phrases*), et des ip en syntagmes intonatifs IP (*Intonation Phrases*), et enfin des IP en structure prosodique PS (*Prosodic Structure*), un groupe accentuel étant constitué d'une groupe de mots dont un seul présente une syllabe accentuée (non-emphatique), en position finale en français.

3 Classes d'évènements prosodiques

Les variations mélodiques à l'endroit des voyelles des syllabes accentuées apparaissent comme paramètre à retenir pour décrire des événements prosodiques de par leur position dans le groupe accentuel et par leur impact perceptif, contrairement aux syllabes non accentuées. Ce choix est de plus validé par des recherches neuro-perceptives récentes, qui suggèrent que les variations mélodiques sont encodées comme des catégories discrètes et contrastives dans le cerveau de l'auditeur (Llanos et al., 2021).

Les caractéristiques acoustiques doivent tenir compte, ne fût-ce qu'approximativement, de leur perception par les auditeurs. Il s'agit en tout cas d'éviter que la courbe acoustique de fréquence fondamentale soit utilisée telle quelle comme seule source de données, et qu'une montée de 5 ou 10 Hz soit transcrite par un ton haut au même titre qu'une montée de 50 Hz par exemple.

Le système de notation choisi est basé sur les cibles tonales ToBI adaptées au français (Delais et al. 2015), et utilisent les classes suivantes :

a. **L*L%↓** : Séquence tonale terminale déclarative

b. **H*H%↑** : Séquence tonale terminale interrogative

c. **LH*↗** : Séquence tonale non-terminale montante

d. **HL*↘** : Séquence tonale non-terminale descendante

e. **H*—** : Montée ou descente mélodique faible perçue comme ton statique, donc inférieure au seuil de perception d'une variation mélodique (seuil de glissando, Rossi, 1971).

f. **H*L#↘** : Séquence tonale non-terminale descendante devant pause. Ce contour, peu ou jamais cité dans la littérature, correspond au contour « de dictée », descendant devant pause, noté H*L#↘, utilisé fréquemment dans le discours politique, et caractérisant le syntagme intonatif IP. Cette séquence est en distribution complémentaire avec la séquence montante LH*↗ (voir exemple ci-dessous).

4 Séquences attestées

Les règles de la grammaire prosodique sont établies à partir des séquences tonales attestées d'un grand nombre d'enregistrements de parole lue et spontanée en français, extraits des corpus SIWIS (5340 phrases lues par 8 locuteurs) et ORFEO (1373 enregistrements spontanés totalisant 4.302.939 mots), complétés par d'autres enregistrements.

Inventaire des séquences attestées de structures prosodiques déclaratives ne comprenant que deux groupes accentuels :

4.1 Déclarative [X - $Y_{L*L\%\downarrow}$] (terminal conclusif déclaratif) :

[**H*— L*L%↓**] attesté : [(*c'est déjA*) H* (*très difficIle*) H* (*pour les autoritÉs*)] L*L% (SIWIS fr_a1_08_008).

*[**HL*↘ L*L%↓**] non attesté dans les phrases déclaratives, **HL*↘** ne peut apparaitre que suivi de **LH*↗** (voir ci-dessous).

[**LH*↗ L*L%↓**] attesté : [(*la clAsse*) LH* (*gAIE*) LH* (*montre le frEIN*)] L*L% (SIWIS fr_a1_08_209)

[**H*L#↘ L*L%↓**] attesté : [(*mais fOrce*) H* (*est de constatER*)] H*L# [(*que la FrANce*) H* (*n'irait pas lOIN*)] H*L# [(*avec le programme socialIste*)] L*L% (F. Fillon, radio). Contours de dictée.

On peut donc avoir plusieurs séquences tonales possibles terminant le premier groupe accentuel, et pas seulement un contour H* montant ou descendant, mais de faible variation et perçu comme un ton statique, terminant un AP dans [A_{AP} L*L%↓], par exemple

[(*le sujEt*) H* (*est très importANt*)] L*L% (fr_b2_06_012, locuteur 06), aussi bien que

[(*le sujEt*) LH* (*est très importANt*)] L*L% (fr_b2_10_012, locuteur 10).

4.2 Déclarative [$Y_{L*L\%\downarrow}$ - X] (terminal conclusif déclaratif) :

*[**L*L%↓ HL*↘**], *[**L*L%↓ LH*↗**] et *[**L*L%↓ H*L#↘**] sont non attestés.

[**L*L%↓ H*—**] attesté : [(*Je vOIs*)] L*L% [(*ce que vous voulez dIre*)] H* (SIWIS C1_12_301)

Cette dernière séquence correspond à une configuration « propos-thème », ou de « focalisation large », s'opposant à [(*Je vOIs*) LH* (*ce que vous voulez dIre*)] L*L% ou à [(*Je vOIis*) H* (*ce que vous voulez dIre*)] L*L%

[**L*L%↓ L*L%↓**] [(*le mot bordEl*) LH* (*c'est du regIstre*) LH* (*populAIre*)] L*L%
[(*comme dIt*) H* (*l'AcadémIe*) H* (*françAIse*)] L*L% (E. Macron, Maubeuge 8/11/18). Cas de deux structures prosodiques indépendantes associées à une seule structure syntaxique (cf. complément rapporté, Bally, 1944).

4.3 Interrogative [X – $Y_{H*H\%\uparrow}$] (terminal conclusif interrogatif) :

[**H*— H*H%↑**] attesté : [(*et ces end**I**ves*) H* (*si blA**N**ches*) H* (*si pr**O**pres*)] H* [(*d'**OÙ***) H*H% (*elles s**O**rtent ?*)] H*H% (SIWIS fr_c_22_195).

[**LH*↗ H*H%↑**] attesté : [(*la CG**T***)] LH* [(*acceptera-t-**E**lle*) H* (*de ne plus tourner le d**O**s*)] LH* [(*à la constructi**ON***) HL* (*europé**E**nne*)] H*H% *?* (SIWIS fr_b_29_194).

[**HL*↘ H*H%↑**] attesté : [(*y a-t-**I**l*) HL* (*des observati**ON**s ?*)] H*H% (SIWIS FR_C1_12_000)

4.4 Interrogative [$Y_{H*H\%\uparrow}$ - X] (terminal conclusif interrogatif) :

*[**H*H%↑ H*—**] non attesté.

[**H*H%↑ H*H%↑**] est attesté : (*C'est bien t**OI***) H*H% (*ma jol**I**e ?*) H*H% (Billières, 2018). Il s'agit ici de différentier la copie du contour terminal interrogatif dans une configuration propos-thème (type (*il est ven**U***) H*H% (*le fact**EU**r ?*) H*H%), de celle impliquant deux structures prosodiques interrogatives indépendantes (type *il est ven**U** ?* **H*H%↑** *Mais à quelle h**EU**re ?* **H*H%↑**). Ce problème a été traité à partir d'observations expérimentales (Ph. Martin, 2008), concluant à une différentiation basée sur l'absence d'un intervalle de moins de 250 ms dans le cas d'une copie de **H*H%↑** dans une configuration propos-thème, en plus d'avoir une montée mélodique plus haute sur le premier contour. Deux structures prosodiques interrogatives indépendantes sont donc séparées par une pause d'au moins 250 ms.

4.5 Séquences déclaratives complexes incluant H*—, HL*↘, LH*↗, H*L#↘

[HL*↘ - LH*↗] : on n'a jamais de contour descendant **HL*↘** sans qu'il soit suivi d'un contour montant **LH*↗** avant le contour terminal déclaratif **L*L%↓** : [**HL*↘ LH*↗ L*L%↓**] mais *[**HL*↘ L*L%↓**] : [(*Nous refus**ON**s*) HL* (*cette réf**O**rme*)] LH* [(*antisoci**A**le*)] L*L% (SIWIS A1_19_010)

[HL*↘ - H*L#↘] : séquences attestées [**HL*↘ H*L#↘ L*L%↓**] mais *[**H*L#↘ HL*↘ L*L%↓**].

[H*L#↘ (le contour de la dictée) **- LH*↗]** : on trouve des séquences [**H*L#↘ LH*↗ L*L%↓**] aussi bien que [**LH*↗ H*L#↘ L*L%↓**], ce qui permet de conclure que **LH*↗** et **H*L#↘** sont en distribution complémentaire : [(*sur la sc**È**ne*) HL* (*internation**A**le*)] H*L# [(*si je suis él**U**e*)] H*L# [(*seront s**IM**ples*) HL* (*et fid**È**les*)] H*L# [(*à notre vocati**ON***) H* (*la plus h**A**Ute*)] L*L% (S. Royal, radio).

Reste **H*—** qui peut être suivi de **L*L%↓**.

4.6 Séquences interrogatives complexes incluant H*—, HL*↘, LH*↗, H*L#↘

On observe dans les phrases interrogatives terminées par **H*H%↑** : [**HL*↘ H*H%↑**], tout comme [**LH*↗ H*H%↑**] ainsi que [**H*— H*H%↑**], mais jamais *[**H*L#↘ H*H%↑**].

4.7 Classement des cibles tonales et des catégories de syntagmes intonatifs

Puisque **L*L%↓** terminal conclusif est le plus proéminent et toujours présent, et que **HL*↘** est toujours suivi de **LH*↗**, on a **HL*↘** < **LH*↗** : HL*↘ est dominé par LH*↗ puisqu'il ne peut apparaitre sans lui, LH*↗ précède donc HL*↘. D'autre part, **H*L#↘** est perceptivement plus saillant que **LH*↗**, mais ces deux séquences tonales étant en distribution complémentaire sont

au même niveau. Enfin, **H*—** est acoustiquement le moins marqué, car perçu comme un ton statique. La hiérarchie des contours est donc **H*—** < **HL*↘** < **LH*↗**, **H*L#↘** < **L*L%↓** pour les phrases déclaratives, et **H*** < {**LH*↗**, **HL*↘**} < **H*H%↑** pour les phrases interrogatives.

4.6 Construction de la structure prosodique

Les règles de réécriture doivent rendre compte des séquences de cibles tonales bien formées :

Continuation majeure : **L*L%↓** ⇨ {**LH*↗**, **H*L#↘**} **L*L%↓**

Continuation mineure déclarative : {**LH*↗**, **H*L#↘**} ⇨ **HL*↘ LH*↗** / **L*L%↓**

Contour neutralisé : **L*L%↓** ⇨ **H*** **L*L%↓** ; {**LH*↗**, **H*L#↘**} ⇨ **H*** **LH*↗** ; **HL*↘** ⇨ **H*** **HL*↘**

Continuation majeure interrogative : **H*H%↑** ⇨ **HL*↘ H*H%↑** ; **H*H%↑** ⇨ **LH*↗ H*H%↑**

4.7 Attribution des séquences tonales

La figure 1 présente la hiérarchie des séquences tonales marquant une structure prosodique marquée des catégories PS, IP, ip et AP, bien que la PS résulte de regroupements des mots prosodiques (AP's) successifs indépendamment de leurs étiquettes (Martin, 1975).

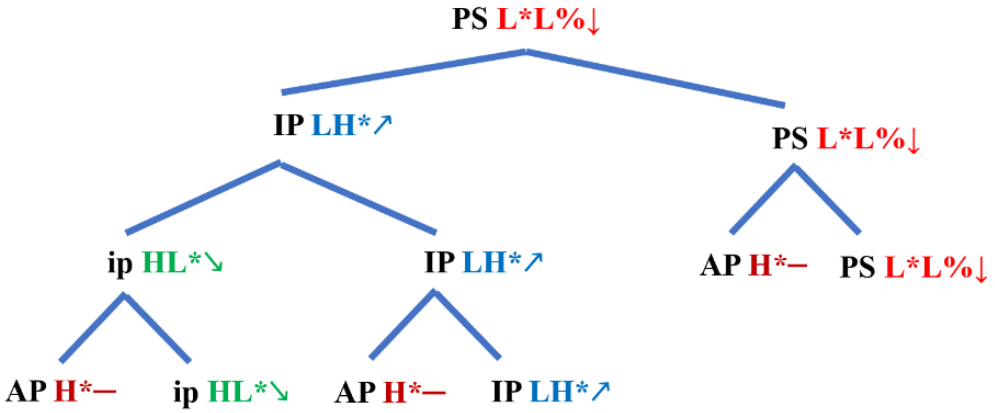


FIGURE 1 Attribution des séquences tonales dans la structure prosodique

Mais peut-on attribuer une séquence tonale unique à une catégorie de syntagme intonatif, ip ou IP ? Les instanciations PS -> **L*L%↓**, IP -> **{LH*↗, H*L#↘}**, ip -> **HL*↘** et AP -> **H*** ne rendent pas compte des variantes de réalisations observées. Par exemple, on peut avoir **H*** **L*L%↓** aussi bien que **LH*↗** **L*L%↓** ou **H*L#↘** **L*L%↓** (voir exemples plus haut). Il faut donc revenir à un principe de base de la phonologie pour trouver une explication, en examinant les conditions nécessaires et suffisantes pour que les séquences tonales indiquent une structure prosodique donnée sans ambiguïté.

4.8 Conditions nécessaires et suffisantes

Pour une structure à deux groupes accentuels déclarative, $[A_{AP}\ B_{AP}]_{SP}$, donc terminée par L*L%↓, la première séquence sur A est celle qui doit se différentier des autres séquence de la phrase, c'est-à-dire B, prises parmi les séquences non encore utilisés dans la liste **H*—** < **HL*↘** < **LH*↗** , **H*L#↘** < **L*L%↓**, soit **H***, **LH*↗** ou **H*L%**. C'est précisément ce qu'on retrouve dans les réalisations attestées.

Pour une structure prosodique déclarative à trois groupes accentuels A, B et C, on peut avoir

$\{[A_{AP}\ B]_{IP}\ C\}_{SP}$, $\{A_{AP}\ B_{AP}\ C\}_{SP}$ ou $\{A_{IP}\ [B_{AP}\ C_{AP}]_{IP}\}_{SP}$ (excluant le cas où B serait une parenthèse prosodique terminée par un contour conclusif $\{A_{AP}\ \{B_{AP}\}_{SP}\ C_{AP}\}_{SP}$).

Ainsi pour [{(*Si ces **ŒU**fs*)$_{AP}$ (*étaient fr**AI**s*)$_{AP}$}$_{IP}$ (*j'en prendr**AI**s*)$_{AP}$]$_{SP}$ en remplaçant les catégories de syntagmes intonatifs par leur instanciations attendues SP -> **L*L%↓** ; IP -> **{LH*—, H*L#↘}** ; ip -> **HL*↘** ; AP -> **H*—**, les variantes possibles sont :

[{(*Si ces **ŒU**fs*) $_{H^{*-}}$ (*étaient fr**AI**s*)}$_{LH^{*}\nearrow}$ (*j'en prendr**AI**s*)] $_{L^{*}L\%\downarrow}$

[{(*Si ces **ŒU**fs*) $_{HL^{*}\searrow}$ (*étaient fr**AI**s*)}$_{LH^{*}\nearrow}$ (*j'en prendr**AI**s*)] $_{L^{*}L\%\downarrow}$

[{(*Si ces **ŒU**fs*) $_{H^{*-}}$ (*étaient fr**AI**s*)}$_{H^{*}L\#}$ (*j'en prendr**AI**s*)] $_{L^{*}L\%\downarrow}$

[{(*Si ces **ŒU**fs*) $_{HL^{*}\searrow}$ (*étaient fr**AI**s*)}$_{H^{*}L\#}$ (*j'en prendr**AI**s*)] $_{L^{*}L\%\downarrow}$

Ces variantes ne sont évidemment pas liées à la structure morphosyntaxique, qui est la même pour les 4 variantes. Le même principe s'applique à n'importe quelle structure prosodique. Il faut et il suffit que les séquences tonales assurent les contrastes entre les catégories de syntagmes intonatifs, c'est-à-dire entre AP, ip et IP, étiquettes dont on pourrait du reste se passer, en ne considérant que la hiérarchie des regroupements des mots prosodiques (AP's) pour une structure donnée (Martin 1975).

4.9 Algorithme d'attribution des séquences tonales

Le mécanisme d'attribution des évènements prosodiques, en partant d'une structure prosodique décrite par ses composantes AP, ip et IP, est donc le suivant :

On commence par la racine SP -> **L*L%↓** (ou **H*H%↑**).

Ensuite, pour les IP, en fait pour les groupes de premier niveau, on sélectionne tous les évènements prosodiques de rang inférieur à **L*L%** dans la liste ordonnée **L*L%↓** > **{H*L#↘, LH*↗}** > **HL*↘** > **H*—**, c'est-à-dire **H*L#**, **LH*** ou **H***, HL* étant exclu car non suivi de LH*.

Pour les ip, groupes de second niveau, même procédure, en retenant **HL*↘** (cette fois suivi de **LH*↗**).

Finalement les AP recevront un **H*—**.

Pour {[A$_{AP}$ B]$_{IP}$ C}$_{SP}$ par exemple on part du contour terminal L*L%↓. Dans l'ordre d'attribution des contours aux frontières des syntagmes intonatifs AP < ip < IP < SP, en fait les groupes selon leurs niveaux, et des évènements prosodiques H* < HL*↘ < {H*L#↘, LH*↗ < L*L%↓, on attribue LH*↗ à IP (Puisque L*L%↓ terminal est déjà pourvu) puis HL*↘ ou plus bas dans la hiérarchie H* à ip : {[A$_{\{H^{*}.\ HL^{*}\searrow\}}$ B$_{LH^{*}\nearrow}$] C}$_{L^{*}L\%\downarrow}$

Même principe pour {A$_{IP}$ [B$_{AP}$ C$_{AP}$] $_{IP}$} $_{SP}$: {A$_{LH^{*}\nearrow}$ [B$_{H^{*}}$ C$_{AP}$] $_{IP}$} $_{L^{*}L\%\downarrow}$. B reçoit un ton H* et non HL*↘ plus haut dans la hiérarchie car HL*↘ ne peut apparaitre sans un LH*↗ qui suit.

Enfin pour {A$_{AP}$ B$_{AP}$ C} $_{SP}$: {A$_{\{H^{*},\ LH^{*}\nearrow\}}$ B$_{\{H^{*},\ LH^{*}\nearrow\}}$ C} $_{L^{*}L\%\downarrow}$ il s'agit d'une énumération, donc A et B doivent être pourvus d'un même contour, soit H* soit LH*↗ soit même H*L#↘ (HL*↘ étant exclu placé immédiatement devant L*L%↓).

5 Exemples de structure prosodique

Les règles rendant compte des séquences bien formées des patrons mélodiques du français permettent de construire sans ambigüité la structure prosodique d'une phrase donnée, et ce à partir des séquences tonales à l'endroit des voyelles des syllabes accentuées. Elles permettent aussi de

rendre compte des variantes possibles. La Figure 2 montre l'analyse de l'exemple {[(*Si ces **ŒU**fs*) $_{AP}$ **HL*↘** (*étaient fr**AI**s*) $_{AP}$] $_{IP}$ **LH*↗** [(*j'en prendr**AI**s*)$_{AP}$]$_{IP}$}$_{PS}$ **L*L%↓,** où (*Si ces **ŒU**fs*) est terminé par **HL***, mais qui pourrait être terminé par le ton neutralisé **H*** en satisfaisant aux conditions nécessaires et suffisantes pour indiquer la même structure prosodique (**H*** pouvant être même réalisé par un F0 montant, mais inférieur au seuil de glissando et donc perçu comme un ton statique).

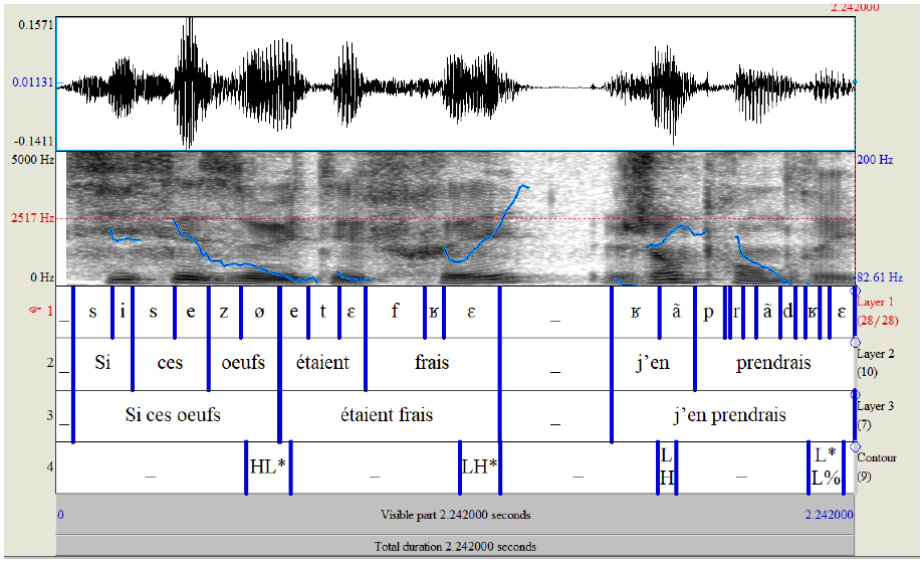


FIGURE 2. Séquence de cibles tonales observées ToBI et structure prosodique de {[(*Si ces **ŒU**fs*) $_{AP}$ **HL*↘** (*étaient fr**AI**s*) $_{AP}$] $_{IP}$ **LH*↗** [(*j'en prendr**AI**s*)$_{AP}$]$_{IP}$}$_{PS}$ **L*L%↓**

Une phrase ne comportant que deux groupes accentuels aura le premier constituant à la fois un AP, un ip et un IP, et pourra porter aussi bien un **H***, un **LH*↗** ou un **H*L#↘**, mais jamais un **HL*↘** (Cette contrainte ne semble jamais avoir été mentionnée dans le cadre autosegmental-métrique). De même, un ip pourra porter un **H*** aussi bien qu'un **LH*↗**

Par contre, dans l'exemple [{[(*Si ces **ŒU**fs*)$_{AP}$ **H*** (*de can**A**rd*)$_{AP}$]$_{ip}$ **HL*↘** (*étaient fr**AI**s*)$_{AP}$]$_{ip}$}$_{IP}$ **LH*↗** [(*j'en prendr**AI**s*)$_{AP}$] $_{ip}$} $_{IP}$]$_{PS}$ **L*L%↓,**

toutes les conditions de contraste entre marqueurs prosodiques sont remplies pour cette structure, et on n'observera donc pas de variantes (Figure 3).

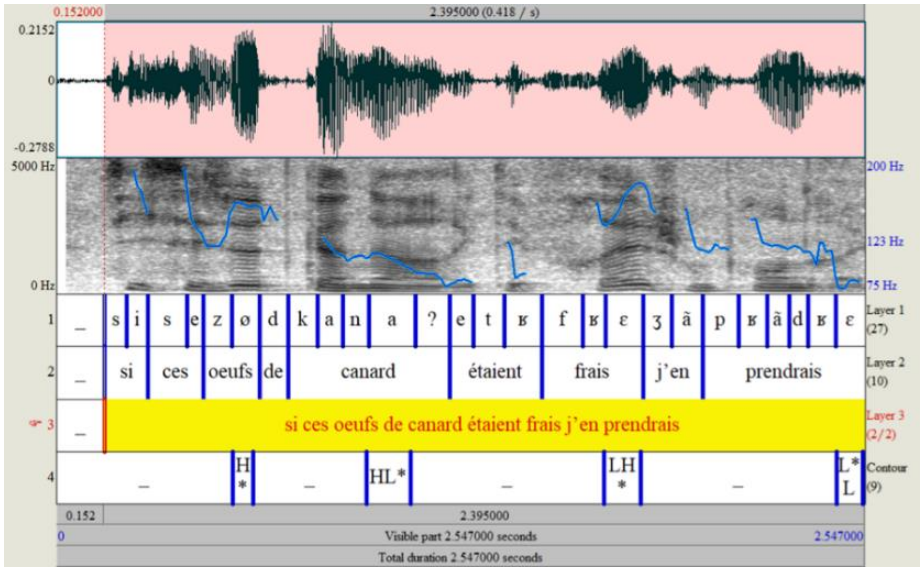


FIGURE 3. Séquence de cibles tonales ToBI et structure prosodique de l'exemple *Si ces œufs de canard étaient frais j'en prendrais*.

6 Discussion

Dans l'exemple de Delais et al. (2020) [{(*les amIs*)$_{AP}$(*du marI*)$_{AP}$ (*de ValérIe*)$_{AP}$↘}ip {.....}ip]$_{IP}$, qu'il y ait un contour descendant sur ip n'a rien à voir avec la syntaxe ou la métrique. Il résulte simplement de l'application des conditions nécessaires et suffisantes pour encoder une structure prosodique donnée, terminée par une séquence montante interrogative située plus loin dans la phrase (principe du contraste de pente en français). Les deux premiers groupes accentuels ne peuvent porter qu'un ton neutralisé H*, la « réserve » de séquences tonales disponibles étant épuisée.

Le même processus s'observe pour l'exemple [{(*les amIs*)$_{H*AP}$(*du marI*)$_{H*AP}$ (*de ValérIe*)$_{HL*↘}$}$_{ip}$ {(*je les ai appelÉs*)AP}$_{ip}$]$_{IP}$ $_{LH*↗}$ [*...et nous nous sommes rencontrÉs..*]$_{IP}$ $_{L*L\%↓}$

Tout dépend de la structure prosodique assignée à la phrase par le locuteur, que ce soit en lecture ou en parole spontanée, donc du mapping fait à partir de la morphosyntaxe. Dans *la mamie des amis de Rémy demandait l'institutrice* (Michelas et D'Imperio 2011, 2015), la structure prosodique peut être :

{[(*la mamIe*)] $_{HL*↘}$ [(*des amIs*) $_{H*}$ (*de RémY*)]}$_{LH*↗}$ {[(*demandAIt*) (*l'institutrIce*)]}$_{L*L\%↓}$
aussi bien que :

{[(*la mamIe*)$_{H*}$ (*des amIs*)]$_{HL*↘}$ [(*de RémY*)]}$_{LH*↗}$ {[(*demandAIt*) (*l'institutrIce*)]}$_{L*L\%↓}$ ou encore

[(*la mamIe*)$_{H*}$ (*des amIs*)$_{H*}$ (*de RémY*)]$_{LH*↗}$ {[(*demandAIt*) (*l'institutrIce*)]}$_{L*L\%↓}$, le locuteur n'étant pas nécessairement syntacticien confirmé pour réaliser ou non la congruence attendue avec la syntaxe.

7 Conclusion

On propose une grammaire prosodique, opérant sur les séquences tonales portées par les voyelles des syllabes accentuées des groupes accentuels, rendant compte de l'indication de la structure prosodique de la phrase. Cette grammaire est totalement indépendante des autres structures morphosyntaxique ou sémantique de la phrase, et ses règles décrivent des séquences tonales attestées dans les données, aussi bien lues que spontanées, ainsi que leurs variantes. Elle met en œuvre les conditions nécessaires et suffisantes pour qu'une structure prosodique donnée puisse être indiquée sans ambiguïté par un ensemble de séquences tonales limité.

La relation avec la syntaxe se trouve ainsi inversée. Au lieu de recenser les configurations de séquences prosodiques à partir de configurations syntaxiques variées, et en particulier des frontières de syntagmes, on procède à l'inverse. À partir d'une structure prosodique donnée, résultant d'une séquence de cibles mélodiques données, on peut ensuite analyser les propriétés syntaxiques des phrases qui peuvent y être associées (ce qui n'est pas abordé ici), et en particulier les propriétés syntaxiques des groupes accentuels, qui apparaissent comme les unités lexicales effectivement utilisées pas les locuteurs et les auditeurs, à la place de mots (Martin, 2018). La fonction des évènements prosodiques apparait alors plus clairement : donner à l'auditeur des balises permettant un pré assemblage rapide des groupes accentuels, indispensable étant donné la mémoire temporelle de la parole continue limitée à 2 à 3 secondes (R. Martin et al., 2014 ; Ph. Martin, 2014).

Cette inversion du processus d'analyse syntaxe -> intonation en intonation -> syntaxe apparait beaucoup plus simple à mettre en œuvre, et rend mieux compte des variantes observées en parole lue et spontanée ainsi que du processus de décodage de la parole par les auditeurs.

Références

BALLY Ch. (1944) *Linguistique générale et linguistique française*, Berne : Francke.
BILLIERES M. (2021) Intonation, prosodie, accentuations, rythme, https://www.verbotonale phonetique.com/accentuations-rythme-intonation-et-tutti-quanti/
COUSTENOBLE H. & ARMSTRONG L. (1934) *Studies in French Intonation*, Cambridge: W. Heffer & Sons.
DELATTRE P. (1966) Les dix intonations de base du français, *French Review* (40) 1-14.
DELAIS-ROUSSARIE E, POST B., YOU H-Y. (2020) Unités prosodiques et grammaire intonative du français : vers une nouvelle approche, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole*, Nancy, France, 08-19 juin 2020.
D'IMPERIO M., MICHELAS A & PORTES C. (2016) École d'Aix et approches tonales de l'intonation du français : Un aperçu des recherches depuis les années 1970, *Langue française* 191 (3/2016), 31-46.
HIRST, D. & DI CRISTO A. (1984) French intonation: A parametric approach, *Die Neueren Sprachen* 83. 554-569.
JUN S-A. & FOUGERON C. (2002) The Realizations of the Accentual Phrase in French Intonation, *Probus* 14, 147-172.
LLANOS F., SNEED GERMAN J., NIKE GNANATEJA G., Chandrasekaran B. (2021) The neural processing of pitch accents in continuous speech, *Neuropsychologia*, 2021, 158.
LÉON P. (1993) *Précis de phonostylistique : Parole et expressivité*, coll. Fac Linguistique, Nathan, Paris, 335 p.
MARTIN Ph. (1975) Analyse phonologique de la phrase française, *Linguistics*, 146 (Fév. 1975), 35-68.
MARTIN Ph. (2008) Postfixes et suffixes interrogatifs : un cas d'ambiguïté prosodique ? Actes de la conférence de la section tchéco-slovaque de l'ISPhS 2008, 111-119.
MARTIN Ph. (2018) Intonation, structure prosodique et ondes cérébrales, London : ISTE, 332 p.
MARTIN, R., YAN H. & SCHNUR T. (2014), Working memory and planning during sentence production, *Acta Psychologica* 152C,120-132.
MERTENS P. (1987) *L'intonation du français : De la description linguistique à la reconnaissance automatique*, Ph. D., U. C. Leuven
MICHELAS A. (2011) *Caractérisation phonétique et phonologique du syntagme intermédiaire en français de la production à la perception*, Thèse de doctorat, Université d'Aix-en-Provence.
MICHELAS A. and D'IMPERIO M-P. (2015) Prosodic boundary strength guides syntactic parsing of French utterances, *Laboratory Phonology* 2015, 6(1), 119-146.
ORFEO (2017) Outils et Recherches sur le Français Écrit et Oral. http://www.projet-orfeo.fr/
POST B. (1999) Restructured Phonologic Phrases in French, evidence from clash resolution, *Linguistics*, 37/1, 1999, 41-63.
ROSSI M. (1971) Le seuil de glissando ou seuil de perception des variations tonales pour la parole, *Phonetica* (23) 1-33.
SIWIS Corpus (2016) Yamagishi, J. et al. The SIWIS French Speech Synthesis Database, https://doi.org/10.7488/ds/1705.
VAISSIÈRE J. (1974) On French Prosody, *Quarterly Progress Report*, M.I.T., Res. Lab. of Electr., (114), 212-223.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Vérification automatique de la voix de locuteurs après conversion à l'aide de PPGs

Thibault Gaudier[1, 2] Marie Tahon[1] Anthony Larcher[1] Yannick Estève[2]
(1) Laboratoire d'Informatique de l'Université du Mans (LIUM), 72100 Le Mans, France
(2) Laboratoire d'Informatique d'Avignon (LIA), 83000 Avignon, France
{prenom}.{nom}@univ-lemans.fr

RÉSUMÉ

La création de contenu journalistique peut être assistée par des outils technologiques comme la synthèse de parole. Cependant l'éditeur doit avoir la possibilité de contrôler la génération du contenu audio comme la prosodie, la prononciation ou le contenu linguistique. Dans ces travaux, un système de conversion de voix génère un signal de locuteur cible à partir d'une représentation temporelle de type Phonetic PosteriorGrams (PPGs) extraite d'un audio source. Les PPGs dêmelent le contenu phonétique du contenu rythmique, et sont généralement considérés indépendants du locuteur. Cet article présente un système de conversion utilisant les PPGs, et son évaluation en qualité audio avec un test perceptif. Nous montrons également qu'un système de vérification du locuteur ne parvient pas à identifier le locuteur source après la conversion, même si le modèle a été entraîné sur des données synthétiques.

ABSTRACT

Automatic Speaker's Voice Verification after Speech Conversion using PPGs

The creation of journalistic content can be assisted by technologies such as speech synthesis. In any cas, the editor needs the possibility to control the audio content generation such as prosody, pronunciation or linguistics. In the present work, a voice conversion system generates a target speaker signal from a temporal representation, using Phonetic PosteriorGrams (PPGs) extracted in the source audio. This representation disentangles rhythmic and phonetic information, and is usually considered speaker-independant. This paper presents a PPGs-based speech conversion system, and its evaluation in terms of general quality. We also demonstrate that a speaker verification model is not able to recover the source speaker after conversion with PPGs, even when the model is trained on synthetic data.

MOTS-CLÉS : synthèse de parole, représentation interprétable de la parole, reconnaissance du locuteur.

KEYWORDS: speech synthesis, interpretable speech representation, speaker recognition.

1 Introduction

Les journalistes et médias ont désormais accès à des flux importants de contenu, venant de différents endroits du monde et dans différentes langues. Dans ce contexte, le projet européen SELMA [1] vise à développer des outils permettant de réaliser du doublage automatique, en générant un signal de parole d'une voix cible à partir d'un texte éventuellement traduit. Une manière de faire serait d'utiliser un

1. https://selma-project.eu/

système de synthèse à partir de texte (TTS) afin de générer le signal correspondant au texte traduit. Cependant, malgré les avancées récentes du domaine, le signal ainsi synthétisé ne correspond pas nécessairement aux besoins des utilisateurs. Il y a donc la nécessité d'avoir des systèmes de conversion de parole permettant un contrôle plus fin de la parole générée selon différents aspects, en utilisant des représentations interprétables. Par exemple, EdiTTS (Tae *et al.*, 2022) utilise le texte comme représentation afin de modifier le contenu linguistique, et (Zhao *et al.*, 2019) utilise les Phonetic PosteriorGrams (PPG) comme représentation permettant de contrôler les contenus rythmique et phonétique.

Les PPGs sont une représentation temporelle des probabilités de présences de différentes unités phonétiques. Ainsi, il est possible, techniquement, de modifier les durées, sans changer les phonèmes (et réciproquement) donnant ainsi du contrôle aux utilisateurs (Zhao *et al.*, 2019) and (Yeh *et al.*, 2018). Evidemment, cette modification est artificielle et ne pourra pas fournir de parole audible si le modèle utilisé pour la conversion ne compense pas certaines modifications (par exemple allonger la durée d'une plosive). Cet aspect sera étudié dans un futur proche.

Les PPGs ont déjà été utilisés pour la tâche de conversion de voix (Levy-Leshem & Giryes, 2021). Cette représentation contient également des informations relatives à l'accent du locuteur, permettant de réaliser de la conversion d'accent (Zhao *et al.*, 2019) (modifier l'accent sans changer la voix). Cependant, cela signifie que certaines informations relatives au locuteur, présentes dans les PPGs, pourraient passer dans le signal audio généré à partir de cette représentation. À long terme, notre objectif est de synthétiser un signal de parole (source) à partir de texte, puis de convertir ce signal vers une voix cible à partir de PPGs corrigés manuellement. Nous cherchons donc à vérifier que le locuteur de l'audio cible ne peut pas être identifié à partir de l'audio source, que cette source soit naturelle (comme présenté ici) ou bien synthétique (comme dans notre objectif à long terme).

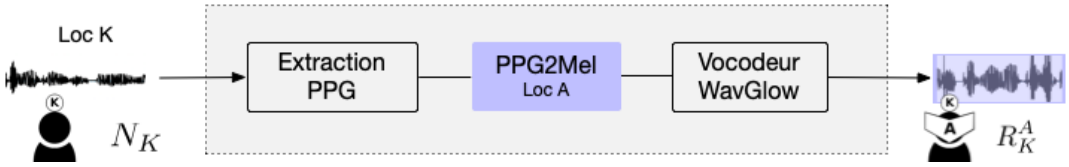


FIGURE 1 – Approche de synthèse utilisant les PPG (PPG2Mel). Les éléments en bleu sont spécifiques à un locuteur. N est un échantillon naturel, R un échantillon resynthétisé (voir Section 4.1)

Pour cela, nous entraînons différents modèles de conversion à partir de PPGs, de manière similaire à (Zhao *et al.*, 2019) and (Levy-Leshem & Giryes, 2021). Plus précisément, comme illustré Figure 1, un système PPG2Mel est entraîné à générer une voix A à partir de PPGs. Ainsi, à l'inférence, lorsqu'un signal de parole N_K provenant du locuteur source K passe dans le système, le signal synthétisé R_K^A est converti vers la voix cible A. Nous réalisons une évaluation subjective de la qualité de la parole générée, comparant notre signal généré à celui obtenu par un système TTS, un signal obtenu par un vocodeur et le signal naturel. L'objectif de cette évaluation, décrite dans la Section 3, est de s'assurer de générer de la parole de qualité convenable à partir de PPGs. Comme notre but n'est pas de proposer un nouveau système de synthèse mais d'en utiliser un existant pour une autre tâche, nous ne cherchons pas à comparer ce système à d'autres à partir d'un score de qualité.

À notre connaissance, aucune étude n'a tenté d'identifier une éventuelle information relative au locuteur original après conversion à partir de PPGs. Notre contribution principale est l'étude de la capacité d'un système de vérification du locuteur (SV) à identifier le locuteur original d'un audio synthétisé. Ceci est réalisé en synthétisant des audios de différents locuteurs source (base de données

VoxCeleb (Nagrani *et al.*, 2017; Chung *et al.*, 2018)) en utilisant deux modèles PPG2Mel entraînés sur deux voix cibles différentes. La Section 4.1 détaille le contenu des bases de données utilisées, leur usage et les notations utilisées. À partir de cette base de données synthétique annotée avec les locuteurs sources, nous pouvons entraîner des systèmes de SV de plusieurs manières. La Section 4.2 détaille notre protocole et la Section 4.3 montre les résultats obtenus. Contrairement au protocole d'évaluation de conversion de voix, nous ne cherchons pas à identifier la similarité entre les audios synthétiques et les locuteurs cibles, mais à vérifier si un système de vérification du locuteur peut identifier les locuteurs sources à partir d'audio synthétique, malgré la présence de la voix cible.

2 Synthèse de parole

Les tâches de synthèse de parole sont souvent divisées en deux étapes : la prédiction d'une représentation fréquentielle (mel-spectrogramme par exemple) à partir de l'entrée, puis l'utilisation d'un vocodeur pour obtenir le signal audio correspondant.

De nos jours, les systèmes de synthèse à partir de texte sont principalement des systèmes neuronaux autorégressifs comme Tacotron2 (Shen *et al.*, 2018), ou des systèmes séquence-vers-séquence, par exemple utilisant des Transformers comme FastSpeech (Ren *et al.*, 2019). L'introduction de contrôle dans ces systèmes est généralement réalisée en conditionnant la génération de parole à un locuteur spécifique (Cooper *et al.*, 2020) and (Valle *et al.*, 2020). Cependant, le contrôle pour d'autres aspects comme la prosodie (Sini *et al.*, 2020), l'intonation (Łańcucki, 2021), le style (Wang *et al.*, 2018) ou l'émotion (Diatlova & Shutov, 2023) ont été introduits dans les systèmes de synthèse.

Le but des système de conversion de voix est généralement de préserver certains aspects provenant d'un audio source (par exemple le contenu linguistique) et d'autres provenant d'un audio cible (les indices acoustiques relatifs à un locuteur). L'utilisation de représentation démêlées (principalement entre locuteur, contenu linguistique et/ou prosodie) extraites du signal cible ont été utilisées par (Qian *et al.*, 2019, 2020), (Polyak *et al.*, 2021). Les PPGs ont également été utilisés pour la conversion de voix (Levy-Leshem & Giryes, 2021), d'accent (Zhao *et al.*, 2019) ou de rythme (Yeh *et al.*, 2018).

Comme beaucoup de systèmes de synthèse génèrent des mel-spectrogrammes à partir de la consigne, le rôle du vocodeur est alors de produire le signal audio correspondant dans le domaine temporel. Aujourd'hui les vocodeurs mainstream sont basés sur des réseaux génératifs comme HifiGan (Kong *et al.*, 2020), ou WaveGlow (Prenger *et al.*, 2019).

3 Synthèse à partir de Phonetic PosteriorGrams (PPG)

3.1 Les Phonetic PosteriorGrams (PPG)

Les Phonetic PosteriorGrams (PPG) (exemple Figure 2) sont une représentation temporelle et probabiliste des phonèmes prononcés dans un audio. Ainsi, pour chaque trame de 30 ms, on obtient la probabilité de présence de chacun des phonèmes. Cette représentation présente certains avantages : elle permet à un utilisateur de contrôler finement l'audio représenté selon certaines caractéristiques, comme le contenu phonétique ou le rythme de parole. Cette représentation contient également une information plus riche qu'une séquence de phonème, car la confusion entre plusieurs classes de phonèmes, qui s'observe par une probabilité non nulle pour différentes classes dans une même trame, peut être interprétée comme une différence de réalisation de ces phonèmes. Ainsi, les PPG contiennent des informations relatives à la phonétique de la phrase et au rythme de celle-ci, mais d'autres informations pourraient être également présentes de manière cachée. L'expérience présentée dans cet article cherche à identifier une éventuelle présence d'information relative au locuteur dans un PPG.

Les PPG sont extraits à partir du modèle présenté dans (Zhao *et al.*, 2019), qui est un Generalized Maxout Network fourni par Kaldi (Zhang *et al.*, 2014), entraîné à imiter un GMM-HMM représentant 5816 unités acoustiques, qui sont ensuite regroupées en 40 phonèmes pour l'anglais. 100 trames de PPG sont extraites chaque seconde.

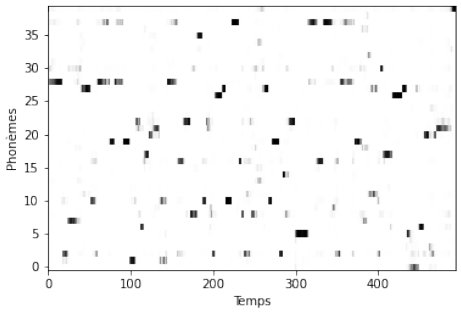


FIGURE 2 – Exemple de PPG pour la phrase : "Such risks can be lessened when the President recognizes the security problem"

3.2 Synthèse de parole à partir de PPG

Les PPG sont une représentation présentant des similitudes avec un encodage one-hot des phonèmes. Ainsi, il est possible de générer un audio correspondant à un PPG de la même manière que pour d'autres représentations de la parole telle une séquence de phonème. Nous avons utilisé une approche similaire à celle présentée dans (Zhao *et al.*, 2019) et (Levy-Leshem & Giryes, 2021). Nous avons donc entraîné Tacotron2 (Shen *et al.*, 2018) en utilisant les PPG extraits des audios en entrée pour prédire les mel-spectrogrammes extraits de ces mêmes audios. Ce système sera désigné par PPG2Mel. Le modèle converge plus rapidement en utilisant des PPG qu'en utilisant du texte, car les PPG sont déjà alignés temporellement avec l'audio. La seule modification que nous avons faite à l'architecture de Tacotron2 est le remplacement de la couche d'embedding de caractères par une couche linéaire transformant les 40 probabilités de présence des phonèmes en une représentation interne de dimension 512.

Comme le système PPG2Mel produit des mel-spectrogrammes, nous devons ensuite les convertir dans le domaine audio. Nous avons utilisé WaveGlow, un vocodeur neuronal décrit dans (Prenger *et al.*, 2019). Notre vocodeur est entraîné sur le dataset LJSpeech (Ito & Johnson, 2017) en utilisant la configuration par défaut à l'exception de la fréquence d'échantillonage, que nous avons passée de 22,05kHz à 16kHz. Nous avons utilisé l'implémentation proposée par Nvidia, disponible sur GitHub [2]

3.3 Évaluation perceptive de la parole

Notre objectif avec cette évaluation est de s'assurer que le système présenté dans la Section 3.2 génère des échantillons audios de qualité suffisante pour étudier la présence d'information relative au locuteur dans la synthèse. Nous avons donc comparé la qualité du système PPG2mel avec une version obtenue par TTS, une version obtenue en utilisant uniquement le vocodeur (analyse-synthèse), ainsi que l'échantillon naturel original. Les systèmes TTS et PPG2Mel ont été entraînés sur la base de données LJSpeech (Ito & Johnson, 2017). Notre baseline TTS utilise l'implémentation de Tacotron2 par Nvidia [3], en changeant uniquement la fréquence d'échantillonage à 16kHz afin de rester consistant avec les autres échantillons. Nous avons utilisé les ensembles d'entraînement, validation et test provenant du même dépôt GitHub. Nous avons ensuite divisé l'ensemble de test en

2. https://github.com/nvidia/waveglow
3. https://github.com/nvidia/tacotron2

trois sous-ensembles en fonction de la durée des audios. 20 segments sont sélectionné dans chaque sous-ensemble afin d'avoir une représentation des audios courts, moyens et longs. Un test Mean Opinion Score (MOS) (1 : mauvais, 5 excellent) a été mis en place avec la plateforme FlexEval (Fayet *et al.*, 2020) à partir de la question suivante : "jugez la qualité de cet échantillon audio". 36 participants sur 44, majoritairement non natifs, ont évalué l'ensemble des 20 segments audios qui leur était présentés. Les 4 versions des 60 segments ont été évalués en moyenne 12 fois.

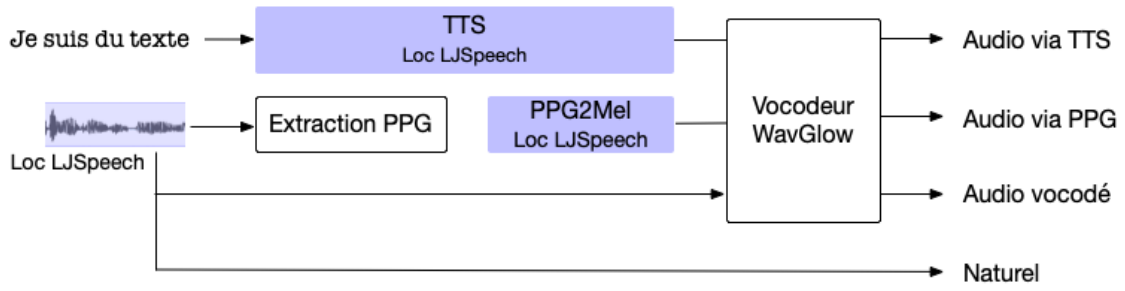


FIGURE 3 – Représentation des 4 versions de l'audio présentées lors du test perceptif. Les blocs bleus sont spécifiques à un locuteur

Les résultats sont décrits dans la Table 1. Nous avons retiré les étapes d'introduction et utilisé toutes les autres réponses, dont celles venant de participants n'ayant pas complété toutes les étapes. À partir de ces résultats, nous pouvons conclure que l'utilisation de PPG pour la synthèse de parole ne dégrade pas la qualité par rapport à la synthèse à partir de texte. Nous observons également qu'une partie importante de la dégradation en qualité provient du vocodeur. Cela peut s'expliquer par le fait que le vocodeur est biaisé par le locuteur de LJspeech. Pour améliorer ce point, il pourrait être intéressant de fine-tuné le vocodeur sur nos données. Nous sommes conscients que nos résultats de MOS sont inférieurs à ceux de la littérature, peut-être parce qu'ils sont non-natifs. Cependant, nous observons également que l'audio naturel n'est pas non plus évalué avec d'aussi bons scores.

TABLE 1 – MOS obtenus lors de l'évaluation. Intervalles de confiance à 95%

Système	Audio naturel	Audio vocodeur	Audio TTS	Audio PPG
MOS	4.35 ± 0.07	3.47 ± 0.07	3.11 ± 0.07	3.24 ± 0.07

4 Identification du locuteur source

Dans cette section, notre objectif est de déterminer si un système de vérification naïf, entraîné sur de la parole naturelle, peut identifier le locuteur source après conversion (Q1). Ensuite, nous utilisons des données synthétiques pour entraîner un modèle informé à identifier ce locuteur source. Nous souhaitons savoir à quel point ce modèle informé parvient à identifier le **locuteur source** dans des échantillons convertis vers la voix cible, mais aussi à partir d'échantillons naturels afin d'identifier le décalage entre parole naturelle et synthétique (Q2). Le modèle informé est supposé apprendre à différencier les locuteurs dans un espace adapté à la voix cible. Si la conversion cache complètement le locuteur source, on s'attend à une forte dégradation des résultats avec les deux modèles (naïf et informé). En revanche, si elle ne cache que partiellement le locuteur source, le modèle naïf devrait obtenir de mauvais résultats, mais le modèle informé devrait parvenir à identifier le locuteur source, et donc obtenir de meilleurs scores. Enfin, nous étudions à quel point les modèles parviennent à lier l'identité du **locuteur cible** provenant des échantillons naturels avec les échantillons convertis vers cette même voix cible (Q3).

4.1 Données et notations

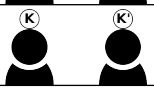
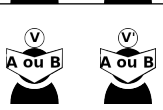
Cette expérience utilise 3 bases de données. La première est la section anglaise de M-AILABS(Solak, 2019), un corpus basé sur LibriVox. Nous avons utilisé 2 locuteurs, E. Klett, notée A, et E. Miller, noté B. Pour chaque locuteur A et B, nous avons 30 à 45 heures de parole, que nous avons divisé en entraînement, validation et test. Au cours de cette expérience, ces deux voix ont servi comme **locuteurs cibles**. Ceci signifie que tous les échantillons synthétiques ont été générés avec l'une de ces voix. Nous avons entraîné deux modèles mono-locuteurs notés PPG2Mel$_A$ et PPG2Mel$_B$, pour les locuteurs A et B (voir Figure 1).

La base de données LibriSpeech-test-clean (Panayotov *et al.*, 2015) est notre base d'enrôlement et de test pour l'expérience de vérification du locuteur. Elle contient 40 locuteurs, équilibrés en terme de genre ($\simeq$ 8 min. de parole par locuteur), notés 1 à 40. Ces locuteurs sont les **locuteurs source** que nous voulons identifier avant et après synthèse. Nous créons deux versions synthétiques de cette base en utilisant les modèles PPG2Mel$_A$ et PPG2Mel$_B$.

Enfin, les bases de données VoxCeleb1&2 (Nagrani *et al.*, 2017; Chung *et al.*, 2018) sont utilisées pour entraîner les systèmes de vérification du locuteur. Les modèles PPG2Mel$_A$ et PPG2Mel$_B$ sont utilisés pour synthétiser tous les échantillons de VoxCeleb vers les **locuteurs cibles**, choisis aléatoirement entre A et B pour chaque échantillon. Les labels de locuteur pour l'apprentissage des modèles sont conservés à l'identique, même si la voix perçue est maintenant différente.

Les échantillons naturels sont notés N_{source}, où $source$ est dans $\{A, B, 1-40\}$. Les échantillons synthétiques sont notés R^{cible}_{source}, où $source$ est identique à précédemment et $cible$ est A ou B selon le modèle PPG2Mel utilisé. Nous ne mentionnerons pas les locuteurs de VoxCeleb. Une illustration des différents cas est présente à la Table 2.

TABLE 2 – Description des notations des différents locuteurs et échantillons synthétiques. Cercle : locuteurs sources ; masques : signal synthétisé en utilisant le modèle lié à la voix indiquée sur le masque. K et K' sont considérés différents.

Données	Locuteurs	Notation	Détails
M-AILABS	4		N_A et N_B. A est E. Klett, B est E. Miller
LibriSpeech-test naturel	40		N_K, $N_{K'}$. $K, K' \in [\![1, 40]\!], K \neq K'$
LibriSpeech-test synthétique	40		$R^A_{K'}$, R^B_K A, B, K, K' décrits précédemment
VoxCeleb1&2	7363		Utilisé pour entraîner le modèle naïf
VoxCeleb1&2 synthétique	7363		Utilisé pour entraîner le modèle informé Synthétisé en utilisant les locuteurs A et B de M-AILABS.

4.2 Modèles de vérification du locuteur

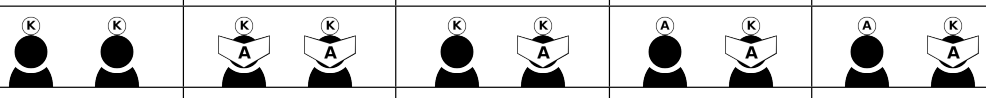
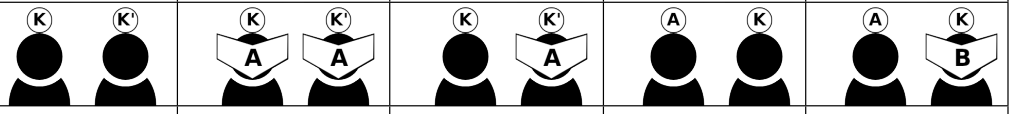
Les modèles de vérification naïf et informé utilisent l'architecture ECAPA-TDNN (Desplanques *et al.*, 2020), avec comme entrée des représentations de l'audio extraites par WavLM-Large [4] et en utilisant l'Additive Angular Margin comme fonction de coût. Nous utilisons des x-vecteurs de dimension 256. Le modèle naïf est entraîné sur les données naturelles de VoxCeleb1&2. Ce modèle obtient un

4. https://github.com/microsoft/unilm/tree/master/wavlm

EER de 1.57% sur VoxCeleb-o après 4 jours d'entraînement sur une carte GPU RTX8000, ce qui est légèrement inférieur à l'état de l'art actuel sur ces données. Le modèle informé est entraîné sur la version synthétique (locuteurs cibles A et B) de VoxCeleb1&2 décrite plus haut. La meilleure version de ce modèle est obtenue après un jour d'entraînement sur la même carte et obtient un EER de seulement 20% sur la version synthétique de VoxCeleb-o.

4.3 Expériences et résultats

TABLE 3 – Définition des expériences par leurs cibles et imposteurs, et taux d'égale erreur (EER) des modèles naïf et informé. Pour chaque expérience, les tests sont définis comme *enrôlement/test*.

Expérience	(1)	(2)	(3)	(4)	(5)
Définition de la cible	N_{1-40}/N_{1-40}	R^A_{1-40}/R^A_{1-40}	N_{1-40}/R^A_{1-40}	N_A/R^A_{1-40}	N_A/R^A_{1-40}
Définition de la cible					
Définition des imposteurs	$N_{1-40}/N_{\overline{1-40}}$	$R^A_{1-40}/R^A_{\overline{1-40}}$	$N_{1-40}/R^A_{\overline{1-40}}$	N_A/N_{1-40}	N_A/R^B_{1-40}
Définition des imposteurs					
Modèle naïf	1.98 %	49.46 %	48.02 %	45.13 %	49.81%
Modèle informé	29.44 %	49.80 %	49.00 %	33.58 %	45.52%

La Table 3 résume les différentes cibles, imposteurs et les résultats obtenus pour 5 expériences de vérification. Les expériences sont décrites par leurs paires enrôlement/test. Un test N_{1-40}/N_{1-40} compare les couples d'échantillons naturels d'un même locuteur parmi 1 à 40, par exemple N_1 et N_1. Un test $N_{1-40}/R^A_{\overline{1-40}}$ compare les échantillons naturels de chaque locuteur 1 à 40 avec les échantillons synthétisés des autres locuteurs, par exemple N_1 avec R^A_2. Chaque expérience donne un EER pour les modèles naïf et informé. Pour les expériences (1), (2) et (3), les labels relatifs au locuteur correspondent au locuteur source parmi $1-40$, tandis que pour les expériences (4) et (5) les labels sont ceux des locuteurs cibles A et B.

La première expérience (1) permet de s'assurer que notre modèle naïf obtient des résultats corrects. Pour cela, nous voulons identifier le locuteur à partir d'audio naturel. Comme attendu, le modèle naïf obtient un bon résultat (EER=1.98%) puisque il s'agit de la tâche d'entraînement de ce modèle. Le modèle informé induit une forte dégradation (EER=29.44%), ce qui indique une différence de domaine entre les données d'entraînement de ce modèle et ce test.

Dans l'expérience (2), on compare les versions converties des audios venant des locuteurs sources $1-40$ avec le modèle PPG2Mel$_A$ (R^A_{1-40}) entre eux, afin de voir si les modèles parviennent à lier les échantillons provenant d'un même locuteur source. Les résultats montrent qu'aucun des modèles ne parvient à réaliser cette tâche (EER $> 49\%$). Cela permet de répondre à la question Q1 : le modèle naïf ne parvient pas à reconnaître des échantillons provenant d'un même locuteur source après synthèse. Une hypothèse est que l'identité du locuteur source a été cachée après conversion. On observe que le modèle informé reconnaît mieux les locuteurs dans l'espace naturel (EER$= 29.44\%$, exp (1)) que dans l'espace synthétique (EER$= 49.80\%$, exp (2)). Durant son entraînement, le modèle informé a

peu convergé, mais il semble que le peu d'éléments discriminants appris permettent uniquement de distinguer les locuteurs dans l'espace naturel, cette tâche étant plus facile. Ceci permet de répondre à la question Q2 : même un modèle informé ne parvient pas à reconnaître le locuteur source après conversion.

L'expérience (3) évalue la capacité des deux modèles à lier un même locuteur dans l'espace naturel et dans l'espace synthétique. Pour cela, nous utilisons les données de LibriSpeech décrites précédemment comme enrôlement, et les versions converties de cette même base utilisant PPG2Mel$_A$ comme données de test. On observe qu'aucun de nos modèles ne parvient à identifier les 40 locuteurs entre ces deux espaces. On peut en conclure que l'approche utilisant les PPGs pour la conversion permet bien de cacher le locuteur source à des modèles de vérification du locuteur, même appris sur des données synthétiques. Les éventuels indices acoustiques permettant d'identifier le locuteur source ne sont pas détectés après la resynthèse.

Les expériences (4), respectivement (5), mesurent la proximité entre les locuteurs source $1-40$ convertis vers la voix A et leur version naturelle (resp. et leur version convertie vers la voix B) par rapport à la proximité avec les échantillons naturels du locuteur A. On conclut de l'expérience (4) que l'identité des échantillons des locuteurs source $1-40$ covnertis avec le modèle PPG2Mel$_A$ ne correspondent pas à l'identité de A, ce qui confirme le fait que le système de conversion ne parvient par à rapprocher le locuteur source du locuteur A. Cependant, les résultats montrent que les échantillons convertis sont plus proches des échantillons naturels de A selon le modèle informé que selon le modèle naïf. On peut donc confirmer que pour le modèle informé, la conversion rapproche les identités naturelle et synthétique. L'expérience (5) montre que les échantillons synthétiques générés avec les modèles PPG2Mel$_A$ et PPG2Mel$_B$ ne sont pas distinguables par le modèle naïf, et sont tous deux éloignés des échantillons naturels de A. Le modèle informé fait une légère distinction entre les échantillons synthétiques générés par PPG2Mel$_A$ et PPG2Mel$_B$. Le modèle de conversion ne permet pas d'atteindre le locuteur cible (ici A ou B) d'un point de vue de ces modèles de vérification. On peut donc répondre à la question Q3 : le modèle informé est légèrement meilleur pour identifier le lien entre la voix cible synthétisée et la voix cible naturelle. Cependant, ce résultat doit être manié avec précaution car réalisé avec deux voix cibles uniquement, et un unique système de vérification.

5 Conclusion

La première expérience présentée vise à s'assurer que notre système de conversion à base de PPGs produit de l'audio de qualité correcte. Le test perceptif réalisé montre que nous obtenons une qualité similaire à celle d'un système TTS habituel, et que le vocodeur utilisé est la source d'une grande partie de la dégradation. L'utilisation d'un meilleur vocodeur ainsi que la réalisation d'un test perceptif de similarité locuteur pourraient être des pistes de recherche pour continuer ces travaux.

Nous avons ensuite entraîné deux systèmes de vérification du locuteur sur de l'audio naturel et sur des données synthétiques afin d'identifier l'information relative au locuteur source qui aurait été cachée par la conversion. Nos expériences montrent que même si le système naïf obtient des résultats comparables avec l'état de l'art sur de la parole naturelle, ni ce système ni le système informé ne parviennent à identifier les locuteurs originaux une fois l'étape de conversion passée. De même, aucun de nos systèmes n'a été capable de lier les versions naturelles et synthétiques d'un même locuteur cible. Nous concluons donc que l'ensemble de la chaine de conversion depuis l'extraction des PPGs à la génération de l'audio permet de cacher les indices acoustiques à un système de vérification du locuteur. Ainsi notre modèle de conversion à partir de PPGs semble pertinent pour controller la génération de parole, tout en cachant les locuteurs sources.

Références

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. In *INTERSPEECH*.

COOPER E., LAI C.-I., YASUDA Y., FANG F., WANG X., CHEN N. & YAMAGISHI J. (2020). Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6184–6188 : IEEE.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). ECAPA-TDNN : Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification. In *INTERSPEECH 2020*, p. 3830–3834 : International Speech Communication Association (ISCA).

DIATLOVA D. & SHUTOV V. (2023). EmoSpeech : guiding FastSpeech2 towards Emotional Text to Speech. In *Proc. 12th ISCA Speech Synthesis Workshop*, p. 106–112. DOI : 10.21437/SSW.2023-17.

FAYET C., BLOND A., COULOMBEL G., SIMON C., LOLIVE D., LECORVÉ G., CHEVELU J. & LE MAGUER S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *Journées d'Études sur la Parole*, p. 22–25, Nancy, France.

ITO K. & JOHNSON L. (2017). The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

KONG J., KIM J. & BAE J. (2020). Hifi-gan : Generative adversarial networks for efficient and high fidelity speech synthesis. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 17022–17033 : Curran Associates, Inc.

ŁAŃCUCKI A. (2021). Fastpitch : Parallel text-to-speech with pitch prediction. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6588–6592 : IEEE.

LEVY-LESHEM R. & GIRYES R. (2021). Taco-vc : A single speaker tacotron based voice conversion with limited data. In *2020 28th European Signal Processing Conference (EUSIPCO)*, p. 391–395. DOI : 10.23919/Eusipco47968.2020.9287448.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : a large-scale speaker identification dataset. *Telephony*, **3**, 33–039.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

POLYAK A., ADI Y., COPET J., KHARITONOV E., LAKHOTIA K., HSU W.-N., MOHAMED A. & DUPOUX E. (2021). Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*, p. 3615–3619. DOI : 10.21437/Interspeech.2021-475.

PRENGER R., VALLE R. & CATANZARO B. (2019). Waveglow : A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 3617–3621 : IEEE.

QIAN K., ZHANG Y., CHANG S., HASEGAWA-JOHNSON M. & COX D. (2020). Unsupervised speech decomposition via triple information bottleneck. In H. D. III & A. SINGH, Éds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, p. 7836–7846 : PMLR.

QIAN K., ZHANG Y., CHANG S., YANG X. & HASEGAWA-JOHNSON M. (2019). AutoVC : Zero-shot voice style transfer with only autoencoder loss. In K. CHAUDHURI & R. SALAKHUTDINOV,

Éds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, p. 5210–5219 : PMLR.

REN Y., RUAN Y., TAN X., QIN T., ZHAO S., ZHAO Z. & LIU T.-Y. (2019). Fastspeech : Fast, robust and controllable text to speech. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 32 : Curran Associates, Inc.

SHEN J., PANG R., WEISS R. J., SCHUSTER M., JAITLY N., YANG Z., CHEN Z., ZHANG Y., WANG Y., SKERRV-RYAN R., SAUROUS R. A., AGIOMVRGIANNAKIS Y. & WU Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4779–4783. DOI : 10.1109/ICASSP.2018.8461368.

SINI A., MAGUER S. L., LOLIVE D. & DELAIS-ROUSSARIE E. (2020). Introducing Prosodic Speaker Identity for a Better Expressive Speech Synthesis Control. In *Speech Prosody 2020*, p. 935–939 : ISCA.

SOLAK I. (2019). The m-ailabs speech dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.

TAE J., KIM H. & KIM T. (2022). EdiTTS : Score-based Editing for Controllable Text-to-Speech. In *Proc. Interspeech 2022*, p. 421–425. DOI : 10.21437/Interspeech.2022-6.

VALLE R., LI J., PRENGER R. & CATANZARO B. (2020). Mellotron : Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6189–6193.

WANG Y., STANTON D., ZHANG Y., RYAN R.-S., BATTENBERG E., SHOR J., XIAO Y., JIA Y., REN F. & SAUROUS R. A. (2018). Style tokens : Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, p. 5180–5189 : PMLR.

YEH C.-C., HSU P.-C., CHOU J.-C., LEE H.-Y. & LEE L.-S. (2018). Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posteriorgram sequences. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 274–281. DOI : 10.1109/SLT.2018.8639647.

ZHANG X., TRMAL J., POVEY D. & KHUDANPUR S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 215–219. DOI : 10.1109/ICASSP.2014.6853589.

ZHAO G., DING S. & GUTIERREZ-OSUNA R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Proc. Interspeech 2019*, p. 2843–2847. DOI : 10.21437/Interspeech.2019-1778.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Voix enfantines, genre et classe sociale : une étude de la fréquence fondamentale

Erwan Pépiot[1]
(1) Laboratoire TransCrit, Université Paris 8, 2 rue de la liberté, 93526 Saint-Denis, France
erwan.pepiot@free.fr

RESUME

Cette étude porte sur les productions d'enfants francophones francilien·nes âgé·es de 8 à 10 ans, en lecture et en parole semi-spontanée. Deux groupes ont été enregistrés : des élèves d'une école privée favorisée (10 filles et 8 garçons), et des enfants scolarisés dans une école publique défavorisée (8 filles et 10 garçons). La F0 moyenne et la modulation de F0 ont été analysées. Les filles présentent une F0 moyenne significativement plus élevée que les garçons dans les deux écoles. La différence étant légèrement plus marquée chez les enfants de l'école favorisée. Aucune corrélation significative n'a été trouvée entre la taille des locuteur·rices et leur F0 moyenne. La modulation de F0 est significativement plus élevée chez les filles au sein de l'école favorisée, mais très similaire entre les deux genres dans l'école défavorisée. Indépendamment du genre, la modulation de F0 est plus forte chez les enfants issus de milieu favorisé.

ABSTRACT

Child speech, gender and social class: a study of fundamental frequency
This study deals with the productions of 8- to 10-year-old French-speaking children living in Paris area in a reading task and in semi-spontaneous speech. Two groups of speakers were recorded: pupils from an upper-class private school (10 girls / 8 boys), and children studying in a lower-class public school (8 girls / 10 boys). Mean fundamental frequency and F0 modulation were measured. Results show that girls from both schools had a significantly higher average F0 than boys. This difference was slightly more pronounced among children from the privileged school. There was no significant correlation found between the speakers' height and their average F0. F0 modulation was significantly higher in girls for the privileged school but was very similar between both genders in the disadvantaged school. Regardless of gender, F0 modulation was stronger in children from privileged backgrounds.

MOTS-CLES : sociophonétique, fréquence fondamentale, intonation, genre, classe sociale, voix enfantines, parole enfantine, français parisien.

KEYWORDS: sociophonetics, fundamental frequency, intonation, gender, social class, children's voices, children's speech, Parisian French.

1 Introduction

La question des différences genrées dans la voix et la parole a été largement étudiée au cours des dernières décennies. Une partie importante de ces recherches se concentre sur la fréquence fondamentale (F0), bien souvent considérée, avec les fréquences de résonance, comme un paramètre crucial dans ce qui constitue une voix féminine ou masculine.

Chez les personnes adultes, la fréquence fondamentale des voix de femmes se situe généralement dans des fréquences plus élevées que celles des hommes, avec des valeurs moyennes se situant respectivement autour de 120 Hz et de 200 Hz (Boë et al., 1975). Ces différences acoustiques sont pour partie dues aux différences sexuées qui émergent lors de la puberté sur les appareils phonatoires. Les taux d'estrogènes, de testostérone et de progestérone varient en fonction de la sexuation du corps et entrainent le développement de plis vocaux plus massifs dans les corps de sexe masculin que dans ceux de sexe féminin (Kahane, 1978 ; Abitbol et al., 1999). Cela explique en partie pourquoi les plis vocaux des hommes vibrent généralement à une fréquence plus basse que ceux des femmes.

La plage de variation et les modulations de F0 sont elles aussi souvent décrites comme dépendantes du genre : les femmes auraient tendance à utiliser des plages de variations plus étendues que les hommes et moduleraient plus leur fréquence fondamentale (p. ex. Austin, 1965 ; Lakoff, 1975, p. 55). Les données varient sensiblement d'une langue à l'autre et selon la méthode utilisée : Henton (1989 ; 1995) a montré qu'en se basant sur une échelle en demi-tons (et non en Hertz), c'est-à-dire une mesure qui reflète la perception humaine des variations de hauteur, alors les différences entre femmes et hommes s'effacent sur l'anglais américain. Pépiot (2014a), en utilisant le même procédé de mesure que Henton a cependant trouvé que les locutrices françaises modulaient significativement plus que les locuteurs français.

Les facteurs sociaux tiennent donc ici un rôle essentiel. Il est désormais clairement établi que les différences vocales femmes-hommes, dans toutes leurs dimensions (F0, formants, VOT, etc.) dépendent de la culture et de la langue parlée (Pépiot & Arnold, 2021). La voix participe donc à la construction sociale des identités de genre (Arnold, 2015 ; Pépiot, 2014b) : chaque locuteur·rice dispose ainsi d'un appareil phonatoire d'une forme donnée, mais peut faire un *usage* de cet appareil en fonction de son genre, à travers des pratiques vocales différentes. La voix n'est ainsi jamais le simple reflet d'une anatomie, mais aussi le résultat d'une performance genrée (Arnold, 2016).

Qu'en est-il alors des enfants pré-pubères ? Si les locuteur·rices adultes ont concentré une bonne partie de l'attention des chercheur·ses, les voix enfantines ont été relativement délaissées. Il s'agit pourtant d'une population particulièrement intéressante à étudier lorsque l'on se penche sur la voix genrée, car contrairement aux adultes, les filles et les garçons pré-pubères ne présentent pas de différence anatomique importante au niveau de l'appareil vocal et toute disparité vocale filles/garçons pourrait donc relever uniquement de conduites articulatoires ayant une origine sociale et culturelle.

Les études sur ces populations sont souvent contradictoires. Plusieurs font état de différences significatives entre filles et garçons, c'est notamment le cas de Hasek et al. (1980), dès 7 ans, et de Whiteside & Hodgson (2000), dès 10 ans ; ces deux études étant menées sur des anglophones. A l'inverse, Bennett (1983), sur des anglophones âgé·es de 7 à 11 ans, et Cornut et al. (1971), sur des francophones âgé·es de 5 à 9 ans, obtiennent des valeurs moyennes de F0 très similaires pour les deux genres. Notons que peu de données sont disponibles quant à la modulation de F0 en fonction du genre chez les enfants.

D'autre part, l'interaction entre le genre et la classe sociale dans les productions vocales enfantines n'a à notre connaissance pas fait l'objet de recherches spécifiques. Or l'on sait que le milieu socio-culturel des locuteur·rices peut grandement influencer leurs productions orales (Labov, 2006), même si la majorité des recherches effectuées à ce sujet portent sur le niveau segmental.

Nous avons donc souhaité, à travers la présente étude, nous pencher sur les pratiques d'enfants francophones pré-pubères, issu·es de deux milieux sociaux très éloignés, en mesurant leur F0 moyenne ainsi que la modulation de leur fréquence fondamentale (plage de variation et écart-type). Nous étudierons ainsi la possible influence des facteurs *genre* et *classe sociale* sur ces paramètres acoustiques chez ces populations.

2 Méthode

2.1 Corpus

La présente étude se fonde sur l'analyse d'un corpus en langue française collecté lors de deux tâches distinctes. Ces deux tâches ont permis d'obtenir des séquences de parole lue et de parole semi-spontanée.

La première tâche consistait en la lecture des 10 phrases suivantes : « *Steven a vendu son vélo hier après-midi.* » ; « *Quand il fait froid et qu'il pleut, je préfère rester chez moi.* » ; « *Ma sœur m'a dit qu'elle allait passer demain.* » ; « *Si tu refais ça, j'appelle la police !* » ; « *J'espère juste qu'un jour on pourra en parler.* » ; « *Franny se réveilla en sursaut.* » ; « *Il l'observa pendant une minute entière.* » ; « *Il y avait un silence étrange à l'autre bout du fil.* » ; « *Est-ce que tu veux aller au cinéma ce soir ?* » et « *Où as-tu trouvé ce livre ?* ».

La seconde tâche, visait à solliciter de la parole (semi-)spontanée. Il était ainsi demandé aux participant·es de relater leur journée de la veille, pendant au moins une minute.

2.2 Participant·e·s

Trente-six enfants (18 filles, 18 garçons) scolarisé·es en classe de CM1 et âgé·es de 8 à 10 ans ont pris part à cette étude. La moyenne d'âge globale est de 9 ans et 6 mois (SD = 9 mois). Elle est de 9 ans et 6 mois pour les filles de l'école CSP+ et pour celles de l'école CSP-, de 9 ans et 5 mois pour les garçons de l'école CSP+ et pour ceux de l'école CSP-. La taille moyenne des filles au moment des enregistrements était de 139,3cm, et celle des garçons de 140,7cm. Ces participant·es sont toutes et tous francophones natif·ves, parlent le français à l'école et dans leur famille, et vivaient en Ile-de-France depuis plus de 3 ans. Aucun·e n'a reporté souffrir de trouble de la parole.

Dix-huit enfants étaient scolarisés dans une école publique du nord-est parisien (8 filles & 10 garçons) : cette école figurait à la 3657ème place sur 3772 dans le classement des écoles d'Ile de France par indice de positionnement social (*IPS*) publié par le Ministère de l'Education Nationale et de la Jeunesse en 2022, avec un IPS de 72,2 (échelle allant de 45 à 185). Selon le Ministère, cet indice "*résume les conditions socio-économiques et culturelles des familles des élèves accueillis dans l'établissement.*" Les 18 autres étaient scolarisés dans une école privée de l'ouest parisien (10 filles & 8 garçons), figurant à la 20ème place du même classement, avec un IPS de 151,5. Les deux écoles sont toutes deux situées à une vingtaine de minutes du centre de la capitale par le RER.

La participation aux enregistrements s'est faite sur la base du volontariat. Aucune contrepartie n'a été donnée aux participant·es. Un formulaire de consentement a été distribué aux parents des enfants en amont des sessions d'enregistrement afin de s'assurer de leur accord. Seul·es les enfants disposant à fois l'accord signé de leurs parents et ayant exprimé eux/elles même leur souhait de participer à l'expérience ont été enregistré·es.

2.3 Procédure d'enregistrement

Les enregistrements se sont déroulés directement dans les écoles, dans des pièces calmes et aménagées pour l'occasion. Chaque session d'enregistrement comprenait les tâches détaillées dans la section 2.1 : en premier lieu, la lecture des phrases avec un débit de parole « normal » (deux lectures pour chaque item), puis la narration portant sur la journée de la veille pendant une à deux minutes. Le maximum a été fait afin de mettre à l'aise les enfants et de diminuer leur niveau de stress.

Lors de la lecture des phrases, en cas d'erreur manifeste, il a été demandé aux enfants de relire l'item, de sorte à ce que deux occurrences exploitables soient obtenues pour chaque phrase. Durant la tâche de narration, chaque enfant a produit au moins 30 secondes de parole exploitable et au maximum 2 minutes.

2.4 Analyse des données

L'analyse acoustique des enregistrements recueillis a été effectuée à l'aide du logiciel *Praat* (Boersma, 2017). Les paramètres suivants ont été mesurés pour chacune des phrases ainsi que pour le discours semi-spontané :

- F0 moyenne.
- Plage de variation de F0, qui correspond à l'écart entre la fréquence la plus basse et la fréquence la plus haute atteinte au sein d'une unité linguistique donnée (ici *phrase* ou *discours*).
- Écart-type de F0, qui indique l'ampleur des variations par rapport à la valeur moyenne. Il constitue le paramètre le plus en mesure de rendre compte de la modulation de F0, en particulier lors de l'étude de longues séquences de parole continue.

Ces données ont été obtenues en générant pour chaque phrase/discours un fichier *Pitch* sur Praat, puis en collectant les valeurs dans la fenêtre *Pitch info*. Afin d'éviter toute erreur d'extraction, les seuils de détection ont été ajustés manuellement pour chaque locuteur·rice et tous les fichiers ont été vérifiés individuellement à posteriori. La plage de variation de F0 ainsi que l'écart-type ont été mesurés en Hertz mais aussi en demi-tons. Cette échelle est en effet particulièrement pertinente car elle rend compte de la variation de hauteur perçue (Henton, 1995). Les données ainsi collectées ont ensuite fait l'objet de tests statistiques de type ANOVA et corrélations, dans le but de tester l'influence du *genre*, de l'école (i.e. du *milieu social*) et de la *taille* des enfants.

3 Résultats

3.1 Phrases lues

La F0 moyenne des locuteur·rices sur les phrases lues est présentée dans le tableau 1 ci-après.

F0 MOYENNE - PHRASES LUES

Ecole CSP-		Ecole CSP+	
Loc.	**F0 moyenne (Hz)**	**Loc.**	**F0 moyenne (Hz)**
F1-	260	**F1+**	295
F2-	244	**F2+**	273
F3-	276	**F3+**	276
F4-	261	**F4+**	252
F5-	264	**F5+**	250
F6-	257	**F6+**	252
F7-	249	**F7+**	278
F8-	313	**F8+**	266
Moy. F	**265,44**	**F9+**	239
G1-	202	**F10+**	256
G2-	250	**Moy. F**	**263,76**
G3-	262	**G1+**	233
G4-	286	**G2+**	225
G5-	213	**G3+**	229
G6-	217	**G4+**	245
G7-	256	**G5+**	235
G8-	223	**G6+**	255
G9-	219	**G7+**	210
G10-	262	**G8+**	248
Moy. G	**239,00**	**Moy. G**	**235,08**

TABLEAU 1 : F0 moyenne en Hertz (Hz) des locutrices et des locuteurs sur les phrases lues (10 x 2 occurrences par participant·e), en fonction du genre (fille -*F*- et garçon -*H*-) et de l'école dans laquelle les enfants sont scolarisé·es.

On constate que pour les deux écoles, les filles présentent une F0 moyenne plus élevée que les garçons. Les chiffres obtenus sont très proches d'une école à l'autre, même si la différence entre les genres semble légèrement plus marquée dans l'école privée dite *CSP+* (28,68Hz de différence entre les deux groupes) que dans l'école publique que nous appellerons *CSP-* (26,44Hz).

Une ANOVA confirme l'influence significative du facteur *genre* sur la F0 moyenne, indépendamment de l'école, avec $F(1,718)=281,476$ avec $p<0,0001$. En considérant séparément les données des deux écoles, la différence filles/garçons demeure largement significative, tant pour l'école CSP- ($F(1,358)=101,982$; $p<0,0001$) que pour l'école CSP+ ($F(1,358)=216,086$; $p<0,0001$). Enfin, l'analyse de détecte pas d'interaction significative entre les facteurs *genre* et *école* ($F(1,716)=0,466$; $p=0,4949$).

Afin de tester si cette différence inter-genres ne relèverait pas de différences anatomiques, un test de corrélation de Pearson a été mené entre la *taille des locuteur·rices* (qui donne une indication de la longueur de leurs plis vocaux) et leur *fréquence fondamentale moyenne* sur l'ensemble des phrases lues : le test ne fait état d'aucune corrélation significative ($r(36)=0,144$; $z=0,830$ avec $p=0,4064$).

La plage de variation de F0, en Hertz et demi-tons, ainsi que l'écart-type moyen de F0 (SD) en Hertz (Hz) et demi-tons (dt) sur les phrases lues sont visibles ci-dessous, dans le tableau 2.

VARIATION DE F0 - PHRASES LUES

Ecole CSP-					Ecole CSP+				
Loc.	**Range (Hz)**	**Range (dt)**	**SD (Hz)**	**SD (st)**	**Loc.**	**Range (Hz)**	**Range (dt)**	**SD (Hz)**	**SD (st)**
F1-	121	**7,51**	21	**1,36**	**F1+**	168	**9,97**	37	**2,12**
F2-	105	**7,32**	16	**1,12**	**F2+**	157	**9,67**	29	**1,82**
F3-	133	**8,01**	23	**1,40**	**F3+**	181	**11,41**	38	**2,37**
F4-	168	**11,85**	33	**2,25**	**F4+**	125	**8,11**	27	**1,76**
F5-	124	**8,33**	23	**1,50**	**F5+**	130	**8,85**	24	**1,67**
F6-	114	**7,85**	19	**1,27**	**F6+**	127	**8,88**	22	**1,57**
F7-	109	**7,82**	22	**1,55**	**F7+**	179	**12,04**	37	**2,44**
F8-	179	**9,52**	36	**1,93**	**F8+**	130	**8,28**	28	**1,77**
Moy. F	**131,65**	**8,53**	**24,03**	**1,55**	**F9+**	173	**13,15**	37	**2,66**
G1-	99	**7,98**	16	**1,31**	**F10+**	151	**11,01**	28	**1,94**
G2-	101	**6,83**	16	**1,08**	**Moy. F**	**152,02**	**10,14**	**30,71**	**2,01**
G3-	161	**10,99**	31	**2,06**	**G1+**	104	**7,79**	21	**1,59**
G4-	158	**10,41**	33	**2,07**	**G2+**	132	**9,99**	27	**2,09**
G5-	84	**6,76**	14	**1,13**	**G3+**	142	**10,98**	31	**2,29**
G6-	105	**8,23**	16	**1,25**	**G4+**	107	**7,53**	20	**1,39**
G7-	129	**8,44**	22	**1,49**	**G5+**	132	**10,15**	27	**2,02**
G8-	111	**8,49**	22	**1,66**	**G6+**	114	**7,72**	22	**1,49**
G9-	120	**9,57**	25	**1,96**	**G7+**	100	**7,94**	19	**1,53**
G10-	141	**9,27**	24	**1,57**	**G8+**	133	**9,11**	26	**1,77**
Moy. G	**120,98**	**8,70**	**21,78**	**1,56**	**Moy. G**	**120,45**	**8,90**	**24,09**	**1,77**

TABLEAU 2 : Valeurs moyennes de la plage de variation de F0 (*Range* ; en Hz et dt) et de l'écart-type de F0 (*SD* ; en Hz et dt) sur chaque phrase lue (10 x 2 occurrences par enfant) en fonction du genre (fille -*F*- et garçon -*H*-) et de l'école dans laquelle les enfants sont scolarisé·es.

Les données recueillies sur ces indicateurs de modulation de F0 montrent une assez forte disparité entre les deux écoles. Globalement, on retrouve plus de modulation (range et SD) chez les enfants de l'école CSP+. Au plan des disparités filles/garçons (avec les valeurs en demi-tons), l'on constate des résultats très proches dans les deux genres pour l'école CSP- mais une assez nette différence dans l'école CSP+, avec plus de modulation chez les filles : la plage de variation à l'échelle de la phrase est

en moyenne 1,24dt plus élevée que chez les garçons (+14%), et l'écart-type est également de 14% plus élevé chez les locutrices.

Une ANOVA a été conduite sur la plage de variation de F0 (en dt) afin de tester l'influence du facteur *genre*. Chez les enfants de l'école CSP-, aucune différence significative filles/garçons n'est détectée ($F(1,358)=0,595$; $p=0,4409$). Chez les élèves de l'école CSP+, l'effet du genre est en revanche très significatif : $F(1,358)=24,622$; $p<0,0001$. Il en va de même sur l'écart-type, exprimé en demi-tons : pas d'influence significative du genre pour l'école CSP- ($F(1,358)=0,052$; $p=0,8203$) mais une influence forte et très significative pour l'école CSP+ ($F(1,358)=17,819$; $p<0,0001$). Une ANOVA à deux facteurs conduite sur l'ensemble des données relève d'ailleurs une interaction significative entre le *genre* et l'*école*, tant sur le *range* ($F(1,716)=17,838$; $p<0,0001$) que sur le *SD* ($F(1,716)=11,727$; $p=0,0007$).

3.2 Parole semi-spontanée

Comme expliqué précédemment, les locutrices et locuteurs ont également eu à produire des séquences de parole semi-spontanée. La fréquence fondamentale moyenne des participant·e·s lors de ces séquences, d'une durée moyenne d'1 minutes et 3 secondes, sont présentées dans le tableau 3.

F0 MOYENNE - DISCOURS SEMI-SPONTANE

Ecole CSP-		Ecole CSP+	
Loc.	**F0 moyenne (Hz)**	**Loc.**	**F0 moyenne (Hz)**
F1-	253	**F1+**	297
F2-	234	**F2+**	256
F3-	259	**F3+**	263
F4-	250	**F4+**	245
F5-	238	**F5+**	256
F6-	238	**F6+**	249
F7-	252	**F7+**	283
F8-	290	**F8+**	260
Moy. F	**251,75**	**F9+**	208
G1-	188	**F10+**	231
G2-	250	**Moy. F**	**254,80**
G3-	253	**G1+**	216
G4-	249	**G2+**	223
G5-	217	**G3+**	216
G6-	208	**G4+**	238
G7-	244	**G5+**	201
G8-	202	**G6+**	248
G9-	198	**G7+**	224
G10-	254	**G8+**	233
Moy. G	**226,30**	**Moy. G**	**224,88**

TABLEAU 3 : F0 moyenne des locutrices et des locuteurs sur le discours semi-spontané, en fonction du genre (fille -*F*- et garçon -*H*-) et de l'école dans laquelle les enfants sont scolarisé·es.

Les résultats obtenus semblent confirmer les tendances observées sur les phrases lues, mais avec des F0 moyens globalement plus bas, indépendamment du genre et de la classe sociale. Les garçons des deux écoles présentent ici encore une fréquence fondamentale moyenne nettement inférieure à celle des filles dans les deux écoles. Dans le détail, on constate qu'à nouveau, cette disparité filles/garçons est plus marquée dans l'école CSP+ (différence de 29,92 Hz) que dans l'école CSP- (25,45 Hz).

L'ANOVA confirme l'influence significative du genre, en considérant l'ensemble des participant·es avec $F(1,70)=31,406$ et $p<0,0001$. En prenant séparément les deux écoles, la différence filles/garçons est significative également, tant pour l'école CSP- ($F(1,34)=11,703$; $p=0,0016$) que pour l'école CSP+, avec pour cette dernière une influence du facteur *genre* plus marquée ($F(1,34)=19,155$; $p=0,0001$). L'ANOVA à deux facteurs ne détecte pas d'interaction significative entre le *genre* et l'*école* ($F(1,68)=0,196$; $p=0,6593$).

Un test de corrélation de Pearson a également été réalisé entre la taille des enfants et leur F0 moyenne. Le test de décèle aucune corrélation significative ($r(720)=-0,269$; $z=-1,584$ avec $p=0,1133$) et vient donc confirmer que ces différences sur la F0 moyenne proviennent de conduites articulatoires socialement construites plutôt que de différences anatomiques.

Le tableau 4, ci-après, présente la plage de variation de F0 (en Hz et dt), ainsi que l'écart-type (*SD* - également mesurée en Hz et dt) sur les séquences de parole semi-spontanée.

VARIATION DE F0 - DISCOURS SEMI-SPONTANEE

Ecole CSP-					Ecole CSP+				
Loc.	**Range (Hz)**	**Range (dt)**	**SD (Hz)**	**SD (st)**	**Loc.**	**Range (Hz)**	**Range (dt)**	**SD (Hz)**	**SD (st)**
F1-	218	**12,67**	24	**1,56**	**F1+**	298	**16,74**	35	**1,99**
F2-	264	**14,9**	23	**1,57**	**F2+**	304	**16,76**	36	**2,16**
F3-	279	**15,44**	33	**1,95**	**F3+**	367	**20,72**	52	**2,95**
F4-	330	**20,31**	45	**2,78**	**F4+**	254	**15,52**	27	**1,86**
F5-	275	**15,93**	30	**1,94**	**F5+**	328	**20,02**	47	**2,82**
F6-	241	**15,52**	28	**1,90**	**F6+**	258	**17,08**	34	**2,36**
F7-	202	**12,98**	27	**1,77**	**F7+**	234	**13,5**	35	**2,22**
F8-	318	**17,56**	46	**2,58**	**F8+**	210	**12,31**	28	**1,91**
Moy. F	**265,84**	**15,66**	**32,16**	**2,01**	**F9+**	258	**20,12**	27	**2,18**
G1-	136	**11,09**	18	**1,61**	**F10+**	263	**17,74**	37	**2,58**
G2-	175	**10,59**	21	**1,44**	**Moy. F**	**277,37**	**17,05**	**35,70**	**2,30**
G3-	253	**14,86**	43	**2,65**	**G1+**	168	**12,61**	23	**1,75**
G4-	268	**16,37**	35	**2,29**	**G2+**	188	**13,33**	34	**2,32**
G5-	172	**12,83**	24	**1,81**	**G3+**	185	**13,27**	29	**2,24**
G6-	125	**9,89**	15	**1,21**	**G4+**	169	**11,47**	22	**1,54**
G7-	243	**14,46**	32	**2,03**	**G5+**	194	**14,18**	22	**1,73**
G8-	224	**15,93**	18	**1,42**	**G6+**	259	**14,92**	30	**1,98**
G9-	231	**16,4**	29	**2,25**	**G7+**	248	**16,92**	39	**2,74**
G10-	195.1	**12.67**	27.89	**1.84**	**G8+**	206	**13,35**	27	**1,86**
Moy. G	**203,04**	**13,60**	**26,09**	**1,86**	**Moy. G**	**201,99**	**13,76**	**28,23**	**2,02**

TABLEAU 4 : Valeurs moyennes de la plage de variation de F0 (*Range* ; en Hz et dt) et de l'écart-type de F0 (*SD* ; en Hz et dt) en fonction du genre (fille -*F*- et garçon -*H*-) et de l'école dans laquelle les enfants sont scolarisé·es.

A l'instar des phrases lues, on observe sur la parole spontanée une modulation globalement plus élevée chez les élèves de l'école CSP+, indépendamment du genre. Concernant les disparités filles/garçons, on constate une modulation plus forte en moyenne chez les filles dans les deux écoles, mais avec un contraste bien plus marqué au sein de l'école CSP+ : la différence y atteint +24% sur le *range* en demi-tons (+15% dans l'autre école) et +14% sur l'écart-type (+8% dans l'école CSP-).

L'ANOVA effectuée sur la plage de variation de F0 en demi-tons fait état d'une différence filles/garçons très significative pour l'école CSP+ ($F(1,34)=18,546$; $p<0,0001$). Dans la seconde école, l'effet du facteur *genre* est également significatif, bien que plus limité ($F(1,34)=7,320$; $p=0,0106$). La même analyse réalisée sur l'écart-type (dt) indique à nouveau une différence inter-

genres significative pour l'école CSP+ : $F(1,34)=5,241$; $p=0,0284$. Chez les élèves de l'école CSP-, aucune différence significative n'est décelée : $F(1,34)=1,067$; $p=0,3089$.

4 Conclusion / discussion

Les données recueillies à travers cette étude sont éclairantes à plusieurs titres. Elles révèlent à la fois des variations inter-genres sur la F0 moyenne et sur la modulation de F0 chez des enfants pré-pubères mais aussi une influence de la classe sociale sur leurs productions vocales.

Concernant la F0 moyenne, des différences genrées significatives ont été observées, avec une fréquence fondamentale moyenne supérieure de l'ordre de 25 à 30 Hz chez les filles, tant sur les phrases lues que sur la parole spontanée, et ce chez les enfants des deux écoles. Ces variations ne semblent pas avoir pour origine des différences anatomiques sexuées. En effet, à cet âge (8 à 10 ans), les transformations majeures de l'appareil vocal des locuteur·rices n'ont pas encore eu lieu. La taille des enfants pourrait être impliquée car l'on sait qu'elle est proportionnelle à la longueur des plis vocaux, mais aucune corrélation significative n'a été trouvée entre la taille des locuteur·rices et leur F0 moyenne, ce qui permet d'écarter cette hypothèse. Cela suggère donc que les enfants pré-pubères auraient tendance à mettre en place des pratiques vocales genrées en adaptant la hauteur de leur voix afin d'aller dans le sens des différences observées chez les adultes. Il est toutefois important de noter que ces différences sont bien moins importantes que celles présentent généralement à l'âge adulte. On remarque enfin que l'ampleur de cette différence genrée est légèrement plus grande au sein des élèves de l'école privée favorisée.

Pour ce qui est des modulations de F0, une différence inter-genres marquée et significative est ressortie au sein des élèves de l'école CSP+. La plage de variation et l'écart-type de F0 ont en effet été significativement plus grands chez les filles que chez les garçons, tant en lecture qu'en discours semi-spontané. Ces résultats vont dans le sens de ceux de Pépiot (2014a) sur des adultes francophones, et suggèrent que dans cette langue les modulations de F0 font partie des pratiques vocales qui varient en fonction du genre. Ces différences pourraient donc émerger avant même la puberté chez les enfants issus de milieux favorisés. Chez les élèves de l'école CSP-, en revanche, les mesures faites sur ces paramètres acoustiques sont très proches chez les filles et les garçons, seule la plage de variation en discours spontané est significativement plus grande chez les filles. L'émergence de cette pratique vocale genrée avant la puberté pourrait donc est un phénomène dépendant de la classe sociale.

Par ailleurs, indépendamment du genre, on constate que les élèves de l'école favorisée ont eu tendance à nettement plus moduler leur F0 que celleux de l'école publique défavorisée. Cela est vrai à la fois pour la plage de variation de F0 et pour son écart-type et quel que soit le type de parole (lue ou spontanée). Ainsi, le fait de moduler fortement la hauteur de sa voix en français pourrait également constituer un marqueur de classe sociale. Une telle hypothèse pourrait être confirmée en répliquant l'expérience chez des locuteur·rices adultes. Cette tendance pourrait aussi s'expliquer par une plus grande aisance et confiance en soi au sein de ce groupe durant les enregistrements.

Cette étude conforte donc l'idée que la F0 n'est pas une caractéristique *essentielle* des locuteur·rices, dépendant uniquement de la forme de leur appareil phonatoire, mais qu'elle résulte également d'un apprentissage et d'une socialisation en tant que membre d'une catégorie de genre et d'une classe sociale spécifique. Cela constitue un nouvel argument pour s'éloigner d'une conception purement *anatomiste* de la F0 que l'on retrouve communément dans la littérature phonétique, et pour une plus grande considération des facteurs sociaux dans l'étude de la voix et de la parole.

Parmi les limites de cette étude, on citera notamment le nombre relativement restreint de participant·es. De plus, cette recherche se limite actuellement au français, il pourrait être intéressant de la répliquer dans d'autres langues.

Remerciements

Un grand merci à Lou Jullien, qui a effectué les enregistrements des enfants de l'école publique dans le cadre de son mémoire de Master à l'ENS Louis Lumière et a réalisé la préparation des fichiers sons en vue de leur analyse acoustique. Merci également à tou·te·s les enfants ayant pris part à cette étude, ainsi qu'au personnel des écoles ayant facilité l'organisation des sessions d'enregistrement.

Références

ABITBOL J., ABITBOL P., ABITBOL B. (1999). Sex hormones and the female voice. *Journal of Voice* 13, 424–446.

ARNOLD A. (2015). Voix et transidentité : changer de voix pour changer de genre ?. *Langage et société* 151(1), 87−105.

ARNOLD A. (2016). Voix. *Encyclopédie critique du genre*, 713−721. Paris : La Découverte.

AUSTIN W. M. (1965). Some social aspects of paralanguage. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 11(1), 31−39.

BENNETT S. (1983). A 3-year longitudinal study of school-aged children's fundamental frequencies. *Journal of Speech and Hearing Research* 26, 137-142.

BOERSMA P., WEENINK D. (2017). Praat: doing phonetics by computer [Logiciel]. Version 6.0.36, publiée le 11 Novembre 2017 sur le site www.praat.org

BOË L.-J., CONTINI M., RAKOTOFIRINGA H. (1975). Etude statistique de la fréquence laryngienne. *Phonetica* 32(1), 1-23.

CORNUT G., RIOU-BOURRET V. & LOUIS M. H. (1971). Contribution à l'étude de la voix parlée et chantée de l'enfant normal de 5 à 9 ans. *Folia Phoniatrica et Logopaedica* 23(6), 381-389.

HASEK, C. S., SINGH, S., MURRY T. (1980). Acoustic attributes of preadolescent voices. *The Journal of the Acoustical Society of America* 68(5), 1262-1265.

HENTON C. (1989). Fact and fiction in the description of female and male pitch. *Language & Communication*, 9(4), 299-311.

HENTON C. (1995). Pitch dynamism in female and male speech. *Language & Communication* 15(1), 43-61.

KAHANE J. C. 1978. A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy* 151, 11-19.

LABOV, W. (2006). A sociolinguistic perspective on sociophonetic research, *Journal of Phonetics* 34, 500-515.

LAKOFF R. (1975). *Language and Woman's Place*. New York : Harper & Row.

PEPIOT E. (2014a). Male and female speech: a study of mean F0, F0 range, phonation type and speech rate in Parisian French and American English speakers. *Proceedings of the 7th International Conference on Speech Prosody*, 305-309.

PEPIOT, E. (2014b). Voix et genre : un état de la question. In Ibrahim, A.H. (éd.), *La langue, la voix, la parole* (pp. 53-86), Paris : CRL.

PEPIOT E. & ARNOLD A. (2021). Cross-gender differences in English/French bilingual speakers: A multiparametric study. *Perceptual and Motor Skills* 128(1), 153-177.

WHITESIDE S. P. & HODGSON C. (2000). Some acoustic characteristics in the voices of 6- to 10-year-old children and adults: a comparative sex and developmental perspective. *Logopedics Phoniatrics Vocology* 25(3), 122-132.

Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

iHist et iScatter, outils en ligne d'exploration interactive de données : application aux valeurs aberrantes de f0 et de formants

Nicolas Audibert[1]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle),
4 rue des Irlandais, 75005 Paris, France
nicolas.audibert@sorbonne-nouvelle.fr

RESUME

Les mesures aberrantes d'un point de vue statistique (outliers) doivent être traitées avec précaution, ce qui peut être compliqué en pratique lorsque la quantité de données devient importante. Afin de faciliter l'inspection des valeurs situées à la marge des distributions, nous proposons deux outils développés avec R/Shiny, disponibles sous forme d'applications en ligne utilisables par des non-spécialistes et distribués gratuitement sous licence GPL. Ces applications permettent de paramétrer la visualisation et d'explorer de façon interactive des distributions via des histogrammes, et les relations entre variables quantitatives via des nuages de points. Deux cas d'utilisation appliqués à des données de parole sont présentés pour illustrer les principales fonctionnalités de ces outils, à partir de mesures acoustiques extraites par Praat : l'ajustement des valeurs limites pour la détection automatique de la fréquence fondamentale, et l'identification de valeurs erronées de formants.

ABSTRACT

Online tools for interactive exploration of speech data: application to outliers in f0 and formants values.
Statistical outliers need to be handled with care, which can be difficult with large datasets. To facilitate the inspection of values at the margins of distributions, we offer two tools developed with R/Shiny, available as online applications for use by non-specialists and freely distributed under the GPL license. These applications let the user set visualization parameters and interactively explore distributions via histograms, and relationships between quantitative variables via scatterplots. Two use cases applied to speech data are presented to illustrate the main functionalities of these tools, based on acoustic measurements extracted with Praat: adjustment of limit values applicable to automatic fundamental frequency detection, and identification of erroneous formant values.

MOTS-CLES : données, valeurs aberrantes, outils en ligne, exploration interactive, f0, formants.
KEYWORDS : data, outliers, online tools, interactive exploration, f0, formants.

1 Introduction

L'analyse de données de parole nécessite fréquemment un « nettoyage » des données afin d'éliminer les mesures erronées, tout particulièrement dans le cas de mesures acoustiques (semi-)automatisées, et ce d'autant plus que les ensembles de données traités sont de grande taille. Plusieurs approches

ont pu être proposées pour cela, la plus courante étant fondée sur la définition de seuils au-delà desquels les valeurs sont considérées comme erronées et donc éliminées. En psycholinguistique, ces seuils sont généralement définis à partir de critères statistiques, le plus souvent en éliminant les valeurs situées à plus de deux écarts-types de la moyenne sous l'hypothèse d'une distribution normale (ce qui revient à éliminer les 2,3% des valeurs les plus petites et les 2,3% les plus grandes). Notons toutefois que de tels seuils statistiques sont fortement dépendants de la distribution des données : dans le cas de mesures de durée pour lesquelles il est courant d'observer des distributions asymétriques, l'utilisation de ce critère qui implique un postulat de normalité conduirait ainsi à considérer à tort des valeurs élevées comme déviantes et des valeurs faibles comme non-déviantes.

Parmi les travaux récents qui se sont penchés sur la question des stratégies applicables face aux données extrêmes, on peut mentionner ceux de Nicklin & Plonski (2020) qui se concentrent plus spécifiquement sur le cas des tâches de lecture en acquisition L2 et comparent deux stratégies statistiques de résolution, l'élagage des valeurs extrêmes ou leur remplacement par des valeurs limites (*winsorizing*), concluant que le choix de l'une ou l'autre de ces stratégies a peu d'incidence sur les résultats. Quant à eux, Osborne & Overbay (2019) soulignent que si l'élimination des valeurs aberrantes permet d'estimer les effets étudiés avec à la fois plus de précision et de robustesse, toutes les valeurs extrêmes ne doivent pas être considérées comme aberrantes et inversement. Si une telle approche se justifie lorsque l'objectif est de caractériser une tendance majoritaire dans un groupe ou une condition expérimentale, les valeurs « déviantes » d'un point de vue statistique peuvent être tout aussi informatives à condition qu'elles ne relèvent pas simplement d'erreurs de liées au recueil des données ou aux mesures effectuées. Par ailleurs les erreurs ou biais de mesure ne sont pas toujours simples à caractériser à partir de la définition de seuils, notamment lorsque ces seuils doivent être établis sur la base de valeurs normatives. En effet dans le cadre de données de parole, de telles normes lorsqu'elles existent et ne se limitent pas à des valeurs moyennes sont souvent difficilement transposables à d'autres conditions de production que celles utilisées pour les établir.

S'il semble raisonnable de recommander une inspection systématique de l'intégralité des données pour identifier les cas qui constituent des erreurs, et le cas échéant distinguer parmi ces erreurs celles susceptibles de faire l'objet d'une correction des cas à éliminer de l'analyse, une telle approche est d'autant plus difficilement applicable que le volume de données traité est important. Or, avec l'évolution des moyens techniques dans les dernières décennies et l'accessibilité croissante des méthodes automatiques en sciences de la parole, la taille des corpus analysés tend à augmenter, notamment lorsque ces corpus consistent en des enregistrements acoustiques. Si cette augmentation de la taille des données a le mérite de permettre la prise en considération dans les travaux en sciences de la parole de phénomènes propres à la parole continue (Liberman, 2019) elle peut aussi accroitre le risque de ne considérer les données de parole que comme des ensembles de valeurs numériques en atténuant voire en perdant le lien direct avec l'interprétation phonétique de ces données.

Lorsque le volume de données devient conséquent et rend impossible l'inspection de l'intégralité des données, on peut recommander d'effectuer ces vérifications sur un échantillon aléatoire, et de façon plus systématique sur certaines plages de valeurs définies à partir de l'observation de la distribution de l'ensemble des données pour un groupe de locuteurs ou une condition particulière. Pour certaines analyses, les données à explorer en priorité peuvent être définies par la combinaison des valeurs prises par plusieurs variables plutôt qu'à partir de la distribution d'une variable unique, afin d'identifier les observations qui ne suivent pas la tendance majoritaire.

Dans de tels cas, un outil tel que le logiciel R combiné au regroupement de paquets *tidyverse* (Wickham et al., 2019) - ou des bibliothèques Python telles que *pandas* ou *Polars* - peut s'avérer précieux pour déterminer quels sous-ensembles de données nécessitent un examen plus approfondi.

Cependant et bien que l'initiation à l'utilisation de ces outils soit de plus en plus largement intégrée dans les formations initiales en linguistique et considérée comme un prérequis pour des formations plus avancées en statistiques, on peut constater que dans les faits leur utilisation reste souvent restreinte à des cas d'utilisation spécifique (le plus souvent la réalisation de tests statistiques), l'exploration des données s'appuyant sur des tableurs moins adaptés au jeux de données de taille conséquente. C'est ce constat qui a en partie motivé le développement des outils présentés dans cet article, initialement mis en place pour l'enseignement des statistiques à un public d'étudiants en licence et master en sciences du langage puis étendus ensuite afin de pouvoir être utiles à la communauté de recherche en sciences de la parole. Les deux applications présentées, qui permettent aux utilisateurs de téléverser leur propre jeu de données afin de l'explorer via des visualisations paramétrables, ont été développées à l'aide du logiciel R (R Core team, 2024) et du paquet shiny (Chang et al., 2024), et peuvent être utilisées directement en ligne ou localement. Nous présentons ici certaines fonctionnalités de ces outils, en les illustrant à travers des applications potentielles à des mesures acoustiques parmi celles les plus fréquemment utilisées dans les études phonétiques, la fréquence fondamentale f0 et les formants.

2 Limites des mesures automatiques de f0 et de formants

2.1 Fréquence fondamentale (f0)

La mesure de la fréquence fondamentale est réputée fiable sur des enregistrements acoustiques de parole normophonique réalisés dans des conditions permettant d'obtenir un rapport signal/bruit élevé. Toutefois, les résultats en condition non-bruitée de l'évaluation réalisée par Jouvet & Laprie (2017) indiquent que même dans ces conditions supposées idéales, l'ensemble des algorithmes évalués avec les paramètres par défaut commettent des erreurs concernant la décision de voisement ou divergent de plus de 20% de la f0 de référence (taux global d'erreur FFE compris entre 2,5% et 13% des trames analysées selon les locuteurs et algorithmes). Notons qu'une étude récente (Vaysse et al., 2022) visant à comparer les performances d'algorithmes de f0 sur des voix pathologiques suggère de combiner deux algorithmes évalués comme optimaux pour respectivement la décision de voisement et l'estimation de f0 afin d'améliorer les performances globales de la détection de f0, ce qui pourrait conduire à un taux d'erreur plus faible également sur la parole normophonique.

La majorité de ces algorithmes de détection de f0 étant paramétrables, il est possible d'obtenir de meilleures performances en ajustant ces paramètres pour tenir compte des spécificités des extraits de parole analysés, notamment en restreignant la plage de valeurs de f0 considérée comme plausibles en fonction du registre propre à chaque locuteur ou enregistrement. Ce constat est à l'origine de la méthode en deux étapes de De Looze (2010), qui propose d'effectuer avec l'algorithme de Praat (Boersma, 1993) une première passe de détection avec une plage de valeurs large (60Hz-600Hz), puis de procéder à une seconde passe de détection avec une plage de valeurs plus restreinte définie à partir de quantiles de la distribution des mesures obtenues lors de la première passe (minimum et maximum fixés respectivement à $0.83*Q_{15\%}$ et $1.92*Q_{65\%}$ des valeurs initiales). Si l'utilisation de valeurs limites ainsi définies permet d'améliorer sensiblement les performances comparativement à l'utilisation de valeurs par défaut, elle repose sur le postulat que la forme de la distribution obtenue lors de la première passe est comparable entre locuteurs et conditions de production.

Pour l'analyse de productions de parole et de chant dirigés vers l'enfant, Falk & Audibert (2021) ont adopté une méthodologie similaire avec une première passe de détection réalisée avec une large

plage de valeurs, mais dans laquelle les valeurs limites ont ensuite été définies pour chaque locutrice*condition à partir de l'inspection visuelle de la distribution. C'est cette dernière approche que nous illustrons à l'aide de l'outil dédié à l'affichage d'histogrammes interactifs.

2.2 Formants

Bien que la mesure automatique des formants ait fait l'objet de nombreux travaux et quand bien même l'analyse se limite à la partie supposée stable des voyelles, une telle tâche reste sujette à de nombreux biais, liés entre autres aux interactions possibles entre fréquence fondamentale et fréquence formantique (cf. Kent & Vorperian (2018) pour un récapitulatif des principaux biais). En raison de ces limites, de nombreuses études ont eu recours à une validation et correction manuelle à partir de l'inspection des spectrogrammes des relevés de fréquences formantiques effectués automatiquement (voir par exemple Van der Harst et al., 2014). Par ailleurs certains auteurs ont opté en remplacement des mesures formantiques pour une paramétrisation du signal de parole à travers des coefficients cepstraux afin de permettre l'automatisation des mesures (Ferragne & Pellegrino, 2010), ou encore pour rendre compte de variations sur les voyelles nasales pour lesquelles la détection automatique de formants est notoirement sujette à erreurs (Hermes et al., 2023), avec toutefois l'inconvénient d'une perte d'interprétabilité du lien avec l'articulation.

Une autre approche applicable à des corpus de grande taille consiste en la méthode proposée par Gendrot & Adda-Decker (2005). Il s'agit de définir un crible permettant d'éliminer les valeurs supposées aberrantes en prenant en considération la catégorie phonologique à laquelle appartient le segment analysé (généralement une voyelle) pour définir des intervalles entre lesquels la détection automatique est considérée comme correcte. La difficulté est de définir un crible suffisamment large pour ne pas assimiler à des erreurs de détection des mesures formantiques qui reflètent simplement la variabilité inhérente à la parole, notamment la parole conversationnelle susceptible de subir d'importants phénomènes de réduction segmentale. Au-delà d'erreurs grossières comme la non-détection du second formant d'où une valeur détectée excessivement élevée, le statut d'une part non-négligeable de mesures effectuées automatiquement est susceptible de rester incertain même après cette étape de filtrage. Ce constat a d'ailleurs récemment conduit Lancien et al. (2023) à opter pour une méthode statistique de filtrage fondée sur l'utilisation de distances de Mahalanobis.

Nous illustrons ici le cas de valeurs formantiques détectées automatiquement sur la voyelle /u/ pour laquelle la détection automatique de formants est problématique en raison de ses propriétés spectrales.

3 Méthodes

3.1 Données

Les données utilisées pour la mise en pratique sur des cas pratiques sont issues du corpus PTSVox (Chanclu et al., 2020), dans lequel 369 locuteurs francophones natifs ont été enregistrés dans des tâches de production de parole lue et de parole spontanée. Un sous-ensemble de 24 locuteurs (12 hommes, 12 femmes) a été enregistré dans les deux conditions lors de multiples sessions. Après une première étape de transcription manuelle et d'alignement forcé, la segmentation en mots et en phones a été corrigée manuellement pour ces 24 locuteurs. Pour les besoins de cet article, nous nous concentrons sur les voyelles produites par une locutrice âgée de 22 ans lors de l'enregistrement.

3.2 Extraction de mesures acoustiques

Les valeurs de fréquence fondamentale ainsi que les fréquences des 3 premiers formants ont été extraites au milieu de chaque voyelle, à l'aide d'un script Praat (Boersma & Weenink, 2024) développé par l'auteur. Pour l'extraction de la fréquence fondamentale, les paramètres par défaut ont été utilisés, avec un pas temporel défini automatiquement et une plage de valeurs fixée à 60-600Hz suivant les recommandations de De Looze (2010). L'extraction des fréquences formantiques a été effectuée avec l'algorithme de Burg implémenté dans Praat, avec également les paramètres par défaut (10 paramètres LPC, trames de 25ms avec recouvrement de 10ms) et une fréquence maximale pour la détection de 5 formants fixée à 5kHz pour les hommes et 5,5kHz pour les femmes.

4 iHist : histogrammes paramétrables et interactifs

4.1 Principales fonctionnalités de l'application

L'application iHist permet à l'utilisateur, après avoir importé un fichier de données au format Excel, TSV ou CSV, de sélectionner une variable quantitative à représenter sous forme d'histogramme, avec la possibilité optionnelle de filtrer les données prises en compte en fonction des valeurs d'autres variables quantitatives ou catégorielles. Parmi les principales fonctionnalités de l'application, l'utilisateur peut modifier les paramètres de l'histogramme, notamment le nombre de classes et les valeurs minimum et maximum affichées. L'application permet également d'afficher en surimpression la courbe correspondant à la densité de la distribution et/ou celle de la distribution normale de même moyenne et même écart-type. Afin de permettre à l'utilisateur de faire le lien visuellement entre la distribution dans son ensemble et les projections les plus couramment utilisées dans les représentations graphiques (graphe en barres avec erreur-type, ou boîte à moustaches), des droites verticales représentant la moyenne et l'intervalle de confiance et/ou la médiane et des quantiles sélectionnés peuvent également être ajoutés. L'application permet en outre une exploration interactive des données de l'histogramme, en accédant au détail des observations regroupées dans une classe via un clic sur la barre correspondante. De cette façon, l'utilisateur peut afficher ces informations, et les exporter dans un fichier structuré pour simplifier leur inspection ultérieure.

La version française de cette application est disponible en ligne à l'URL suivant : https://shiny.laboratoirephonetiquephonologie.fr/iHist_fr/ . Le code source de l'application ainsi que les fichiers de localisation en français et en anglais sont distribués sous licence GPL via GitHub[1].

4.2 Application aux valeurs de f0

Sur les mesures de f0 obtenues à partir des 4 972 voyelles produites par la locutrice sélectionnée, les valeurs limites obtenues pour la seconde passe de détection avec la méthode de De Looze (2010) sont de $0,83*Q_{15\%}$=131Hz et de $1,92*Q_{65\%}$=303Hz. Pour cette illustration (Figure 1) nous nous concentrons sur la limite basse, sous laquelle on trouve un nombre de valeurs plus important qu'attendu comme le montre l'histogramme. Si l'inspection de la première classe (70,4Hz-96,2Hz) confirme que les valeurs de f0 s'y trouvant correspondent bien à des erreurs de détection, les valeurs

[1] https://github.com/nicolasaudibert/iHist.git

proches de cette limite basse nécessitent une inspection plus approfondie afin d'affiner cette valeur limite, et le cas échéant définir des sous-groupes. Cette inspection est ici réalisée via l'exportation au format TSV des valeurs sélectionnées après affinage du nombre de classes, utilisée ensuite comme entrée d'un script Praat dédié à l'affichage et l'évaluation d'extraits sélectionnés.

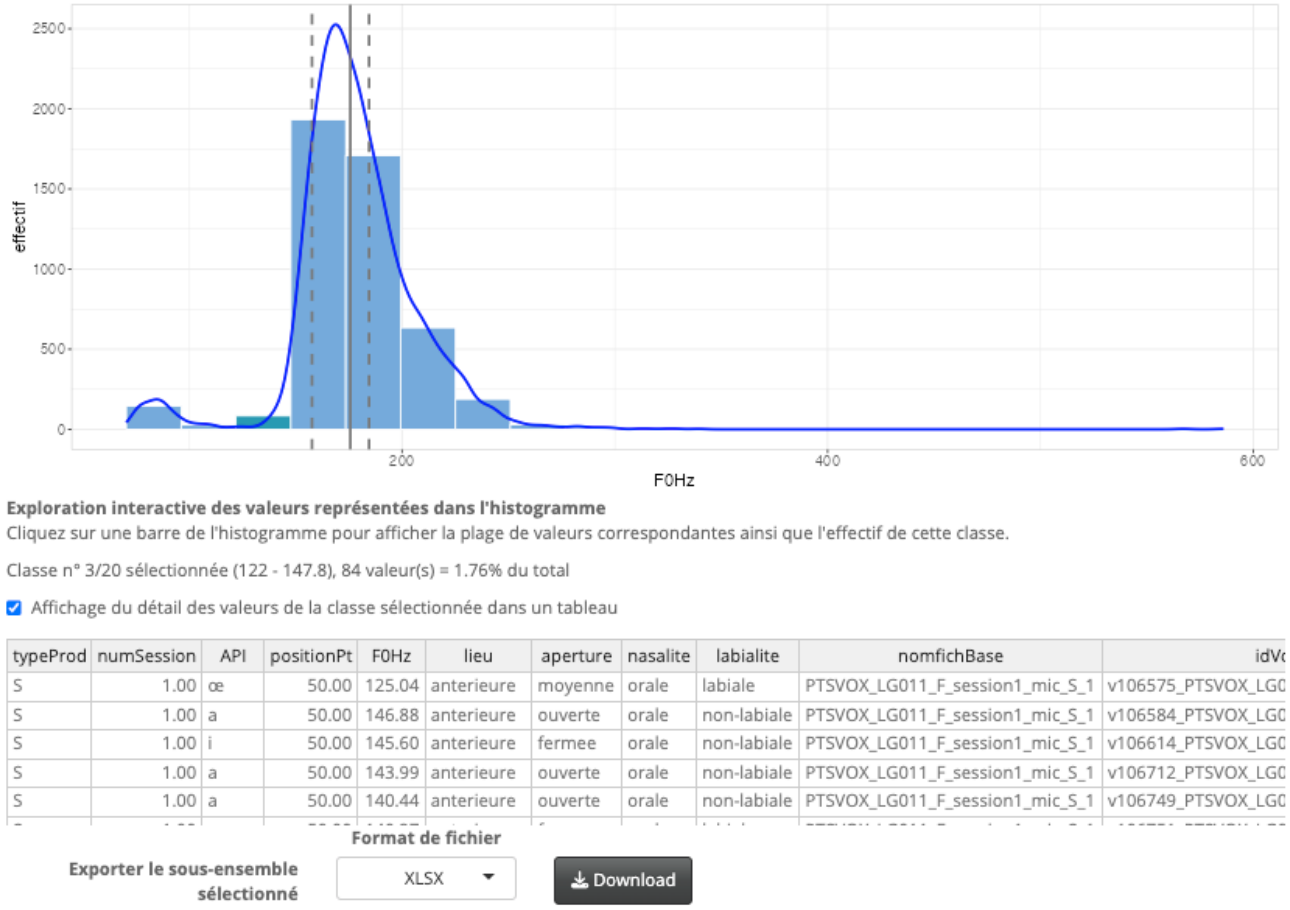


FIGURE 1 : Illustration de l'application *iHist* appliquée aux valeurs de f0 de la locutrice LG011, après sélection d'une classe (en vert) contenant 1,76% de l'ensemble des valeurs dont le détail est affiché dans le tableau en bas de page. Outre l'adaptation du nombre de classes affichées, l'affichage de la densité de distribution (courbe bleue) est activé, ainsi que celui de la médiane et de quantiles sélectionnés (droites grises, ici en pointillé quantiles à 15% et 65%).

5 iScatter : nuages de points interactifs

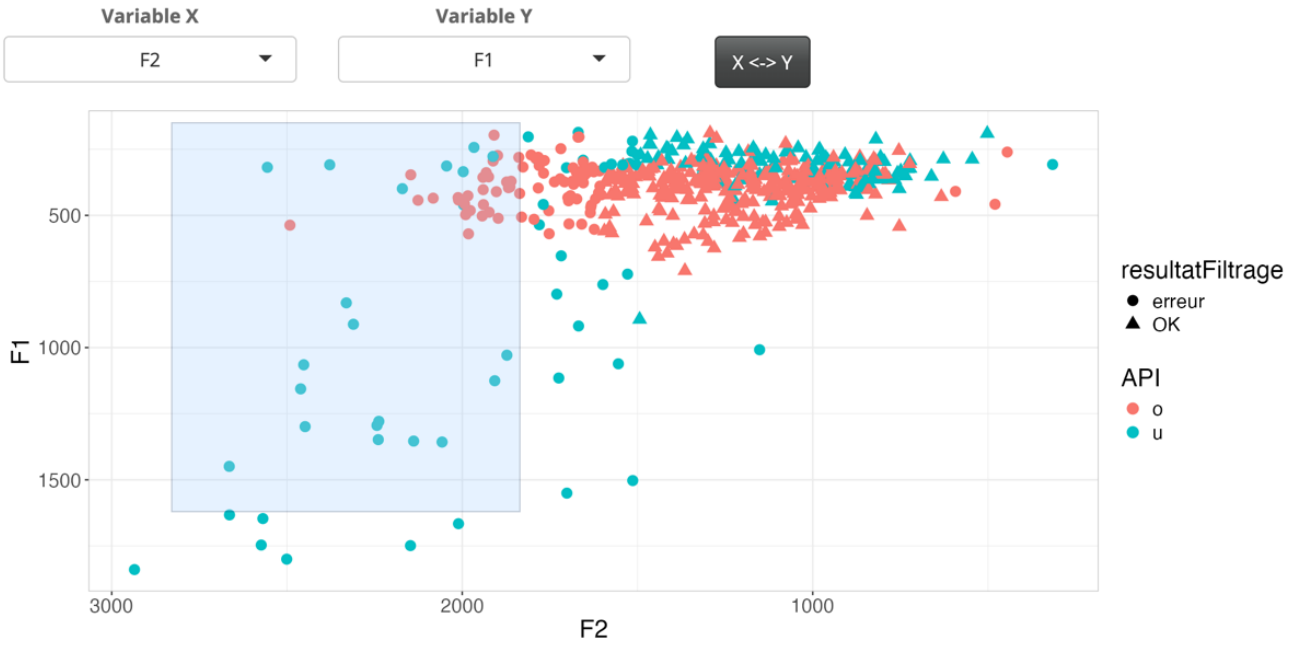
5.1 Principales fonctionnalités de l'application

Outre les fonctionnalités d'importation et de filtrage des données équivalentes à celles de iHist, l'application iScatter dédiée à l'affichage de nuages de points interactifs permet à l'utilisateur de sélectionner une ou deux variables indépendantes qui définissent la couleur et/ou la forme des points, et de façon optionnelle d'afficher les droites de régression pour chaque sous-groupe défini par les variables indépendantes. Les corrélations de Pearson et Spearman sont en outre affichées pour chaque sous-groupe. L'utilisateur a la possibilité de sélectionner le point le plus proche par un clic ou d'effectuer une sélection rectangulaire pour afficher les détails des points sélectionnés dans un

tableau exportable. La sélection d'une ligne dans ce tableau permet de visualiser la position du point correspondant dans le nuage de points.

La version française de l'application iScatter est disponible en ligne à l'URL suivant : https://shiny.laboratoirephonetiquephonologie.fr/iScatter_fr/. Le code source de l'application et les fichiers de localisation sont également distribués sous licence GPL via GitHub[2].

5.2 Application aux valeurs de formants : cas de la voyelle /u/



Affichage du détail des valeurs sélectionnées dans un tableau

loc	sexeLoc	typeProd	duree	API	ctxG_API	ctxD_API	F1	F2	F3	F4	resultatFiltrage
LG011	F	S	0.02	o	j	n	481.46	1977.24	3302.63	4323.47	erreur
LG011	F	S	0.02	u	t	ʒ	911.89	2310.52	3511.31	4428.10	erreur
LG011	F	S	0.04	u	t	ʃ	1348.10	2239.53	3840.36	4665.96	erreur
LG011	F	S	0.04	o	s	m	389.39	1864.96	2868.84	4674.40	erreur
LG011	F	S	0.06	o	n	k	392.51	1863.79	3104.94	4076.96	erreur

FIGURE 2 : Illustration de l'application *iScatter* appliquée aux valeurs de formants de la locutrice LG011, après application d'un filtre pour n'afficher que les occurrences de /u/ et /o/ et sélection rectangulaire d'une zone incluant une majorité de valeurs étiquetées comme erronées suite à l'application d'un crible. La première variable indépendante (couleurs distinctes) est utilisée pour distinguer les catégories vocaliques, et la seconde (formes des points) pour distinguer les exemplaires étiquetés par le crible comme ayant des valeurs formantiques correctes ou non

La figure 2 illustre l'utilisation de l'application iScatter pour explorer les valeurs de F1 et F2 considérées comme erronées, suite à l'application d'un crible avec les mêmes valeurs limites que celles utilisées par Audibert et al. (2015) sur les mesures formantiques des 4 255 voyelles orales produites par la locutrice LG011. La sélection effectuée pour afficher les détails des caractéristiques de certaines voyelles se concentre ici sur les occurrences de la voyelle /u/ pour lesquelles les valeurs de F2 détectées automatiquement dépassent la limite supérieure de 1500Hz définie par le crible

2 https://github.com/nicolasaudibert/iScatter.git

utilisé pour les /u/ produits par des femmes (par ailleurs pour de nombreux exemplaires de /u/, la valeur de F1 détectée dépasse également la limite supérieure de 1000Hz).

Notons ici que parmi les options de paramétrage de l'affichage du nuage de point proposées par l'application, l'inversion de l'orientation des axes *x* et *y* a été utilisée afin d'obtenir une représentation conforme à celle couramment utilisée pour représenter les voyelles dans le plan F1/F2. Par ailleurs afin d'éviter d'afficher un nombre de points superposés trop important, un filtre a été appliqué afin de n'afficher que les occurrences de /u/, ainsi qu'à titre de comparaison les occurrences de la voyelle /o/ avec laquelle le recouvrement est particulièrement important dans les données de cette locutrice. De même qu'avec l'application iHist, en complément de l'inspection immédiate des caractéristiques documentées dans le jeu de données importé, le sous-ensemble sélectionné peut-être exporté pour faire l'objet d'une inspection guidée des signaux de parole (en l'occurrence également à l'aide d'un script Praat).

6 Discussion et conclusion

Les cas d'utilisation exposés dans cet article se concentrent sur l'utilisation des outils proposés pour identifier des erreurs de mesure. Cependant ces cas d'utilisation constituent une simple illustration, ces outils pouvant être utiles pour de nombreuses autres applications potentielles liées à l'étude de la voix et de la parole. En particulier, l'application *iScatter* dédiée aux nuages de points interactifs utilisable pour identifier les observations qui ne suivent pas la tendance générale dans le cadre de l'exploration des liens entre variables quantitatives, plus spécifiquement les relations entre mesures acoustiques et/ou perceptives supposées capturer différentes facettes d'un même phénomène. À ce titre, cette application est déjà utilisée régulièrement par des collègues pour l'inspection de données de voix et de parole, mais n'avait jamais fait l'objet d'une diffusion plus large dans la communauté.

Sans avoir la prétention de fédérer une communauté autour du développement de ces outils dont de nombreux aspects restent perfectibles, la distribution de leur code source sous licence GPL permet à tout un chacun de les modifier pour les adapter à des besoins spécifiques, et surtout de les utiliser localement avec des fichiers plus volumineux que ne le permet le serveur utilisé pour rendre ces applications directement accessibles en ligne. En effet, leur utilisation sur un ordinateur personnel nécessite uniquement l'installation des logiciels R et RStudio (complétés par un ensemble de paquets) auxquels la majorité des collègues et étudiants amenés à procéder à l'analyse quantitative de données de voix et de parole sont déjà familiarisés.

La vocation des outils proposés est avant tout d'offrir aux collègues et étudiants un complément à l'utilisation de tableurs pour l'exploration des données quantitatives, afin de faciliter la mise en œuvre de l'injonction fréquente dans les formations en statistiques quel que soit le niveau visé : « N'oubliez pas de regarder vos données ! ». En revanche il ne s'agit en aucun cas de se substituer aux formations à l'utilisation de R et autres outils avancés d'analyse de données, qui restent indispensables et doivent continuer à être encouragées.

Remerciements

Ce travail a été soutenu par le projet ANR PASDCODE (ANR-21-CE28-0015) et par le Laboratoire d'Excellence Empirical Foundations of Linguistics (LabEx EFL, ANR-10-LABX-0083). Il contribue à l'IdEx Université de Paris (ANR-18-IDEX-0001).

Références

AUDIBERT, N., FOUGERON, C., GENDROT, C., & ADDA-DECKER, M. (2015). Duration-vs. style-dependent vowel variation: A multiparametric investigation. In *Proceedings of the 18th International Congress of Phonetic Sciences*, p. 5. HAL: hal-01251372.

CHANG W., CHENG J., ALLAIRE J., SIEVERT C., SCHLOERKE B., XIE Y., ALLEN J., MCPHERSON J., DIPERT A. & BORGES B. (2024). *shiny: Web Application Framework for R*. R package version 1.8.0.9000, https://github.com/rstudio/shiny, https://shiny.posit.co/.

BOERSMA P. (1993). 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* (Vol. 17, No. 1193, pp. 97-110).

BOERSMA P. & WEENINK D. (2024). Praat: doing phonetics by computer [Programme informatique]. Version 6.4.05, téléchargé le 27 janvier 2024 depuis http://www.praat.org/

CHANCLU A., GEORGETON L., FREDOUILLE C. & BONASTRE J. F. (2020). PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire. In *Actes de la 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* (pp. 73-81). HAL: hal-02798519.

DE LOOZE C. (2010). Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais. Thèse de doctorat. Université de Provence - Aix-Marseille I. HAL: tel-00470641.

FALK S. & AUDIBERT N. (2021). Acoustic signatures of communicative dimensions in codified mother-infant interactions. *The Journal of the Acoustical Society of America*, *150*(6), 4429-4437. DOI: 10.1121/10.0008977. HAL: hal-03592269.

FERRAGNE E. & PELLEGRINO F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*, *38*(4), 526-539. DOI: 10.1016/j.wocn.2010.07.002. HAL: hal-01240095.

GENDROT C. & ADDA-DECKER, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech 2005*, p. 2453-2456. DOI: 10.21437/Interspeech.2005-753. HAL: halshs-00188096.

HERMES A., AUDIBERT N. & BOURBON A. (2023). Age-related vowel variation in French. In *Proceedings of the 20th International Congress of Phonetic Sciences*, p. 2045-2049. Guarant International. HAL: hal-04193397.

JOUVET D. & LAPRIE Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, p. 1614-1618. IEEE. DOI: 10.23919/EUSIPCO.2017.8081482. HAL: hal-01585554.

KENT R. D. & VORPERIAN H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders*, *74*, 74-97. DOI: 10.1016/j.jcomdis.2018.05.004.

LANCIEN M., ADDA-DECKER M. & STUART-SMITH J. (2023). Knowledge-driven vs. data-driven methods for filtering acoustic measures in phonetics corpora. In: Skarnitzl R. & Volín J. (Eds.), In *Proceedings of the 20th International Congress of Phonetic Sciences*, p. 3166-3170. Guarant International.

LIBERMAN M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, *5*, 91-107. DOI: 10.1146/annurev-linguistics-011516-033830.

NICKLIN C. & PLONSKI L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 26-55. DOI: 10.1017/S0267190520000057.

OSBORNE J. W. & OVERBAY A. (2019). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6. DOI: 10.7275/qf69-7k43.

R CORE TEAM (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

VAN DER HARST S., VAN DE VELDE H. & VAN HOUT R. (2014). Variation in Standard Dutch vowels: The impact of formant measurement methods on identifying the speaker's regional origin. *Language Variation and Change*, *26*(2), 247-272. DOI: 10.1017/S0954394514000040.

VAYSSE R., ASTÉSANO C. & FARINAS J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *The Journal of the Acoustical Society of America*, *152*(5), 3091-3101. DOI: 10.1121/10.0015143. HAL: hal-03879676.

WICKHAM H., AVERICK M., BRYAN J., CHANG W., MCGOWAN L.D., FRANÇOIS R., GROLEMUND G., HAYES A., HENRY L., HESTER J., KUHN M., PEDERSEN TL., MILLER E., BACHE S.M., MÜLLER K., OOMS J., ROBINSON D., SEIDEL D.P., SPINU V., TAKAHASHI K., VAUGHAN D., WILKE C., WOO K. & YUTANI H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686. DOI: 10.21105/joss.01686.