Traitement automatique et analyse de la variation dans la parole : des mesures phonétiques sur grands corpus aux réseaux de neurones profonds

Cédric Gendrot

TRAITEMENT AUTOMATIQUE ET ANALYSE DE LA VARIATION DANS LA PAROLE : DES MESURES PHONETIQUES SUR GRANDS CORPUS AUX RESEAUX DE NEURONES PROFONDS

Dossier présenté pour l'obtention d'une

Habilitation à Diriger les Recherches

Par

Cédric Gendrot

Maître de Conférences à l'Université Sorbonne Nouvelle – Paris 3

Laboratoire de Phonétique et Phonologie – UMR 7018

Soutenue le 8 juillet 2021 devant :

Laurent Besacier, Président, Professeur, Université Grenoble Alpes et Naver Labs Europe

Ann R. Bradlow, Rapporteur, Professeur, Northwestern University, Department of Linguistics

Corinne Fredouille, Rapporteur, Maitre de Conférences HDR, Avignon Université, UR 4128 Laboratoire Informatique d'Avignon

Kim Gerdes, Examinateur, Professeur, Université Paris-Saclay, UMR9015 Laboratoire Interdisciplinaire des Sciences du Numérique

Christine Meunier, Rapporteur, Directeur de Recherche CNRS, Aix-Marseille Université, UMR7309 Laboratoire Parole et Langage

François Pellegrino, Garant, Directeur de Recherche CNRS, Université Lumière Lyon 2, UMR5596 Dynamique du Langage

Remerciements

Ce document de synthèse est le résumé de ma carrière d'enseignant-chercheur, un métier complexe et passionnant, fait de rencontres parfois décisives. Je vais essayer de les relater ici :

Je remercie avant toute chose mon garant d'habilitation, François Pellegrino qui a su me guider et me préparant en me laissant toute liberté.

Je remercie les membres du jury, Laurent Besacier, Ann Bradlow, Corinne Fredouille, Kim Gerdes, et Christine Meunier qui ont accepté spontanément de relire et d'évaluer ce document de synthèse.

Je remercie Jacqueline Vaissière, mon ancienne directrice de thèse, qui m'a formé et tout appris.

Je remercie l'ensemble des membres du Laboratoire de Phonétique et Phonologie, un environnement humain et stimulant est un tel atout pour un enseignant-chercheur, et plus particulièrement Angélique, Barbara, Cécile, Didier, Martine, Nicolas, Rachid.

Je remercie mes collègues de TAL dans mon département à l'Université Sorbonne Nouvelle qui m'ont permis de m'épanouir dans le Master PluriTAL que j'ai vu grandir, avec une mention spéciale à Serge, Kim et Pascal.

J'ai eu de nombreuses responsabilités académiques depuis mon recrutement. Outre les nombreuses heures passées *à faire du mail*, ces activités m'ont permis également de rencontrer de nombreuses personnes avec qui il a souvent été agréable d'échanger et d'apprendre.

Je remercie Emmanuel, mon ami de 15 ans, qui a soutenu son habilitation quelques jours avant moi. Que de chemin parcouru …

Je remercie Olivier, mon ami de 25 ans qui m'a jeté dans la recherche il y a si longtemps, il est de ces rencontres qui changent une vie.

Table des matières

1 Introduction générale

J'ai été recruté en 2006 à l'Université Sorbonne Nouvelle sur un poste de Maître de Conférences dont l'intitulé était « phonétique et traitement automatique des langues ». Ce poste est donc à l'intersection de deux disciplines : la linguistique et l'informatique. Ce parallèle entre les deux disciplines, dans mes recherches bien sûr, mais également dans mes responsabilités pédagogiques, ainsi qu'administratives sera le fil rouge de ce document de synthèse, où ces trois volets sont intimement liés.

Mes responsabilités d'ordre pédagogique et académique se sont accrues progressivement au fil de ma carrière. Après ma titularisation obtenue en 2007, j'ai intégré les différents conseils de mon UFR qui m'ont permis de mieux comprendre le fonctionnement de mon université. Puis, j'ai rapidement pu accéder à des charges avec plus d'implications (responsabilité du Master, responsabilité d'axe du LabEx EFL 'Empirical Foundations of Linguistics', mandat d'élu aux commissions centrales, direction d'UMR). Ces activités sont non négligeables pour un enseignant-chercheur puisqu'elles représentent souvent plus de la moitié du temps de travail. Tous ces aspects de ma carrière sont développés dans les deux premiers volets de ce document. J'essaye d'y expliquer notamment comment j'ai essayé de les combiner à mes activités de recherche en termes de contenu.

Dès mon recrutement, j'ai intégré le Laboratoire de Phonétique et Phonologie, un laboratoire mixte avec le CNRS (UMR7018) très dynamique, où j'ai pu participer à des projets de recherche financés (ANR, IDEX, PI, etc.) en rédigeant certains modules, puis en dirigeant moi-même des projets que je présenterai dans le troisième et dernier volet. Ce document d'Habilitation à Diriger les Recherches a été pour moi l'occasion de revenir sur 15 années de recherches après l'obtention de mon doctorat. Considérer l'ensemble de mes travaux comme une suite cohérente de questionnements scientifiques fut une introspection agréable et enrichissante. Ces recherches ont eu pour fil rouge l'utilisation de grands corpus de parole non préparée et segmentés automatiquement. Le but premier était d'avoir accès à une grande quantité de données, mais ce point seul ne pouvait être un objectif. Mes travaux se sont concentrés sur la variation des segments et principalement les voyelles dans une parole non préparée. Quels sont les critères linguistiques qui permettent de prédire la réalisation des segments alors que la parole est caractérisée par une forte variabilité ?

Dans la première section, je présente les premiers travaux que j'ai effectués sur grands corpus, incluant les premières mesures quantitatives et observations, ainsi que les précautions méthodologiques nécessaires. Le premier corpus avec lequel j'ai travaillé était le corpus ESTER[1] : un corpus de parole journalistique, et après des premières observations très générales (réduction en fonction de la durée phonétique), j'ai souhaité valider et approfondir ces résultats sur d'autres langues et d'autres styles de parole.

Dans la deuxième section, j'ai développé cette analyse de la variation des voyelles en m'attachant notamment aux catégories prosodiques que l'on peut identifier dans la parole. Une nouvelle fois, j'ai appliqué ces analyses à d'autres langues et à d'autre styles.

Dans la troisième section, je me suis appliqué à traiter des phénomènes linguistiques dont la variation soulève des questions sur la séparation entre phonétique et phonologie, et j'ai montré comment les grands corpus pouvaient de différentes façons apporter des réponses concrètes.

[1] *http://catalogue.elra.info/en-us/repository/browse/ELRA-S0241/*

Dans la quatrième et dernière section, une discussion est proposée, notamment au regard des évolutions théoriques et technologiques en recherche phonétique. L'utilisation des grands corpus y est comparée à celle des petits corpus de parole lue. Une remise en question des méthodes, tant pour les données que pour les analyses est également avancée et des solutions sont proposées.

2 Activités d'enseignement et responsabilités pédagogiques

Dans mes activités d'enseignant, j'ai eu la chance de pouvoir rapidement constituer des cours nouveaux au sein de la maquette du département de linguistique. Très rapidement, j'ai pu avoir la responsabilité d'une UE de Phonétique (CM+TD), d'une UE de statistiques, ainsi que d'un cours de Linguistique de Corpus en Licence.

Dans le cadre du Master pluriTAL (http://plurital.org), Master cohabilité avec l'université Paris Nanterre et l'INALCO, j'ai eu dès 2007 l'opportunité de dispenser un cours combinant la phonétique et l'informatique, où j'ai enseigné le traitement automatisé de données orales, incluant son analyse statistique, puis les bases du traitement du signal acoustique. J'ai décidé plus récemment de réorienter ce cours pour lui donner une visée plus appliquée en invitant chaque étudiant à créer un système de synthèse de la parole à partir des enregistrements de sa propre voix, ce qui l'oblige à programmer plusieurs scripts qui visent à réaliser du traitement du signal automatique ainsi qu'un travail d'analyse et d'annotation. Depuis 2019 pour la maquette actuelle, j'ai initié la création d'un cours innovant et très demandé, d'abord financé par l'INALCO, puis rapatrié au sein de mon université. Ce cours très lié aux préoccupations du moment permet aux étudiants de réaliser un système de reconnaissance de l'oral (locuteurs, formes verbales, phonèmes, parole, etc.) à partir de Réseaux de Neurones.

De par ces responsabilités, j'ai eu l'opportunité de diriger et co-diriger un nombre important de mémoires de Master (cf. Annexes), sur des sujets proches de mes thématiques de recherche, mais également sur des sujets que les étudiants avaient choisis eux-mêmes.

2.1 Activités pédagogiques

J'ai eu l'opportunité de dispenser dès le début de ma carrière des cours en Licence et en Master, de par les besoins au sein de notre département pour des cours transversaux qui allient informatique et linguistique. Tout en continuant à dispenser des cours fondamentaux de linguistique, phonétique et phonologie en Licence, j'ai enseigné au sein du Master PluriTAL et du Master de Phonétique et Phonologie. Les cours que j'ai dispensés ont régulièrement été amenés à évoluer vers une plus grande interaction où les étudiants sont invités à travailler sur des données qu'ils ont eux-mêmes créées. Mon intégration au sein de différentes spécialités m'a obligé à régulièrement dépasser mon service d'enseignement de façon significative. Depuis 2014, la souplesse accordée pour les conditions d'enseignement aux doctorants m'a permis de redistribuer un certain nombre de cours de licence aux doctorants intégrés dans mon équipe de recherche.

2.1.1 Cours de Licence spécifiques

Les cours de Licence en phonétique et phonologie s'intègrent dans la maquette de Licence de Sciences du Langage, une maquette spécialisée en Linguistique qui s'adresse à de futurs professeurs des écoles, mais également à des étudiants désireux de passer un concours de l'enseignement secondaire, ou des étudiants qui passent en parallèle les concours d'orthophonie. On y trouve également une proportion non négligeable d'étudiants qui souhaitent poursuivre en Master TAL. Les cours de phonétique et phonologie s'inscrivent dans une logique progressive qui nous permet de pousser les exigences scientifiques jusqu'à un très bon niveau dans chaque discipline.

2.1.2 Cours de Licence transversaux

Ces cours spécifiques sont complétés par des cours transversaux que l'équipe pédagogique a progressivement élaborés. J'ai ainsi pu avoir la responsabilité d'un cours de Linguistique de Corpus. Ce dernier est un cours à vocation pluridisciplinaire qui permet d'intégrer simultanément des notions de syntaxe, sémantique, acquisition et bien sûr phonétique sur des corpus oraux que les étudiants enregistrent eux-mêmes afin de les analyser de façon semi-automatique. L'utilisation d'une analyse semi-automatique est cruciale ici puisqu'elle oblige les étudiants à comprendre ce qu'ils font, tout en les formant à des rudiments de programmation informatique.

Dès 2006 j'ai également pu mettre en place des cours de statistiques pour les étudiants de Licence devant la nécessité croissante de maitriser ces outils pour toute recherche scientifique. Après avoir dispensé ces cours plusieurs années et après avoir incité à la création de cours semblables en Master, j'ai pu passer la main en laissant ce cours à d'autres membres de mon laboratoire.

J'ai également créé au sein de la maquette de licence 2014 le cours pour le Bureau des Enseignements Transversaux (B.E.T.) intitulé Voix, Rythme et communication. Ce cours qui est toujours dispensé dans la maquette en vigueur est né de ma motivation toujours intacte de rapprocher des disciplines qui se côtoient trop souvent sans bénéficier l'une de l'autre. Je fais découvrir chaque année la phonétique à des étudiants de langues (notons qu'ils ont déjà des cours de phonétique pour l'apprentissage d'une langue seconde), mais surtout à des étudiants de l'UFR Arts et Media pour que ceux-ci puissent avoir une facette plus technique de leur voix et de la façon dont ils l'utilisent de façon inconsciente dans la communication quotidienne.

2.1.3 Cours de Master et articulation avec la licence ou la Recherche

Le Master de Phonétique et Phonologie est un Master Recherche totalement imbriqué dans le laboratoire CNRS de Phonétique et Phonologie. Il propose dans ce cadre des cours approfondis qui permettent de former des étudiants à la recherche, mais plus récemment à accéder aux métiers de « data-scientists ». Le Laboratoire est un élément international moteur dans l'analyse acoustique de la parole, ainsi que sa modélisation spectrale, et les cours visent à former les étudiants dans cette optique. Nous sommes également équipés d'une plateforme physiologique complète permettant l'analyse de la parole (électro-glotto-graphe, électromyographe, capteur de mouvements Qualysis, et plus récemment electro-magnetic-articulograph, etc.) et nos cours visent à former les étudiants à utiliser ces outils dans leurs recherches. Il est à noter que de nombreux chercheurs nationaux et internationaux viennent également se former à ces instruments.

2.1.4 Cours à distance (ENEAD)

Depuis l'année universitaire 2019-2020, le département de Linguistique a intégré un volet à distance (ENEAD) pour la Licence et le Master, et je dispense un cours de Master à distance en phonétique au sein de la maquette de Linguistique LLTS (Langage, Langues, Textes, Société). Les enjeux sont importants puisqu'un nombre croissant d'étudiants souhaitent continuer à se former alors qu'ils doivent commencer une activité professionnelle, ou bien parce qu'ils n'ont pas la possibilité de se déplacer en île de France. J'ai pu me former très rapidement à l'utilisation d'outils de forums et de tableaux interactifs pour mieux répondre aux attentes spécifiques de ces étudiants. La pandémie qui est apparue au début de l'année 2020 nous a tous contraints à utiliser ces outils. L'enseignement de l'utilisation d'outils informatiques n'est pas la plus aisée et demande certains efforts de la part des enseignants et des étudiants, qui les ont acceptés pour une grande majorité. Comme pour le travail dans une équipe de recherche, je ne crois pas que le travail à distance soit efficace sur le long terme cependant.

2.2 Présentation des formations suivies concernant mes activités pédagogiques

Les formations que j'ai suivies m'ont permis de réactualiser le contenu de mes enseignements afin que les étudiants puissent bénéficier des avancées en termes de contenus et de méthodes. Bien sûr, toutes les formations ne peuvent être transmises sans un certain recul qui permet de réfléchir à la meilleure façon de les mettre en forme pour un public étudiant. Les conférences et workshops auxquels je participe chaque année peuvent également être considérés comme des formations dans la mesure où ils permettent de comprendre quels sont les courants théoriques en vigueur ainsi que les méthodes d'analyses utilisées. Si des formations proposées par l'université, le LabEx EFL ou le CNRS font partie de mon quotidien, elles ne font qu'initier et ne peuvent en aucun cas remplacer la pratique pour ses propres recherches qui permet d'intégrer pleinement l'acquisition de nouvelles connaissances.

Dans le domaine de l'informatique, les technologies évoluent très rapidement :

- en 2006 pour mon recrutement le langage le plus utilisé en TAL était « perl », alors que « python » est devenu actuellement la référence. Il est devenu indispensable aux étudiants pour leur formation et pour leur recherche d'emplois au sortir de l'université, il était donc nécessaire que je me forme auprès de collègues à l'intérieur et à l'extérieur de mon université. Je l'utilise désormais dans mes cours de Master 1 et 2, citons notamment la librairie « Parselmouth » qui permet d'intégrer des commandes Praat en Python.

- Au début des années 2000, les statistiques étaient réalisées à l'aide de logiciels payants tels que Statistica ou SPSS, mais R, un logiciel gratuit a progressivement pris une place fondamentale dans ce domaine avec des mises à jour quotidiennes des librairies qui implémentent tous types de calculs. J'ai rapidement suivi des formations au sein de mon laboratoire pour me familiariser avec ce nouveau langage de programmation intégré qui peut s'enseigner à des étudiants, y compris en Sciences Humaines et Sociales.

- Les réseaux de Neurones en intelligence artificielle ont explosé en 2014 lors de challenges de reconnaissance d'images. Dans le cadre du Master, mais ceci peut s'appliquer pour des étudiants de licence, il est nécessaire que les enseignants-chercheurs se forment quotidiennement dans leurs recherches afin de pouvoir initier les étudiants à ces nouvelles technologies et les inciter à s'y impliquer

pleinement. J'ai pu mettre en application ce principe avec la création d'un nouveau cours mentionné au début de cette section.

Dans le domaine de la phonétique et de la phonologie, les avancées technologiques ont permis d'analyser la parole plus facilement et de façon plus visuelle. Il va de soi qu'au sein d'un laboratoire de recherche avec des doctorants qui suivent les évolutions et les mettent en pratique immédiatement, il est plus facile de pouvoir se former de façon continue.

2.3 Responsabilités pédagogiques

Après avoir intégré le conseil d'UFR de linguistique (à l'époque le regroupement des UFRs n'avait pas eu lieu) en 2008, j'ai été chargé de la responsabilité du parcours « professorat des écoles » en collaboration avec l'Université Paris 6. Ce parcours sélectif, toujours en vigueur, qui comptait à l'époque environ 35 étudiants pendant les trois années de Licence nous permet d'intégrer des bons étudiants motivés, mais avec des exigences d'emploi du temps complexes car ils suivent une partie de leur formation à l'Université UPMC (Paris 6, et désormais Sorbonne Université). Cette expérience m'a permis de découvrir les responsabilités pédagogiques avec la rédaction de maquettes, les conseils de perfectionnement, et les interactions avec les différents services de l'université.

En 2010, j'ai pris la responsabilité du Master de Phonétique et Phonologie, et dans ce cadre j'ai constitué deux maquettes pédagogiques. Pour la première, j'ai créé une co-accréditation avec le Master de Phonétique de l'université Paris 7. Cette première maquette ambitieuse m'a permis de créer une formation cohérente en multipliant les forces de deux universités parisiennes. Cette co-accréditation s'inscrivait à l'époque dans le cadre de la COMUE Sorbonne Paris Cité et a été citée à de nombreuses reprises comme un modèle pédagogique au sein de la COMUE de par sa capacité à mutualiser les forces et les étudiants dans un esprit parfaitement collaboratif. J'ai également travaillé dès cette première maquette avec les forces enseignantes du Master PluriTAL afin de proposer à leurs étudiants un nombre important de cours de Phonétique en M2 pour ceux qui souhaitaient s'orienter vers l'informatique sur des données orales (le TAL étant souvent concentré sur les ressource écrites). Pour la deuxième maquette, j'ai pu mettre en place pour les étudiants du Master de Phonétique et Phonologie des modules de spécialisation où nos étudiants sont encouragés à se perfectionner en développement informatique au sein du Master PluriTAL, ce qui permet aux étudiants de ces deux formations de se perfectionner en sélectionnant les spécialisations de leur choix dans l'autre formation. Cet aspect avait pour objectif d'être une ébauche de Master professionnalisant puisque tous les étudiants ne se destinent pas nécessairement à faire de la recherche après un Master, et il apparaissait judicieux de les préparer à intégrer rapidement le milieu professionnel. En effet, les assistants vocaux commençant à se multiplier à l'époque, il y avait déjà une forte demande pour des compétences alliant linguistique et informatique, et plus spécifiquement sur l'oral.

Depuis 2010, je suis également responsable E-Candidat de la licence sciences du langage. J'examine et je sélectionne tous les dossiers d'inscription des étudiants en réorientation vers notre licence (L1 à L3), et depuis l'instauration de Parcours-Sup des étudiants de L2 et L3 seulement.

2.4 Corpus et diffusion de la recherche

Les corpus enregistrés par les étudiants pendant les cours de Linguistique de Corpus sont utilisés avec leur accord, dans le cadre de la protection des informateurs par la CNIL (Commission Nationale de l'Informatique et des Libertés) et plus récemment le RGPD (Règlement Général sur la Protection de Données). Ces corpus sont utilisés pour compléter et constituer une base de données d'annotations en dépendances sur le français spontané (Arborator / tree-tagger : avec Kim Gerdes et Sylvain Kahane) afin de faciliter l'analyse syntaxique automatique.

Si j'ai eu l'opportunité d'encadrer un grand nombre d'étudiants en Master 1 et 2 (voir détail en annexes) de par ma présence au sein du Master PluriTAL, ainsi qu'au sein du Master de Phonétique et Phonologie, je n'ai malheureusement pas encore pu suivre un nombre significatif de doctorants de manière officielle, cette tâche incombant le plus souvent aux chercheurs CNRS de mon équipe ainsi qu'au professeur de Phonétique en place. Depuis les réglementations mises en place et appliquées par l'Ecole Doctorale 268, les Maîtres de Conférences de mon équipe se voient de plus en plus offrir l'opportunité de co-diriger des doctorants, ce qui me permet de profiter de cette opportunité avec plaisir (avec notamment la co-direction de Gabriele Chignoli depuis 2018). J'avoue avoir hâte de pouvoir valider mon Habilitation à Diriger des Recherches pour multiplier ces expériences.

3 Responsabilités académiques

Après ma titularisation obtenue en 2007, j'ai été élu dès 2008 au conseil de l'UFR (devenue département depuis), et j'ai intégré la commission des postes et le collège de spécialistes, commissions que je n'ai pas quittées à ce jour. Ces commissions sont le squelette de tout département/UFR et il est nécessaire, sinon indispensable d'y participer. Je ne m'étendrai donc pas sur cet aspect en préférant me concentrer sur les démarches supplémentaires que j'ai pu effectuer dans mon parcours administratif. On pourra relever, au-delà des nombreuses participations à des comités de sélection, la présidence de deux comités de sélection pour des postes de Maitres de Conférences à pourvoir au sein de mon département, deux recrutements (Nicolas Audibert et Naomi Yamaguchi) qui ont donné entière satisfaction à tous les collègues de l'université.

3.1 Responsabilités administratives

Dès la création du LabEx EFL en 2010, j'ai été responsable de plusieurs opérations au sein de l'axe 1 (Phonetic and Phonological Complexity) mais surtout co-responsable de l'axe 6 qui traitait spécifiquement des corpus en linguistique. A partir de 2015, j'ai été élu responsable principal de cet axe (http://www.labex-efl.com/wordpress/recherche/ressources-langagieres/), dans lequel j'ai géré une quinzaine d'opérations traitant de l'oral et de l'écrit pour tout le périmètre du LabEx, j'ai dans ce cadre milité et obtenu l'intégration de membres individuels qui n'appartenaient pas à des UMRs et qui de par ce statut n'auraient pas pu bénéficier de ce cadre de recherche. Cette responsabilité m'a amené à participer mensuellement aux Conseils Scientifiques Restreints du LabEx EFL.

A partir de 2016, suite au départ en retraite de mon directeur d'unité Pierre Hallé (DR2 au CNRS), j'ai été élu à la direction du laboratoire de Phonétique et Phonologie (UMR 7018 CNRS : http://lpp.in2p3.fr). Le Mandat de Pierre Hallé n'ayant pu aller à son terme, j'ai endossé la responsabilité de rédiger le rapport HCERES, en collaboration avec le reste du laboratoire (évaluation en mars 2018 consultable ici : https://www.hceres.fr/fr/rechercher-une-publication/lpp-laboratoire-

de-phonetique-et-phonologie). A ce poste, et avec l'aide de la directrice adjointe Cécile Fougeron, nous avons essayé de poursuivre la dynamique de l'équipe avec notre touche personnelle, et en impliquant toujours plus les chercheurs CNRS dans les activités pédagogiques et académiques de l'université, en intégrant les étudiants le plus tôt possible (dès le Master 1, voire la Licence 3) dans les activités de recherche du laboratoire. Le laboratoire a été félicité dans ce sens par l'évaluation de l'HCERES en mars 2018. Il eût été plus logique qu'un professeur ou tout au moins un maître de conférences déjà titulaire d'une HDR ait cette responsabilité, mais les personnes susceptibles d'être à ce poste avaient depuis peu accepté des responsabilités académiques importantes et non cumulables avec le poste de directeur d'équipe. Le choix d'avoir à la tête de notre UMR un enseignant-chercheur, et une chercheuse CNRS comme directrice-adjointe s'expliquait par la stratégie mise en place depuis la création du laboratoire d'imbriquer parfaitement la recherche et l'enseignement dans une même unité.

3.2 Responsabilités et mandats locaux ou centraux

En 2014, après l'obtention d'une ANR Jeunes Chercheurs, et ayant été en contact avec le Vice-Président Recherche de l'Université M. Carle Bonafous-Murat, celui-ci m'a contacté pour faire partie de son équipe au sein de la Commission de la Recherche (et donc du Conseil Académique) pour son briguer la présidence de l'Université.

J'ai accepté cette responsabilité avec de grands projets en tête : je souhaitais que les doctorants puissent proposer des projets financés (« projets innovants ») sans avoir recours à un membre permanent comme référent ; j'envisageais la création d'un comité d'éthique qui puisse attribuer aux recherches expérimentales un agrément sans avoir à passer par les démarches complexes d'un Comité de Protection des Personnes ; et surtout j'envisageais de réfléchir à des démarches pour faciliter les recherches pluridisciplinaires entre des disciplines comme la linguistique et la communication ou le théâtre. Malheureusement, si le premier point a été facilement obtenu, le deuxième a dû attendre le début de la vague suivante pour voir le jour. Les deux autres ont été rapidement noyés dans les dissensions qui existaient au sein de cette commission, et il est vrai à cause de décisions tantôt incomprises, tantôt peu claires de la part de l'équipe présidentielle. Tout n'est pas négatif dans cette expérience puisque j'ai participé à un certain nombre d'activités qui m'ont permis de comprendre les rouages de l'université. Si je n'ai pas poursuivi mon investissement au sein de la Commission de la Recherche en 2019, je n'ai pas abandonné l'idée d'un comité d'éthique puisque c'est en tant que directeur d'équipe que j'ai demandé la mise en place de cette instance au sein de la commission de la recherche, en essayant de stimuler son activité pour qu'elle se pérennise (comité validé en fin d'année 2019 et mis en place au cours de l'année 2020).

Au cours de cette expérience au sein de ces conseils centraux, j'ai été membre de la commission de reclassement de l'université, et je fais également partie depuis 2016 de la Commission des Locaux en tant que représentant de mon département.

4 Activités scientifiques

L'objectif de ma thèse (soutenue en décembre 2005) était de valider le lien existant entre la structure intonative et rythmique d'un énoncé et la réalisation physiologique et acoustique des sons. J'ai

continué à travailler dans cette voie pendant dix ans, en essayant de l'appliquer à des domaines variés tels que l'apprentissage des langues, la syntaxe, etc. La nouveauté de ces travaux a consisté à utiliser des corpus de grande taille (plusieurs dizaines d'heures de parole) et d'élaborer les possibilités d'automatisation de ces analyses. En effet, il était courant jusque-là en phonétique (et en linguistique de façon plus générale) de constituer un corpus ad-hoc qui permettait de confirmer ou d'infirmer une question théorique. Bien que ces corpus prouvent encore leur utilité, il est utile d'avoir recours à des données « écologiques », où des individus produisent de la parole spontanée et naturelle dans des conditions libres de toute contrainte. L'inconvénient majeur de ce type de données est qu'il est difficile d'isoler les contextes d'analyse de façon automatique : il est donc nécessaire de repenser la stratégie de détection des unités telles que la fin de phrase, l'accent de phrase, etc. Je reviendrai sur cet aspect tout au long de mon manuscrit, ainsi que dans la discussion générale.

J'ai commencé à travailler avec des chercheurs du LIMSI (Laboratoire pour l'Ingénierie et la Mécanique des Sciences de l'Industrie) de l'université Paris 11 à la fin de mon Doctorat. Je n'ai eu de cesse d'approfondir ces collaborations depuis, et un des chercheurs CNRS du LIMSI (Martine Adda-Decker) est venu rejoindre notre UMR, de même qu'un enseignant-chercheur en Traitement Automatique des Langues (Kim Gerdes). Depuis 2018, avec des chercheurs et enseignants-chercheurs de l'université d'Avignon (Laboratoire d'informatique d'Avignon), mais également de Toulouse (Institut de Recherche en Informatique de Toulouse), j'essaye de tirer profit des capacités de l'apprentissage profond (Les Réseaux de Neurones Convolutifs) pour l'appliquer à la phonétique, et à la linguistique dans son ensemble.

Je présenterai ci-dessous de façon quasi-chronologique, les recherches que j'ai menées depuis la fin de ma thèse, essentiellement axées sur la variabilité en parole continue, notamment sur les voyelles orales du français. Je commencerai en présentant les mesures acoustiques automatiques effectuées sur les voyelles, avec les précautions mises en place. Dans un deuxième temps, ces analyses seront approfondies en abordant la modélisation statistique de ces voyelles, ainsi que leur comparaison à travers plusieurs styles de parole, et plusieurs langues. La troisième partie de mes activités scientifiques poursuivra le fil logique de ces investigations en abordant les aspects linguistiques, à savoir la réalisation des voyelles en fonction de leur position morpho-syntaxique, ainsi que des analyses plus spécifiques à la mélodie. Des comparaisons interlangues sont à nouveau effectuées afin de délimiter les caractéristiques propres à chaque langue et à son système. Dans la quatrième section, trois exemples caractéristiques par leur problématique située entre la phonétique et la phonologie sont présentés : le cas du schwa, du /ʁ/, et un exemple de variation diachronique. Enfin, je terminerai par une discussion de ces résultats, de leurs implications et de leurs limites, et je me projetterai sur une nouvelle thématique de recherche abordée depuis 2019 : la recherche d'invariants chez le locuteur, avec l'inclusion de méthodes issues de l'apprentissage profond.

4.1 Analyses vocaliques automatiques et valeurs de référence du français

Après mon doctorat, je me suis intéressé à l'hypo- et l'hyper-articulation des voyelles : les positions dans l'énoncé où se concentrent ces phénomènes, et leurs origines. Mes travaux dans cette optique ont commencé par une collaboration avec Martine Adda-Decker qui effectuait ses recherches au LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Industrie) à l'époque dans l'équipe TLP (Traitement du Langage Parlé). Nos objectifs étaient semblables bien que provenant de disciplines complémentaires. En effet, je souhaitais obtenir plus de données à analyser afin de confirmer des hypothèses : par exemple, j'avais analysé des catégories prosodiques sur de la parole de laboratoire

dans ma thèse et j'avais à cœur de vérifier si ces catégories pouvaient être retrouvées dans des corpus de parole non contrôlée. Martine Adda-Decker, quant à elle, voulait obtenir plus de connaissances sur la parole ; elle avait surtout accès à des corpus conséquents (d'environ 2 à 5 heures pour commencer) et pour lesquels une transcription manuelle avait été effectuée et vérifiée, et un alignement au niveau du phonème très précis, et renommé comme tel au niveau international.

4.1.1 Présentation des premiers corpus de travail

Les données de parole utilisées correspondaient dans un premier temps à deux heures de parole journalistique issus de journaux télévisées de France Inter. Ils n'ont eu de cesse d'être augmentés et nous nous sommes servis rapidement par la suite du corpus ESTER (Galliano et al., 2006).

Lors de la campagne ESTER (Evaluation des Systèmes de Transcription Enrichie d'émissions Radiophoniques), financée par le programme interministériel français TECHNOLANGUE et organisée conjointement par l'Association Française de la Communication Parlée (AFCP), la Direction Générale à l'Armement (DGA) et *Evaluations & Language Resources Distribution Agency* (ELDA), un corpus d'environ 100 heures de parole journalistique d'émissions radiophoniques, de différentes stations de radio, a été distribué aux participants. Même si une bonne partie des enregistrements correspondent à de la parole préparée, i.e. produite à partir d'un support écrit, elle est « convertie » à l'oral par des présentateurs et des présentatrices professionnels. Un pourcentage non négligeable des émissions correspond également à des interventions d'invités ou d'auditeurs, pour lesquelles il y a souvent peu ou pas de préparation écrite. On a dans ce cas une langue orale, certes influencée par l'écrit, mais qui s'éloigne d'une simple oralisation de l'écrit. Le passage de la lecture à la parole radiophonique a un impact au niveau des prononciations (entre autres), avec des réalisations qui peuvent s'écarter de manière importante de prononciations canoniques. Par opposition à une tâche de lecture sans auditoire, qui consiste à prononcer chaque mot écrit de manière canonique, les émissions radiophoniques sont destinées à un large public dispersé et distant, et le souci de compréhension globale prévaut certainement ici sur celui d'une production orale reflétant fidèlement une forme écrite. La parole est produite à un rythme soutenu mais quelques hésitations, répétitions et faux-départs viennent ponctuer des structures syntaxiques qui restent souvent proches de l'écrit. Afin de permettre d'estimer des modèles de langue adaptés à l'oral journalistique (par opposition aux journaux écrits), la DGA, en collaboration avec le LIMSI, a entrepris la transcription manuelle de dizaines voire centaines d'heures de journaux radiodiffusés à la fin des années 1990. Cette transcription orthographique s'est faite de façon normée, sans trucages orthographiques (comme « i'coup'nt les arb'es »). Ce principe permet de converger au mieux vers des transcriptions stables indépendantes du transcripteur et également plus rapides. Les modèles acoustiques de mots devront ensuite être adaptés pour prendre en compte les variations de production possibles pour une seule orthographe.

La transcription orthographique en séquences n'étant pas suffisante pour une analyse phonétique, il est nécessaire d'avoir recours à un alignement phonétique (cf. Figure 1). La transcription orthographique est utilisée par le système d'alignement pour localiser les frontières de début et de fin de mots, pour choisir parmi les alternatives potentielles de prononciation (en particulier les liaisons et la présence de schwas), et pour identifier les silences, respirations et autres bruits. Des modèles de phones indépendants du contexte sont utilisés pour l'alignement. Alors que des modèles acoustiques dépendants du contexte (triphones) produisent des transcriptions de meilleure qualité (c'est-à-dire des taux d'erreur mot plus faibles), les modèles acoustiques indépendants du contexte sont plus

efficaces pour repérer les frontières de phonèmes (Bürki et al., 2008). Pour des raisons techniques, la précision de la segmentation est limitée à 10 ms et la durée minimale d'un phonème est de 30 ms. L'étiquetage ainsi produit n'est pas un étiquetage phonétique, mais plutôt un étiquetage phonémique guidé par la transcription standard des mots. Les variations dans la prononciation des sons pourront donc être évaluées par des mesures acoustiques telles que des mesures de formants.

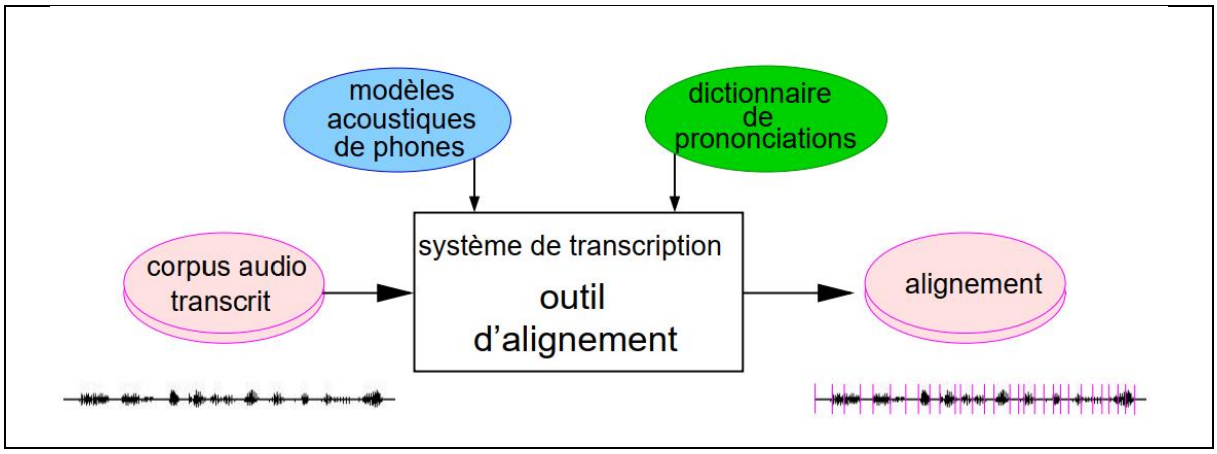


Figure 1 : alignement du corpus audio. La transcription orthographique du corpus permet de sélectionner les transcriptions phonémiques possibles dans le dictionnaire de prononciation. La segmentation est optimisée à l'aide des modèles acoustiques et des variantes de prononciation, d'après Gendrot et Adda-Decker (2004)

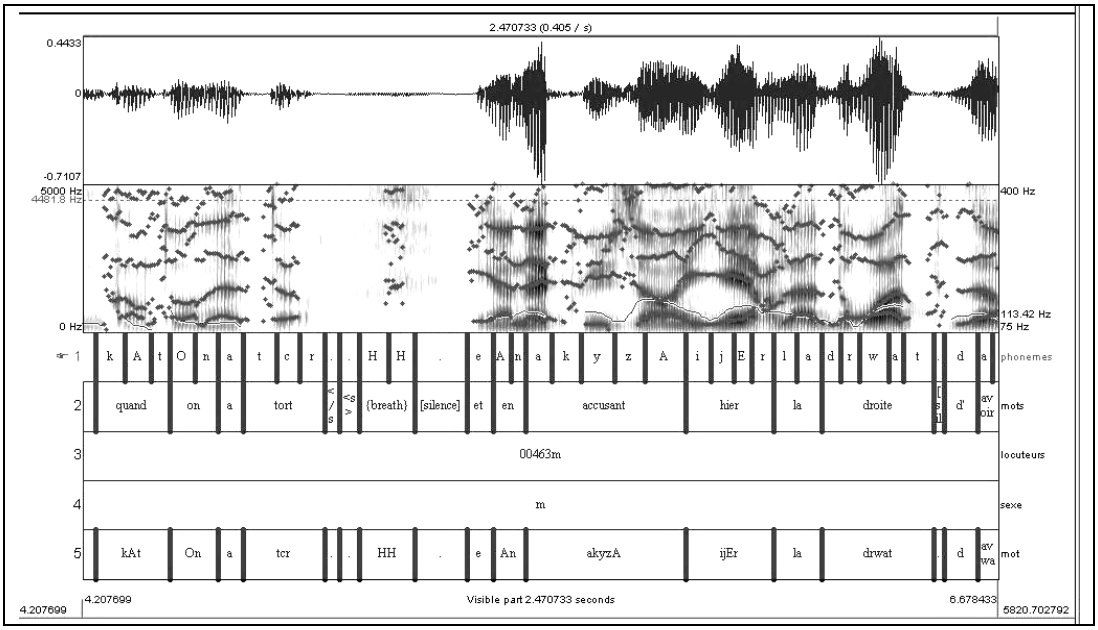


Figure 2 : illustration de l'alignement et de l'annotation réalisées automatiquement. L'information fournie par la segmentation est de haut en bas : (1) phonèmes ; (2) mots ; (3) code du locuteur ; (4) sexe du locuteur ; transcription phonémique du mot

Pour les travaux effectués sur le corpus ESTER, 30 heures de parole radiophonique (environ 500 hommes et 300 femmes) ont été extraites du corpus ESTER, car cette quantité correspondait à la segmentation fournie par l'IRISA ("Institut de Recherche en Informatique et Systèmes Aléatoires") sur le site de l'AFCP (http://www.afcp-parole.org/ressources/corpus-et-outils/, non disponible

désormais). Les procédures de transcription et de segmentation, et surtout leur fiabilité, sont détaillées ci-dessous.

4.1.2 Présentation phonétique du corpus ESTER et préparation des analyses

La première approche que nous avons menée avec Martine Adda-Decker (Gendrot et Adda-Decker, 2005) avait pour but de présenter les distributions et réalisations des différentes voyelles du français dans un corpus de parole non contrôlée. Le but était de fournir une référence réalisée sur un corpus conséquent et non plus sur quelques occurrences lues ou bien quelques dizaines de minutes de parole soigneusement choisies et annotées. Il est nécessaire de garder en mémoire que la parole journalistique conserve des spécificités propres à son style de parole. Par exemple, l'usage du pronom « vous » est de rigueur et à l'inverse le pronom « tu » n'aura que très peu d'occurrences. Cette différence parmi de nombreuses autres pourra avoir des conséquences sur la fréquence de la voyelle /u/ par opposition à /y/, ainsi que sa réalisation acoustique puisqu'influencée par son contexte phonémique immédiat. Dans la suite de ce mémoire, nous ferons un travail de comparaison des différents styles de parole, mais cette première approche était nécessaire et permettait d'établir une référence qui manquait dans la littérature au milieu des années 2000.

Le Tableau 6 montre le nombre de voyelles dans le corpus brut (dont un peu plus de 40 % sont des voyelles). Le schwa (/ə/) et le /œ/ y sont comptabilisés ensemble. Ceci est dû à la représentation dans le système de reconnaissance, où ils correspondent à un seul modèle acoustique. Les contextes consonantiques ont été regroupés en quatre catégories choisies pour leur influence coarticulatoire sur la réalisation acoustique de la voyelle, à savoir pour les labiales : p,b,f,v,m, pour les dentales : t,d,n,s,z, pour les palato-vélaires : k,g,ʃ,ʒ,ɲ,ŋ, et pour les uvulaires, ʀ. Nous pouvons observer au premier abord que les voyelles les plus fréquentes sont dans l'ordre a, i, e, œ (regroupant œ et ə) et ɛ. Ensuite seulement nous pouvons trouver ɑ̃, ɔ, u, y, ɔ̃, ɛ̃, o et ø.

	a	e	i	o	u	y	ɔ	ɛ	ø	œ	total
p-p	565	251	294	240	246	293	445	237	102	198	2871
p-t	3563	1352	4408	1073	405	1038	1513	2384	269	850	16855
p-k	849	168	940	302	196	50	246	315	35	97	3198
p-r	2190	1396	448	38	480	96	1672	2171	17	157	8665
t-t	6133	4857	8578	1038	622	2045	2123	4267	185	2100	31948
t-p	3719	2698	3470	294	678	569	1524	572	132	2576	16233
t-k	1860	1977	2531	132	222	357	512	1128	18	597	9334
t-r	1278	920	1536	33	487	823	1365	2645	226	1506	10819
k-k	473	109	121	68	20	68	68	123	6	73	1129
k-p	601	209	427	79	420	130	648	230	12	164	2920
k-t	2416	827	835	444	375	907	1144	1009	69	399	8425
k-r	1035	249	142	10	1275	183	521	364	10	182	3971
r-r	99	7	120	5	–	5	77	130	30	88	564
r-p	1276	708	692	161	456	39	1043	556	20	1586	6537
r-t	1946	1565	2263	546	150	273	650	2122	93	1786	11394
r-k	778	619	1355	83	45	128	544	502	16	585	4665

Tableau 1 : nombre d'occurrences des voyelles orales dans différents contextes consonantiques observés dans le corpus ESTER (p : labial ; t : dental ; k : palatovélaire ; r : uvulaire), d'après Gendrot et Adda-Decker (2010)

Notre travail avait également pour but de contribuer à l'établissement de valeurs de formants et de leur variabilité en français, qui étaient encore peu documentées à l'époque. Il était possible de trouver des valeurs formantiques dans Calliope (Calliope, 1989) mais celles-ci étaient mesurées sur de la parole lue et en contexte consonantique p _ R, ce qui avait pour conséquence, notamment de par la présence du /R/ de modifier l'articulation de la voyelle de façon conséquente. Notre intérêt s'est porté principalement sur les variations de F1 et de F2 qui peuvent être interprétées – si on ne tient pas compte de l'effet des lèvres – en termes d'aperture/fermeture (corrélée à F1) et d'antériorité/postériorité (corrélée à F2). F3 a également été mesuré puisqu'il est plus globalement corrélé à l'arrondissement des lèvres. Notons cependant que l'arrondissement des lèvres a pour effet d'allonger le conduit vocal et aboutira à une baisse de l'ensemble des formants. Lorsque des quantités de voyelles à analyser dépassent les quelques dizaines de milliers, il devient impossible d'envisager de les analyser manuellement. En 2005, il était encore peu fréquent d'analyser automatiquement les formants des voyelles et nous devions nous assurer de respecter nombre de précautions méthodologiques afin de valider notre approche. Les voyelles nasales ont été retirées par la suite de ces analyses puisque l'ouverture de la cavité nasale génère des anti-formants dans les zones des deux premiers formants dont l'analyse devient de facto hasardeuse (M. Y. Chen, 1997 ; Maeda et al., 1993). Nous avons utilisé le logiciel PRAAT (Boersma, 2001) qui permettait une automatisation relativement aisée grâce à son langage de programmation intégré. La totalité des scripts mis en place furent d'ailleurs mis à disposition sur ma page web afin de lancer la communauté de parole dans cet essor qu'était l'analyse automatique de la parole. PRAAT proposait déjà en 2005 un algorithme de détection automatique des formants qui pouvait être adapté selon le sexe du locuteur. La détection de pics d'amplitude -déterminés dans une plage inférieure à 5kHz pour les hommes et inférieure à 5.5kHz pour les femmes- était ainsi effectuée grâce à l'algorithme Burg. Les paramètres choisis impliquent une fenêtre d'analyse de 25 ms[2]. Trois mesures ont été effectuées respectivement à 1/3, ½ et 2/3 de la voyelle, puis moyennés pour fournir une valeur unique. Cette méthode avait l'avantage de fournir une valeur moyennée pour chaque voyelle, mais en cas de détection erronée sur l'une de ces trois valeurs, la mesure obtenue pouvait être erronée dans son ensemble. L'interprétation de ces pics d'amplitude en tant que formants peut poser problème sur un nombre non négligeable de segments (Gendrot et Adda-Decker, 2005) :

- des bruits ou de la musique peuvent se superposer sur la parole gênant ainsi la détection des formants
- une fréquence fondamentale trop élevée pour les voix de femmes et d'enfants, ou bien dans le cadre de la voix émotionnelle, qui en espaçant les harmoniques va diminuer le renforcement des pics d'amplitude spectrale.
- la nasalité d'une voix ou une voyelle coarticulée.

Nous avons eu recours à un filtrage des valeurs de formants (Gendrot et Adda-Decker, 2005) obtenues afin de rejeter les résultats aberrants par rapport à l'acoustique du conduit vocal. Nous avons établi des fourchettes larges (cf. Tableau 2) pour les hommes et les femmes basées sur les connaissances acoustiques mais également sur la base d'erreurs de détection fréquemment observées lors de

[2] La fréquence d'échantillonnage utilisée pour le corpus ESTER est de 16000 Hz, fréquence souvent utilisée dans le domaine de la reconnaissance automatique de la parole comme étant un juste compromis entre la taille des fichiers et la quantité d'information spectrale ; dans la terminologie du traitement du signal, une Transformée de Fourier Rapide classique sur 512 points correspond donc pour cette fréquence d'échantillonnage à une fenêtre de signal de 32 ms

l'utilisation du logiciel. Si l'une des valeurs pour ces 3 formants est mesurée en dehors de sa fourchette, la voyelle correspondante est rejetée par le filtre. Une centaine de vérifications manuelles par voyelle a également été opérée afin de s'assurer que les voyelles rejetées étaient bien causées par des problèmes importants de détection de formants et non des valeurs éloignées de leur cible qui pourraient être dues à un phénomène de coarticulation, ou à des caractéristiques du locuteur par exemple. Le concept de cible acoustique, et de la non atteinte de cette cible ('target undershoot') étant un phénomène que l'on souhaitait analyser par la suite (Lindblom, 1963 ; Perrier et al., 1996).

Hommes										
	i	y	e	ɛ	a	œ	ø	ɔ	o	u
F1	<750	<800	<800	<1000	<1000	<1000	<900	<900	<900	<900
F2	1500 - 2500	1300 – 2200	1100 - 2400	1200 - 2300	800 - 2300	800 - 2000	700 - 2000	600 – 1800	600 - 1600	400- 1500
F3	> 2000	> 1700	> 2000	> 2000	> 1800	> 2000	> 1700	> 1500	> 1500	> 1400
Femmes										
F1	< 900	< 900	< 900	< 1100	< 1100	< 1100	< 1000	< 1000	< 1000	< 1000
F2	1600 – 3100	1400 -2800	1400 - 3000	1400 - 2700	900 - 2300	800 - 2400	700 - 2300	600 - 2000	600 - 1600	400-1500
F3	> 2500	> 1800	> 2200	> 2000	> 1900	> 2000	> 1800	> 2000	> 2100	> 1800

Tableau 2 : fourchettes de valeurs utilisées pour le premier filtrage de formants (minimum et maximum) en Hertz, d'après Gendrot et Adda-Decker (2005)

Après ce filtrage, approximativement 4 % de voyelles ont été rejetées. La majeure partie de ces rejets correspondent à des segments de très courte durée (60 % des segments rejetés ont une durée inférieure à 50 ms). A l'écoute de ces voyelles, il nous est apparu que la segmentation n'était pas en cause, à fortiori pour les voyelles dont la durée effective est supérieure est à 30 ms. D'autres raisons peuvent expliquer ces rejets, en particulier un dévoisement partiel ou total des voyelles qui rend la détection des formants, y compris manuelle, difficile voire impossible, et produit ainsi des valeurs de formants aberrantes. Une deuxième raison qui vient expliquer les erreurs de détection est la proximité de deux formants pour les voyelles dites « focales », particulièrement dans les basses fréquences comme pour les voyelles hautes postérieures par exemple. L'algorithme détecte ainsi un seul formant au lieu de deux, ce qui provoque un décalage d'attribution vers les formants réels supérieurs. La voyelle /u/ est particulièrement concernée par ces rejets, puisque ses formants F1 et F2 sont très bas (en dessous de 1000 Hz) et proches l'un de l'autre. Mais on peut également mentionner F2 et F3 pour /y/. Pour /i/, on note une proximité entre F3 et F4, mais notre filtrage n'a pas porté sur les formants supérieurs à F3. Les proportions rejetées pour chaque voyelle et présentées dans le Tableau 3 illustrent ce point avec des taux plus élevés par ordre décroissant pour /u/, /y/ et /o/.

voyelle	i	y	e	ɛ	a	œ	ø	ɔ	o	u
Taux de rejet (en %)	5.0	15.0	1.0	0.3	0.6	4.0	0.4	1.0	4.9	25.0

Tableau 3 : proportion des segments rejetés en fonction de leur identité, d'après Gendrot et Adda-Decker (2005)

Nous avions émis l'hypothèse que la détection automatique de formants serait plus problématique pour les femmes que pour les hommes puisque pour les femmes, la séparation plus importante des

harmoniques dans le spectre ne favorise pas un renforcement aussi fort des zones spectrales correspondant aux cavités de résonance. Mais les résultats ont montré qu'au contraire, les taux de rejets sont plus élevés pour les hommes que pour les femmes (4.76 % vs. 2.9 % au sein de chaque classe). Une explication possible (Labov, 1972 ; Traunmüller, 1984) serait que l'articulation des voyelles produites par les femmes serait plus précise afin de compenser le manque de renforcement des pics formantiques. Une autre explication possible serait la plus grande distanciation des formants chez les femmes par la petitesse des cavités et qui se retrouvent sur l'espace vocalique beaucoup plus grand occupé en superficie par les voyelles des femmes.

Le Tableau 4 montre le taux de segments rejetés en fonction de leur durée segmentée automatiquement répartis en 3 catégories [30-50 ; 60-80 ; 90-110], ces catégories seront d'ailleurs reprises par la suite pour l'interprétation des mesures de formants. Ces intervalles ont été déterminés comme la distribution la plus régulière au sein de chaque catégorie. Rappelons que la durée minimale d'un segment lors de l'alignement automatique est de 30 ms et la précision est de 10 ms (3 états HMMs de 10 ms chacun). Sans surprise, les résultats montrent que le filtrage élimine les voyelles plus courtes de façon prépondérante. Au-delà des raisons précédemment invoquées pour expliquer le filtrage des valeurs erronées, nous pouvons ajouter que lors de prononciations rapides et peu articulées, plusieurs segments sont automatiquement segmentés à la durée minimale de 30 ms ce qui peut entrainer des erreurs de segmentation et donc des erreurs de mesures.

Catégorie durée (en ms)	[30 - 50]	[60 - 80]	[90 – 110]
proportion (en %)	39	38.5	22.5
Taux de rejet (en %)	6.1	2.8	2.4

Tableau 4 : proportion de segments rejetés en fonction de leur durée, d'après Gendrot et Adda-Decker (2005)

Le Tableau 5 (d'après *Gendrot et Adda-Decker (2010)*) montre la durée des voyelles analysées et leur répartition au sein de chaque catégorie de durée et révèle des différences de durée d'une voyelle à une autre. Il faut garder en tête que les différences observées ici peuvent être issues des caractéristiques articulatoires de chaque voyelle, les voyelles arrondies sont en moyenne plus longues que les voyelles étirées par exemple, mais également de par leur contexte d'utilisation (Keating, 1985 ; Lehiste, 1970). La voyelle /a/ qui se trouve fréquemment au sein du mot grammatical « la » verra sa durée fortement raccourcie dans ces contextes, ce qui influera en retour sur sa durée moyenne.

voyelle	a	e	i	o	u	y	ɔ	ɛ	ø	œ	moyenne
durée moyenne	70.0	74.0	76.0	104.0	87.0	84.0	69.0	76.0	84.0	68.0	75.0
cat. courte en %	39.2	39.2	30.4	35.5	20.6	23.1	37.3	35.6	32.0	46.6	35.3
cat. moyenne en %	40.6	37.5	42.7	35.6	36.4	42.0	42.9	37.2	32.7	34.4	39.4
cat. longue en %	20.2	23.4	26.8	51.6	37.6	34.9	19.9	27.3	35.3	19.1	25.4

Tableau 5 : durée moyenne des voyelles orales et leur proportion au sein de chaque catégorie de durée

4.1.3 Précautions méthodologiques pour l'analyse automatique

Dans cette section, je relate un travail effectué (Büerki et Gendrot, 2007 ; Bürki et al., 2008) en collaboration avec Audrey Bürki et Cécile Fougeron, et différents collaborateurs de laboratoires d'informatique (Georges Linares, Guillaume Gravier et Martine Adda-Decker). Nous nous sommes intéressés en particulier aux outils automatiques d'alignement en phonèmes et à leur utilisation dans le cadre d'études phonético-phonologiques. L'alignement en phonèmes prend comme source un signal de parole et détermine la succession des phones produits ainsi que leurs frontières. Sa réalisation manuelle étant extrêmement laborieuse et parfois sujette à subjectivité, le recours à un alignement automatique permet un gain de temps considérable, mais améliore également la consistance des résultats. Plusieurs corpus alignés automatiquement – ou outils d'alignement – ont été mis à disposition et utilisés par des phonéticiens à partir du milieu des années 2000. (voir entre autres Fougeron et al., 2007 ; Gendrot et Adda-Decker, 2005 ; Kuperman et al., 2007). La question abordée ici concerne la fiabilité d'une telle démarche et plusieurs points sont ici à considérer. Il existe tout d'abord différents systèmes d'alignement dont les caractéristiques sont susceptibles de donner lieu à des performances inégales (Van Bael et al., 2007). Ces systèmes ont par ailleurs, pour nombre d'entre eux, été développés dans le cadre de la reconnaissance de la parole. Or, un fort taux de reconnaissance correcte des mots ne va pas toujours de pair avec un alignement en phonèmes des plus adéquats (Kessens et Strik, 2004). Il est donc nécessaire de s'assurer de la fiabilité d'un alignement avant d'effectuer des analyses acoustiques. Les données utilisées ici correspondent à 30 heures extraites du corpus ESTER, et donc à une parole de type journalistique. Les alignements obtenus ici seront à priori plus fiables sur de la parole journalistique que sur de la parole spontanée qui comporte beaucoup plus de phénomènes de réductions. Dans le cadre de l'étude que je résume ici, les alignements en phonèmes de trois systèmes automatiques sont évalués et comparés, relativement à leur pertinence pour l'étude linguistique d'une voyelle susceptible d'être élidée, le schwa en français (voir section 4.3.1 pour plus d'informations sur l'état de l'art quant à l'élision du schwa). Il s'agit tout d'abord de déterminer dans quelle mesure les segmentations obtenues sont à même de rendre compte de cet objet linguistique, et, par-là, d'être utilisées pour des analyses linguistiques fines. Nous tentons également de dégager les régularités et les facteurs qui régissent les décisions des systèmes et leurs erreurs au regard d'un alignement manuel de référence : la pertinence d'une généralisation des données obtenues est évaluée.

La difficulté majeure est liée à la nécessité de pouvoir comparer le résultat obtenu avec un alignement de référence. Un alignement manuel est en général utilisé à ces fins. Or, plusieurs études ont comparé les performances de différents transcripteurs manuels et montrent des disparités parfois importantes. Deux options ont été proposées dans la littérature pour établir une référence acceptable. La première consiste à ne considérer dans les évaluations que le matériel linguistique sur lequel tous les transcripteurs s'accordent. C'est la méthode du consensus. La seconde option consiste à construire une transcription de référence fondée sur le vote majoritaire. Ces deux propositions présentent cependant des inconvénients majeurs lorsque l'on s'intéresse aux corpus à des fins d'analyses linguistiques. Les désaccords entre transcripteurs peuvent être porteurs d'informations linguistiques pertinentes et la non prise en compte des « cas limites » est susceptible d'orienter sensiblement les résultats. De même, les frontières difficiles à placer pour les humains sont également source de difficultés pour la machine. Nguyen et Espesser (2004) ont montré que la congruence entre les alignements manuel et automatique est meilleure en début qu'en fin de voyelle, la durée attribuée aux segments par l'aligneur automatique étant plus courte. Ils ont également montré que, si le milieu de la voyelle est localisé avec une bonne précision (écart inférieur à 20 ms) dans 75 % des cas, la précision varie en fonction de la position de la voyelle dans le mot et du contexte segmental droit. Ces

différents résultats suggèrent que les écarts entre alignements manuel et automatique ne concernent pas de manière univoque tous les phonèmes ni les contextes, que le type d'erreur considéré ou la mesure rapportée influencent le résultat et que différents facteurs semblent gouverner ces écarts. Les auteurs ont également rapporté que le schwa est le phonème le plus fréquemment impliqué dans les erreurs touchant les voyelles.

Dans notre étude (Büerki et Gendrot, 2007 ; Bürki et al., 2008), plusieurs systèmes d'alignement automatique ont été évalués dans leur capacité à rendre compte de la voyelle schwa en français. La voyelle schwa (ou « e » muet) a la particularité phonologique de pouvoir être élidée. Autrement dit, un mot comportant un schwa, comme « semaine », peut être prononcé avec ce dernier ([səmɛn]) ou sans ([smɛn]). Les raisons de choisir le schwa comme angle d'attaque étaient multiples. Il était alors difficile de brosser un tableau consensuel concernant la réalisation du schwa. Par ailleurs, le schwa pose des problèmes non négligeables aux systèmes d'alignement utilisés pour la synthèse de la parole (Lanchantin et al., 2008), ainsi qu'aux systèmes de reconnaissance. Adda-Decker relève en 2007 (Adda-Decker, 2007) que 5 % des erreurs de reconnaissance impliquent des omissions, insertions ou confusions liées au schwa. Mieux cerner sa gestion par les systèmes d'alignement pouvait donc être profitable à la recherche en Traitement Automatique de la Parole. De plus, comprendre comment les différents systèmes d'alignement se comportent vis-à-vis de l'élision du schwa pouvait nous offrir un premier aperçu des problèmes que peut poser le phénomène plus large de réduction/effacement de segments dans la parole continue en français et dans d'autres langues.

4.1.3.1 Comparaison des alignements de deux juges

Les items choisis pour étudier la fiabilité des systèmes d'alignement, tant dans la détection du schwa que dans la précision de son alignement, sont présentés ci-dessous (Bürki et al., 2008 ; Büerki et Gendrot, 2007). 22 773 occurrences pour un total de 583 mots distincts contenant un schwa ont été extraites du corpus ESTER. Seuls les mots segmentés de manière alternante dans le corpus tel que le proposait l'alignement automatique ont été retenus pour l'analyse (mots ayant été segmentés avec et sans schwa, par exemple le mot « semaine » segmenté 84 fois [səmɛn] et 36 fois [smɛn]). 5 016 occurrences ont été soumises à une analyse auditive et acoustique. 230 séquences impropres à l'analyse (5 % des données) ont été exclues : 135 occurrences produites par des locuteurs jugés non francophones sur la base de leur accent (22 locuteurs) et 95 séquences inintelligibles ou correspondant à des erreurs de reconnaissance. Une première correction manuelle de la segmentation automatique (effectuée par un juge et entérinée par un second) a permis d'éliminer 67 mots (479 occurrences) ayant été produits de manière non variable (uniquement avec ou uniquement sans schwa) dans le corpus. Au final, 4 307 occurrences ont été retenues pour l'analyse.

Afin d'établir un alignement de référence, nous avons opté pour une démarche qui tienne compte des différentes considérations évoquées dans la littérature. Un premier juge a corrigé la segmentation automatique fondé sur des monophones, l'objectif étant de pouvoir ensuite utiliser cet alignement pour l'étude du schwa. Un second juge a fait de même sur 47 % des occurrences, en ayant accès à la segmentation du premier juge. Un degré d'accord a été calculé afin de pouvoir ensuite estimer les résultats des systèmes automatiques. En effet, afin de déterminer si un alignement phonétique obtenu automatiquement est satisfaisant ou non, il est nécessaire de disposer d'une ligne de base. Plusieurs auteurs utilisent à cet effet le degré d'accord entre deux alignements manuels (Cucchiarini et Strik, 2003). Si le degré d'accord entre alignement automatique et alignement manuel est similaire à celui observé entre deux alignements manuels, le système d'alignement est jugé satisfaisant.

Dans un premier temps, nous avons évalué la présence vs. absence du schwa. À partir de l'alignement automatique de l'IRISA fondé sur des monophones, plusieurs modifications ont été effectuées par chacun des juges. Les schwas non détectés ont été ajoutés, les schwas détectés par erreur (insertions)

éliminés. Un critère uniforme a été appliqué : un schwa a été considéré comme réalisé en présence à la fois de périodicité dans le signal et d'une structure formantique. La correction des estimations de durée a ensuite été effectuée sur la base de l'apparition/disparition de périodicité sur le signal acoustique et d'un deuxième formant sur le spectrogramme. Le degré d'uniformité entre les deux alignements manuels a porté sur la présence de la voyelle, l'estimation de sa durée et le placement des frontières. Des pourcentages d'accord ont tout d'abord été calculés. L'accord global (nombre d'occurrences pour lesquelles le jugement des deux juges est identique sur le nombre total de jugements) équivaut à 99 %. L'accord est de 98 % sur les schwas présents et de 99 % sur les schwas absents. La relation entre les deux jugements est significative ($\chi^2(1) = 1914.9$, $p < 0.0001$) et de force importante ($\phi = 0.977$). La seconde mesure souvent utilisée pour mesurer le degré d'accord entre deux juges sur des données catégorielles est le coefficient de Kappa. Ce dernier permet de corriger les pourcentages en tenant compte du hasard et rend possible une comparaison entre des valeurs issues de différentes conditions. Le coefficient obtenu ici (0.98), particulièrement élevé, témoigne d'un accord « presque parfait ».

Dans un second temps, nous avons évalué la précision de la durée des schwas. La durée moyenne des 1 420 schwas jugés présents par les deux juges a été calculée. Elle est de 52 ms, pour le premier et de 53 ms pour le second. Le coefficient de corrélation entre les estimations des deux juges est élevé ($r = 0.966$, $n = 1\,420$, $p < 0.01$) et ce malgré une différence de durée significative ($t(1\,419) = 8.38$, $p < 0.001$). 95 % des frontières placées par les deux juges tombent dans un intervalle de 0 à 10 ms et 99 % tombent dans un intervalle de 0 à 20 ms. Ces valeurs sont plus hautes que celles rapportées notamment par Wesenick et Kipp (1996), pour des segments consonantiques (87 % des frontières dans un intervalle de 10 ms et 96 % dans un intervalle de 20 ms). La déviation moyenne est de 9 ms pour le début de la voyelle et de 6 ms pour sa fin. Ces valeurs sont plus basses que ce qui a été obtenu par Pitt et al. (2005) dont l'évaluation est fondée sur l'alignement de quatre transcripteurs et concerne l'ensemble des segments vocaliques et consonantiques (écart moyen de 16 ms). Par ailleurs, le second juge tend à placer ses frontières de début de voyelle plus à gauche alors que les frontières de fin de voyelle sont décalées vers la droite (i.e. les frontières sont considérées plus « vocaliques » que par l'autre juge).

Dans leur ensemble, ces résultats révèlent un degré d'accord extrêmement important entre les deux alignements manuels, généralement plus haut que ce qui est rapporté dans la littérature. Les comparaisons restent cependant difficiles, étant donné les différences importantes de matériel et de méthodologie. Le haut degré d'accord obtenu ici est probablement à rapporter à la méthode hybride utilisée, entre des alignements indépendants (qui donneraient des résultats plus faibles) et un consensus ; n'oublions pas également que nous ne sommes concentrés que sur un phonème, la tâche des juges était donc réduite. Ces résultats nous permettent en définitive de recourir, pour l'évaluation des alignements automatiques, à l'alignement du premier juge, portant sur la totalité des occurrences (Bürki et al., 2008).

4.1.3.2 Présentation des systèmes d'alignement

Trois systèmes d'alignement sont évalués dans la présente étude (Büerki et Gendrot, 2007 ; Bürki et al., 2008), tous développés dans le cadre de la reconnaissance de la parole continue. Un système a été développé au LIA (Laboratoire Informatique d'Avignon) et les deux autres à l'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires). Les deux systèmes de l'IRISA, que l'on appelle « IRISA monophones » et « IRISA triphones », sont en fait deux variantes d'un même système. Les trois systèmes d'alignement ont recours à des HMM. Ils se distinguent en revanche sur deux points : le dictionnaire de prononciations et les modèles acoustiques utilisés. Pour les trois systèmes, les dictionnaires de prononciations ont été construits à partir de phonétiseurs automatiques et corrigés à la main. Pour un mot donné, les différentes prononciations, appelées variantes, sont considérées

comme équiprobables et ne dépendent pas du contexte syntaxique dans lequel un mot est prononcé. Dans les deux systèmes de l'IRISA, les prononciations ont été établies à partir du dictionnaire ILPho (Boula de Mareüil et al., 2000), fondé sur des règles de phonétisation. Seules les variantes de prononciations les plus probables ont été conservées, avec un nombre moyen de variantes de prononciations par mot de 1.8. Le système du LIA, quant à lui, utilise le phonétiseur automatique à base de règles LIA_PHON (Béchet, 2001) dont les sorties ont été corrigées manuellement, pour générer une moyenne de 4.54 variantes de prononciation par mot.

Les modèles acoustiques utilisés par les systèmes du LIA et de l'IRISA ont été estimés sur le corpus d'apprentissage de la base ESTER (Galliano et al., 2006), composé de 80 heures d'émissions radiophoniques. Ces modèles diffèrent par leur structure, leur prise en compte ou non du contexte phonétique et leur taille (nombre de composantes gaussiennes utilisées). Les deux systèmes de l'IRISA utilisent un ensemble de 35 phonèmes représentés par des HMM à 3 états qui peuvent être, selon le système, soit indépendants du contexte phonétique (système « IRISA monophones »), soit dépendants du contexte phonétique et du sexe du locuteur (système « IRISA triphones »). Dans les deux cas, la structure des modèles utilisés impose une durée minimale des phonèmes de 30 ms. Les modèles indépendants du contexte possèdent au total 114 états modélisés chacun par 128 gaussiennes, soit un total de 14 592 gaussiennes. Les modèles dépendants du contexte possèdent, quant à eux, un total de 8 140 états correspondant à 260 480 gaussiennes, pour les voix d'hommes comme pour les voix de femmes. Les modèles du système d'alignement du LIA comportent 3 400 états, soit environ 230 000 gaussiennes. Ils sont dépendants du contexte phonétique à l'intérieur des mots, mais indépendants de ce dernier en frontière de mots. Par ailleurs, la structure des modèles n'impose pas de contrainte de durée minimale aux phonèmes.

4.1.3.3 Evaluation des alignements automatiques

Chaque système d'alignement automatique a été évalué en termes de présence/absence de la voyelle, d'estimation de la durée et du placement des frontières au regard de l'alignement effectué par le premier juge. Outre les taux et types d'erreurs, l'influence du contexte consonantique et de facteurs acoustiques intrinsèques au schwa a été mesurée pour chacun des systèmes (Büerki et Gendrot, 2007 ; Bürki et al., 2008).

En raison d'écarts temporels trop importants entre l'alignement manuel et l'alignement automatique, un certain nombre d'occurrences n'ont pas pu être considérées dans les analyses. Les comparaisons de chaque système avec l'alignement manuel ne se font donc pas toujours sur le même nombre d'occurrences : 4 287 occurrences pour la comparaison des deux systèmes de l'IRISA avec l'alignement manuel (qu'on nommera « Manuel 1 ») et 4 130 occurrences pour la comparaison du système du LIA avec l'alignement manuel (« Manuel 2 »).

Les résultats concernant la présence vs. l'absence du schwa sont présentés ci-dessous. De manière similaire à ce qui a été entrepris pour la comparaison entre les deux alignements manuels, plusieurs mesures de fiabilité ont été calculées afin de déterminer le degré d'accord entre chacun des alignements automatiques et l'alignement manuel de référence. L'alignement du LIA diffère davantage de l'alignement manuel de référence que les alignements du système de l'IRISA. Le coefficient de Kappa révèle un accord substantiel (0.78) sur l'échelle proposée par pour le système du LIA, alors qu'il est « presque parfait » pour les deux alignements du système de l'IRISA.

Deux types d'erreurs interviennent dans les divergences entre alignement manuel et automatique. Des schwas absents peuvent être segmentés par le système automatique (insertions), des schwas présents peuvent ne pas être détectés. Une analyse distincte de ces erreurs est présentée ci-dessous.

Les trois systèmes se distinguent par le taux d'insertion de schwas dans leurs alignements (nombre d'occurrences alignées avec un schwa sur nombre total d'occurrences sans schwa dans l'alignement manuel). L'alignement du LIA présente le plus fort taux d'insertion (9 %, n = 1 220), suivi de l'alignement IRISA triphones (5 %, n = 1 250) et de l'alignement IRISA monophones (2 %, n = 1 250). Une analyse de régression logistique binomiale confirme l'impact du système sur le taux d'insertion (Chi2 Omnibus : $\chi^2(2) = 46.85$, $p < 0.0001$). La probabilité qu'un schwa soit inséré est plus faible pour le système IRISA monophones que pour le système IRISA triphones ($z = 3.05$, $p < 0.01$) ou pour celui du LIA ($z = 6.19$, $p < 0.0001$). Elle est également plus faible pour le système IRISA triphones que pour le système du LIA ($z = 3.67$, $p < 0.0001$).

Une analyse de régression logistique binomiale a été conduite afin d'évaluer l'influence du contexte consonantique (sonorité et mode d'articulation) sur la catégorie de détection (correct vs. insertion). Pour le système du LIA (n = 1 220), le modèle statistique complet prédit de manière significative la détection (Chi2 Omnibus : $\chi^2(8) = 99.8$, $p < 0.0001$) et explique entre 8 % (pseudo R2 de Cox et Snell) et 18 % (pseudo R2 de Nagelkerke) de la variance.

Système d'alignement	Consonne gauche	Consonne droite
LIA	voisée > sonante > sourde nasale > occlusive, liquide > fricative	liquide, fricative > nasale
IRISA monophones	fricative > occlusive, liquide	nasale > liquide, fricative sonante > sourde
IRISA triphones	voisée > sourde	voisée > sonante

Tableau 6 : influence des propriétés des consonnes environnantes sur le taux d'insertion pour chacun des systèmes d'alignement (les contextes à gauche du signe « > » génèrent davantage d'insertions que les contextes à sa droite)

Le Tableau 6 (Bürki et al., 2008)résume l'influence du contexte consonantique sur le taux d'insertion des schwas. Nous remarquons en particulier des influences contraires en ce qui concerne le mode d'articulation pour le système du LIA et celui de l'IRISA monophones, mais plus globalement la faible proportion des données expliquées par l'entourage consonantique. Nous présenterons dans la section 4.3.1. une analyse acoustique et perceptive de l'élision du schwa qui nous a permis d'expliquer en partie ces résultats obtenus par un système d'alignement.

Les trois systèmes se distinguent également par le taux de détection correcte/incorrecte pour les schwas présents dans l'alignement manuel de référence. Le système du LIA présente le plus grand taux de non-détection (10.4 %, n = 2 910) suivi de l'alignement « monophones » de l'IRISA (9.6 %, n = 3 037), puis de l'alignement « triphones » de l'IRISA (8 %, n = 3 037). Une analyse de régression logistique binomiale confirme l'impact du système sur le taux de non-détection (Chi2 Omnibus : $\chi^2(2) = 16.3$, $p < 0.001$). La probabilité qu'un schwa ne soit pas détecté est plus faible pour le système IRISA triphones que pour le système IRISA monophones ($z = 2.93$, $p < 0.01$) et que pour le système du LIA ($z = 3.87$, $p = 0.0001$). Une analyse de régression logistique binomiale a été conduite pour chaque système afin d'estimer la contribution de différents facteurs sur l'adéquation de la détection pour les schwas présents. Aux caractéristiques des consonnes entourant le schwa a été ajoutée la durée du schwa telle qu'elle a été obtenue manuellement. Ces résultats sont présentés dans le Tableau 7 et montrent que les trois systèmes sont fortement similaires en ce qui concerne les variables qui influent sur la détection des schwas présents. La durée est le facteur le plus influent et les traits analysés (mode d'articulation et sonorité) montrent toujours une configuration similaire lorsque leur rôle est significatif, qu'ils appartiennent à la consonne de gauche ou de droite.

Système d'alignement	Consonne gauche	Consonne droite	Durée
LIA	Sonante > voisée > sourde Liquide, fricative > nasale	Sonante > voisée > sourde	–
IRISA monophones	Sonante > voisée > sourde	Sonante > voisée, sourde Fricative, liquide > nasale	–
IRISA triphones	Sonante > voisée > sourde	Sonante > voisée > sourde Fricative, liquide > nasale	–

Tableau 7 : résumé des variables ayant une influence sur le taux de non-détection pour chacun des systèmes. Les contextes à gauche du signe « > » génèrent davantage d'erreurs que les contextes à sa droite. Pour la durée un «–» indique qu'une augmentation de la durée entraîne une diminution du taux de non-détection

4.1.3.4 Estimation de la durée du schwa

Outre la détection de la voyelle, il est également intéressant de connaître le comportement d'un système d'alignement quant au placement des frontières des phones. Un décalage important de ces dernières peut avoir des incidences sur les mesures de durée. Les analyses ci-dessous ont été effectuées sur les schwas détectés à la fois par l'alignement automatique et l'alignement manuel de référence et sont résumées dans le Tableau 8.

Le système du LIA attribue au schwa une durée plus importante que celle de l'alignement manuel de référence (55 ms vs. 52 ms (sd = 17)). La différence de durée, évaluée par un test-t apparié, est significative ($t(2\,607) = 10.47$, $p < 0.0001$). Le système IRISA triphones attribue lui aussi une durée plus importante au schwa que l'alignement manuel (61 ms vs. 52 ms (sd = 18), $t(2\,808) = 26.2$, $p < 0.0001$). Le système IRISA monophones se distingue en attribuant au schwa une durée inférieure à celle de l'alignement manuel (49 ms vs. 52 ms (sd = 18, $t(2\,744) = 12.9$, $p < 0.0001$).

Système d'alignement	x(s)	% diff. > 10 ms	% diff. > 20 ms	Écarts positifs	Écarts négatifs
LIA (n = 2 608)	55 (17)	34,4 %	9,8 %	21 9 % > 10 6,2 % > 20	12,5 % > 10 3,6 % > 20
IRISA mono. (n = 2 745)	49 (19)	32,8 %	9,6 %	8,5 % > 10 1,7 % > 20	24 3 % > 10 7,9 % > 20
IRISA tri. (n= 2 809)	61 (21)	44,2 %	17,9 %	37,7 % > 10 15,9 % > 20	6,5 % > 10 2 % > 20

Tableau 8 : durées moyennes (et écarts-types), pourcentages d'écarts supérieurs à 10 et à 20 ms et pourcentages d'écarts positifs et négatifs pour chacun des alignements automatiques

Nous avons souhaité évaluer dans quelle mesure ces écarts de durée étaient influencés par la nature des consonnes entourant le schwa. Pour ce faire, les écarts ont été classés en trois catégories : durée similaire (à +/– 10 ms) à la durée de référence (= catégorie « correct »), surestimation (> 10 ms) et sous-estimation (< 10 ms). La Figure 3 ci-dessous présente le taux d'occurrences dans chaque catégorie pour les trois systèmes et le Tableau 9 (Büerki et Gendrot, 2007 ; Bürki et al., 2008) résume les propriétés des consonnes ayant une influence sur les taux de surestimation et de sous-estimation. Un « + » signifie que les surestimations sont favorisées par ce contexte, un « – » que les sous-estimations sont favorisées par ce contexte.

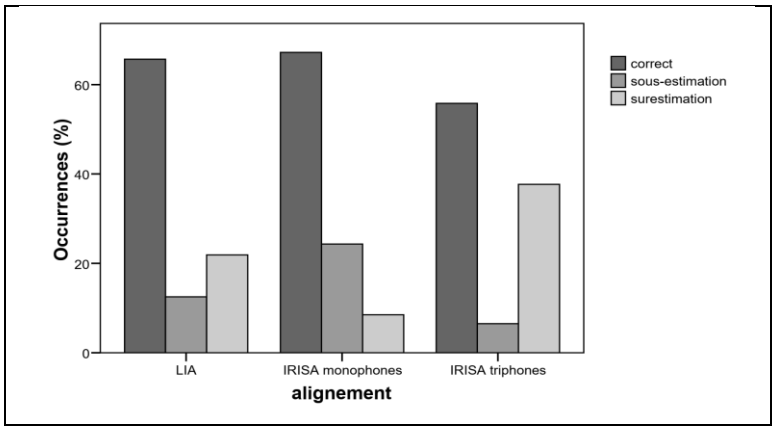


Figure 3 : pourcentage de surestimations, sous-estimations et estimations correctes de la durée pour les trois systèmes d'alignement évalués, d'après Bürki et al. (2008)

	Sonorité gauche			Sonorité droite			Mode gauche				Mode droit			
	Sou.	Son.	V.	Sou.	Son.	V.	F	O	N	L	F	O	N	L
LIA	+	–				+	+			–				
IRISA mono.	+	–			+		+							+
IRISA tri.	+	–				–	+			–	–			

Tableau 9 : résumé des propriétés consonantiques ayant une influence sur la catégorie d'estimation de la durée (Sou. = sourde, Son. = sonante, V = voisée, F = fricative, O = occlusive, N = nasale, L = liquide)

Comme précédemment, les alignements des systèmes du LIA et de l'IRISA triphones ont un profil similaire avec un faible nombre de sous-estimations par rapport aux surestimations. L'alignement de l'IRISA monophones se démarque, quant à lui, par un nombre très important de sous-estimations. En résumé, nous constatons que les trois systèmes sont globalement similaires en ce qui concerne l'influence du type de consonne précédant le schwa sur les taux et types d'erreurs. En revanche, les erreurs des trois systèmes diffèrent dans leurs relations au type de consonne se trouvant après la voyelle.

Une analyse du placement des frontières a été entreprise afin de déterminer si les frontières sont décalées de manière égale en début et en fin de voyelle pour chaque système et de comparer ces derniers. La Figure 4 ci-dessous présente, pour chaque système d'alignement, le pourcentage des écarts à gauche et à droite, supérieurs à 10 ms et à 20 ms, au regard de l'alignement manuel de référence. Le pourcentage global d'écarts en début de voyelle est élevé pour l'alignement IRISA triphones : 50 % d'écarts supérieurs à 10 ms et 14 % d'écarts supérieurs à 20 ms, comparativement aux performances des deux autres systèmes (LIA : 30 % > 10 ms et 6 % > 20 ms, IRISA monophones : 23 % > 10 ms et 4 % > 20 ms). Pour chacun des systèmes, les écarts sont plus nombreux vers la gauche que vers la droite, la frontière de la voyelle est donc placée plus précocement.

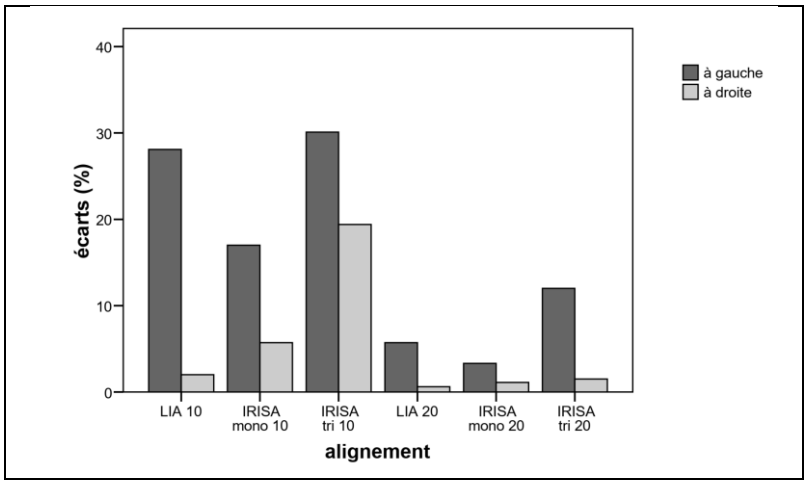


Figure 4 : pourcentage des écarts à gauche et à droite supérieurs à 10 et à 20 ms en début de voyelle pour chacun des alignements, d'après Bürki et al. (2008)

La Figure 5 ci-dessous présente les pourcentages de frontières situées dans un intervalle supérieur à 10 et à 20 ms, à gauche et à droite en fin de voyelle. Les taux d'écarts globaux en fin de voyelle sont plus bas qu'en début de voyelle pour l'alignement de l'IRISA triphones (32 % > 10 ms, 8 % > 20 ms). Dans l'alignement de l'IRISA monophones, la configuration inverse est observée, les écarts étant plus nombreux en fin de voyelle (31 % > 10 ms, 9 % > 20 ms). En ce qui concerne l'alignement du LIA, le pourcentage d'écarts en fin de voyelle est similaire à celui observé en début de voyelle (30 % > 10 ms, 6 % > 20 ms). L'alignement du LIA et celui de l'IRISA monophones ont un profil similaire, avec davantage d'écarts vers la gauche que vers la droite. L'alignement « triphones » présente lui davantage d'écarts vers la droite que vers la gauche. La plus grande durée attribuée aux voyelles par IRISA triphones s'explique donc par un décalage des frontières de début à gauche et de fin à droite. Les alignements du LIA et de l'IRISA monophones ont le même profil, mais la quantité relative des écarts à gauche amène pour l'un à une durée surestimée, pour l'autre à une durée sous-estimée.

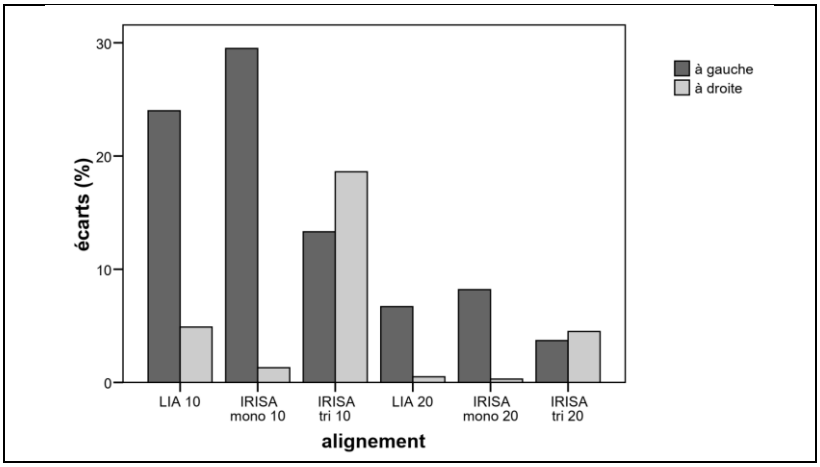


Figure 5 : pourcentage des écarts à gauche et à droite supérieurs à 10 et à 20 ms en fin de voyelle pour chacun des alignements, d'après Bürki et al. (2008)

4.1.3.5 Résumé de l'évaluation des systèmes automatiques

Concernant la capacité des systèmes d'alignement automatique évalués à décider de la présence/absence d'un phone, nos données montrent que les systèmes considérés ne sont pas égaux sur ce point. L'alignement effectué par le système de l'IRISA triphones est le plus proche de l'alignement manuel de référence, suivi par celui réalisé par son homologue « monophones ». La comparaison de deux systèmes similaires sur tous les points excepté le type de modèles acoustiques, nous permet d'évaluer la contribution de la nature de ces derniers dans l'adéquation de la détection du schwa. L'utilisation de triphones (modèles de phones dépendants du contexte) plutôt que de monophones en reconnaissance des mots est largement répandue étant donné les meilleures performances qu'elle permet d'obtenir. Le lien entre taux de reconnaissance et détection de phones est relativement direct, une bonne capacité à détecter les phones présents et à ne pas en insérer est cruciale pour éviter les erreurs de reconnaissance. De manière générale cependant, les trois systèmes témoignent de performances, en termes de détection de phones, tout à fait acceptables avec des pourcentages de désaccord impliquant insertions et effacements allant de 7 à 10 %.

L'emplacement des frontières en particulier peut s'avérer fondamental suivant les analyses envisagées. L'une des conséquences d'une détermination inadéquate de l'emplacement des frontières du phone peut être une mauvaise évaluation de la durée de ce dernier. Lorsque des modèles de phones dépendants du contexte sont utilisés, les frontières entre les phones sont plus « floues » : il est difficile de savoir quelle partie du phone va être considérée comme telle ou comme partie du contexte lors de l'apprentissage. On peut dès lors s'attendre à davantage d'erreurs d'estimation de la durée.

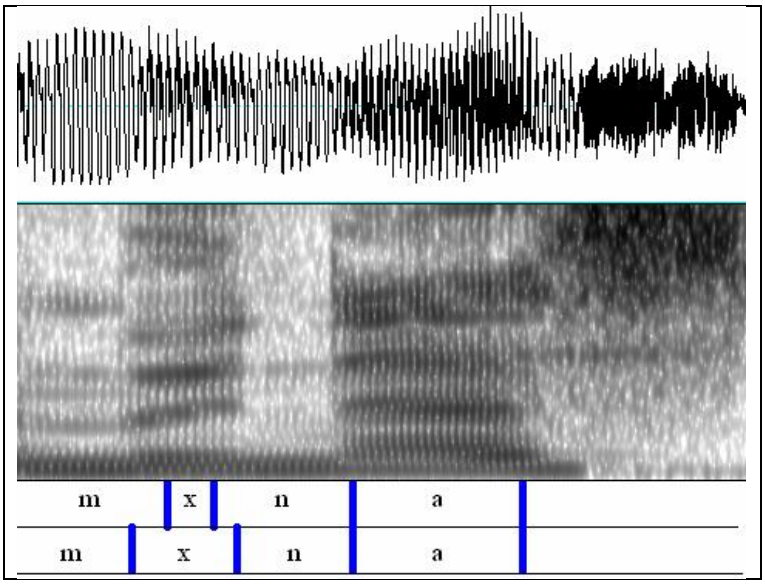


Figure 6 : exemple de différences dans le placement des frontières du schwa (« x ») entre l'alignement automatique effectué par le système « triphones » de l'IRISA (en haut) et l'alignement manuel (en bas) pour une occurrence du mot « menace », d'après Bürki et al. (2008)

S'il s'agit d'étudier les caractéristiques temporelles de la voyelle, les données issues d'un alignement automatique devront être considérées avec prudence. Nous avons vu en effet que les durées estimées automatiquement diffèrent parfois fortement des durées segmentées manuellement et qu'elles sont influencées par les consonnes suivantes et précédentes. L'étude de l'influence du contexte segmental sur la durée des voyelles en particulier risque d'être fortement biaisée si elle s'appuie sur un alignement non vérifié manuellement. Par ailleurs, il s'agit de garder à l'esprit les limitations imposées par le système à la durée d'un segment, que cette limite soit ou non imposée par les modèles de phones. Dans les alignements automatiques évalués ici, la voyelle ne se voit jamais attribuer une durée

inférieure à 30 ms, or la durée minimale attribuée au schwa par l'alignement manuel est de 8 ms. En ce qui concerne l'impact des divergences temporelles entre l'alignement manuel et l'alignement automatique sur des analyses formantiques, nous avons entrepris d'autres analyses (Adda-Decker et al., 2008).

S'il est souvent admis que les imprécisions d'alignement peuvent être compensées par une grande quantité de données analysées, certaines précautions méthodologiques s'avèrent utiles malgré tout. Par exemple, une analyse acoustique des voyelles dans la partie médiane (du premier au dernier tiers) restera assez peu sensible aux imprécisions de la segmentation, mais il n'en sera pas de même pour une analyse visant à analyser des voyelles plus courtes ou des parties spécifiques des voyelles (transition consonne-voyelle ou voyelle-consonne, par exemple).

4.1.3.6 Evaluation des mesures acoustiques automatiques

Afin de mesurer la variabilité des mesures acoustiques obtenues en fonction du système d'alignement, nous avons utilisé les segmentations produites par différents systèmes d'alignements, les trois systèmes précédents auquel nous avons ajouté celui du LIMSI (Adda-Decker et al., 2008). Les mesures présentées ici sont la durée des voyelles et leurs valeurs de formants. La Figure 7 établie à partir de ces différents alignements montre des variations de durée en fonction de la position de la syllabe, dans des mots trisyllabiques (à gauche) et quadrisyllabiques (à droite). Concernant les mots trisyllabiques, il est possible d'observer pour tous les systèmes que la voyelle de la syllabe initiale est un peu plus courte que celle de la syllabe intermédiaire, les deux étant beaucoup plus courtes que celle de la syllabe finale (voir Vaissière (2010) pour une revue globale). Ces tendances sont identiques pour tous les systèmes, bien que les mesures brutes varient légèrement d'un système à l'autre. Si l'on analyse chaque voyelle séparément, il est possible alors de mettre en évidence quelques différences, notamment pour les voyelles /ø/ et /œ/, induites quant à elles par des différences du dictionnaire de prononciation. Pour les mots quadrisyllabiques, la syllabe initiale, fréquemment porteuse d'un accent initial (que cet accent soit purement rythmique ou qualifié de « journalistique »), est légèrement plus longue que les syllabes en deuxième position (la tendance est moins nette pour le système IRISA_monophones malgré tout). La forte cohérence entre les différents systèmes considérés permet également d'accroître la confiance apportée vis-à-vis des observations présentées ici.

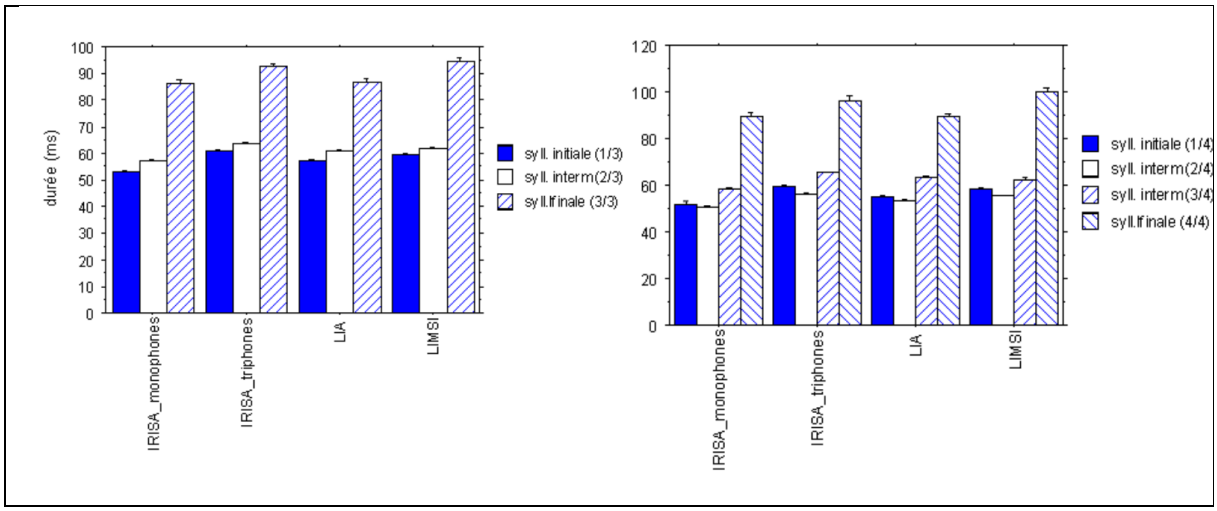
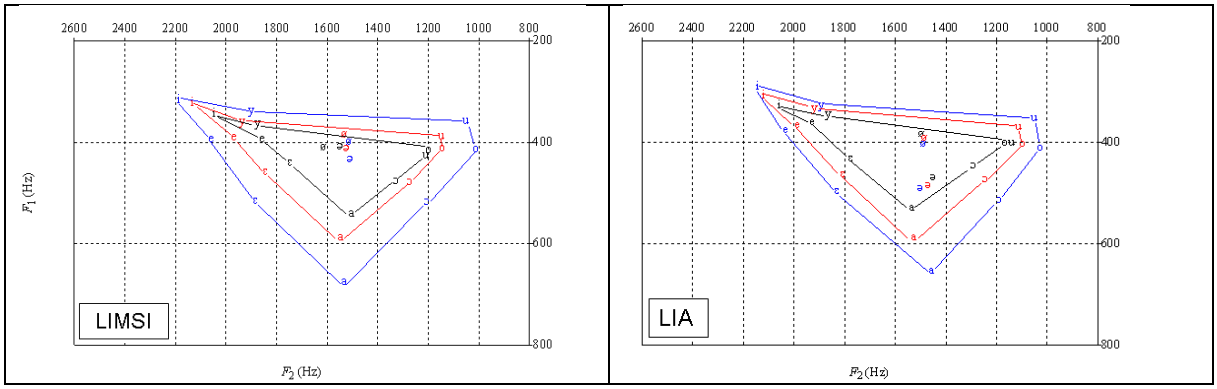


Figure 7 : variations de durée en fonction de la position de la syllabe pour différents systèmes d'alignement (de gauche à droite : IRISA monophones, IRISA triphones, LIA, LIMSI), d'après Adda-Decker et al. (2008)

Pour l'analyse phonémique, et plus particulièrement l'analyse des voyelles orales, les mesures qui nous intéressent sont les mesures formantiques. Une vérification auprès des résultats déjà mentionnés dans la littérature ou bien après une vérification manuelle sur un échantillon de ces mêmes données s'avère précieuse. Une solution peut consister à rejeter des mesures aberrantes par l'établissement de filtres pour les valeurs de formants ou de f0 établis sur la base de connaissances acoustiques ou bien de vérifications visuelles, telles qu'effectuées par Gendrot et Adda-Decker (2004) et Woehrling et Mareüil (2007). Les mesures automatiques fournies par le biais de logiciels libres tels que PRAAT (http ://www.fon.hum.uva.nl/praat) ou openSMILE (https://www.audeering.com/opensmile/) sont de plus en plus performantes ; les erreurs de mesures ne sont pas dues au hasard, elles peuvent être justifiées et révèlent des phénomènes intéressants. Par exemple, pour les formants, le /i/ du français est une voyelle mieux perçue par le rapprochement des 3ème et 4ème formants, ce qui peut contribuer à favoriser la non détection du 2ème formant. Il en va de même pour les deux premiers formants de /u/, ou bien pour les mesures de f0 sur de la voix craquée, fréquente en parole continue. Dans le cas précis de la mesure des formants de /u/, les auteurs de PRAAT suggèrent, pour une meilleure détection de ces formants, de modifier légèrement les paramètres d'analyse en abaissant la limite supérieure du seuil de détection des cinq premiers formants. Cette procédure a permis sur des données de parole radiophonique de réduire les taux d'erreurs de détection de 45 % à 19 % pour /u/ (Gendrot et Adda-Decker, 2005).

Pour les quatre figures ci-après (voir Figure 8), représentant quatre types d'alignements différents, nous présentons les triangles vocaliques des voyelles orales du français en fonction de leur durée. Les mesures de formants, prises entre 1/3 et 2/3 de la voyelle, même si elles varient légèrement en fonction des différents systèmes d'alignement, montrent un même comportement centripète en fonction d'une durée segmentale décroissante. Les ellipses de variation (non affichées ici) y sont semblables aussi. Ce comportement stable à travers différentes configurations montre que la précision de segmentation n'est pas en cause pour ce type de mesures impliquant la partie centrale des segments. La seule figure qui diffère sensiblement des autres est celle construite à partir de la segmentation « IRISA_triphones », dont nous avons pu montrer que la précision était moins importante. Nous observons également des différences pour les voyelles centrales, mais cela est dépendant du dictionnaire de prononciation, et de la façon dont sont traitées les hésitations par exemple. Les taux de rejets sont également analysés et indiquent des valeurs semblables : LIA (5.5 %), IRISA_monophones (4.8 %), IRISA_triphones (5.5 %) et LIMSI (4.1 %).



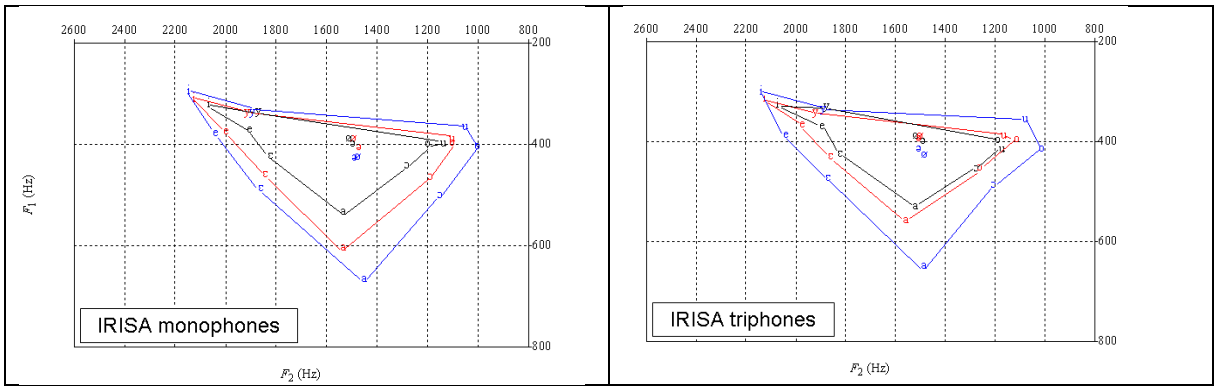


Figure 8 : valeurs moyennes de F1 et F2 des voyelles orales du français en fonction de la segmentation utilisée et de la durée vocalique. De l'intérieur à l'extérieur, segments de durée (en ms) : [30 - 50], [60 - 80], [90 - 110], d'après Adda-Decker et al. (2008)

Il est évident qu'à l'heure où ces lignes sont écrites, de nombreux travaux ont permis de reconsidérer tant les alignements que les analyses automatiques sous un regard nouveau, que nous présenterons en discussion générale dans la section 4.4. Si les outils ont pu évoluer pour aller vers des mesures plus fiables et rapides, les résultats et précautions abordés ici restent des principes à respecter aujourd'hui (Adda-Decker et al., 2008).

4.1.4 Premières statistiques descriptives sur les valeurs formantiques du corpus ESTER

Le Tableau 10 et le Tableau 11 nous fournissent les valeurs moyennes des quatre premiers formants ainsi que leurs écarts-types après filtrage pour les hommes et pour les femmes. Il est nécessaire de se rappeler ici que les valeurs moyennes ne fournissent qu'une idée très imprécise de ce que les valeurs de formants peuvent inclure de variations puisque les valeurs de formants dépendent non seulement du locuteur, mais également du contexte d'articulation et leur durée. Ces premières analyses ont donc naturellement été approfondies en considérant ces deux derniers paramètres comme facteurs de variations.

	i	y	e	ɛ	a	œ	ø	ɔ	o	u
F1	310	336	370	438	557	400	384	456	397	371
st. dev.	74	120	54	68	97	80	57	77	56	78
F2	2005	1803	1850	1717	1444	1445	1474	1203	1041	1105
st. dev.	188	135	169	162	178	186	166	196	208	213
F3	2784	2425	2545	2490	2438	2440	2405	2420	2477	2470
st. dev.	193	178	156	157	167	163	141	164	195	179
F4	3492	3270	3435	3406	3427	3304	3213	3306	3371	3394
st. dev.	247	200	254	270	262	230	190	213	222	235

Tableau 10 : valeurs moyennes des quatre premiers formants en fonction de la voyelle pour les hommes ; st. dev: écart-type, d'après Gendrot et Adda-Decker (2010)

	i	y	e	ɛ	a	œ	ø	ɔ	o	u
F1	348	371	423	526	685	436	420	528	438	404
st. dev.	83	126	75	106	125	94	65	110	75	90

F2	2365	2063	2176	2016	1677	1643	1693	1347	1140	1153
st. dev.	186	187	181	181	223	233	159	217	203	188
F3	3130	2745	2860	2800	2735	2715	2687	2743	2790	2742
st. dev.	238	210	175	182	200	170	150	164	205	206
F4	4159	3786	4039	3986	3937	3856	3850	3929	3853	3872
st. dev.	300	251	350	392	368	302	277	280	255	294

Tableau 11 : valeurs moyennes des quatre premiers formants en fonction de la voyelle pour les femmes ; st. dev: écart-type, d'après Gendrot et Adda-Decker (2010)

La différence entre les formants mesurés pour les hommes et pour les femmes s'avère moins importante (non illustrée ici) pour les voyelles fermées que pour les voyelles ouvertes comme souvent mentionné dans la littérature, par la présence de résonances de Helmholtz pour ces voyelles (F1 pour /i/, /y/ et /u/, et également F2 pour /u/). Les valeurs de référence mentionnées jusqu'alors pour le français (Calliope, 1989) et représentées dans la Figure 9 y apparaissent nettement plus extrêmes sur l'axe F1-F2, et particulièrement pour les voyelles postérieures, de par le contexte uvulaire utilisé pour établir ces données.

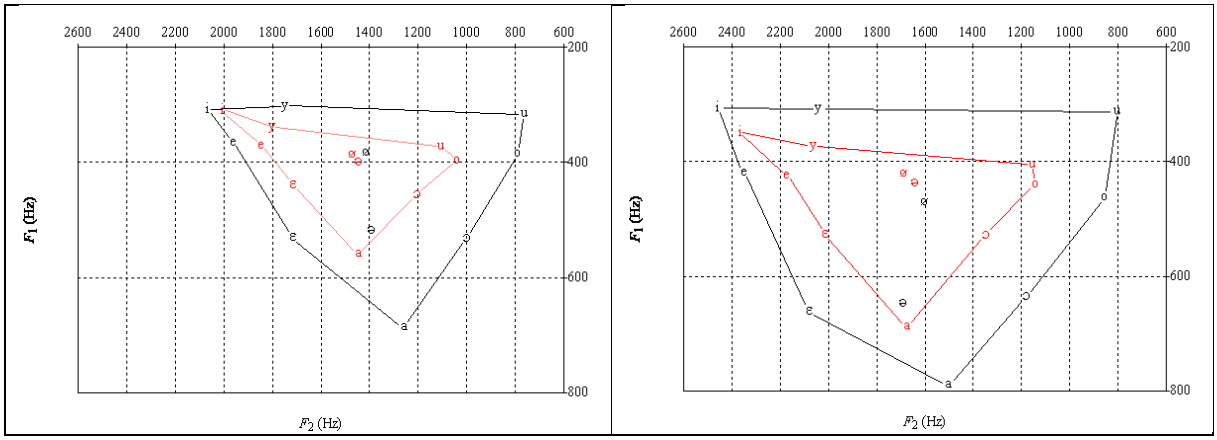


Figure 9 : valeurs moyennes des formants F1-F2 pour les hommes (gauche) et pour les femmes (droite). En rouge (à l'intérieur), les valeurs moyennes pour nos données, et en noir (extérieur), les valeurs fournies par Calliope, d'après Gendrot et Adda-Decker (2010)

4.1.4.1 Réduction acoustique des voyelles en fonction de la durée vocalique

Le premier facteur de variation que nous avons analysé ensuite a été la durée représentée par les trois catégories [30-50], [60-80], et [90-110] (rappelons que les durées varient selon un pas de 10 ms de par le système d'alignement, donc aucune valeur n'est obtenue entre chaque dizaine) utilisées dans la section 4.1.2. Les résultats sont illustrés par la Figure 10, le Tableau 12 et le Tableau 13. Nous avons observé que l'espace vocalique formé par les formants F1 et F2 diminue progressivement avec la durée des segments analysés ; avec des variations plus faibles pour les voyelles fermées /i/ et /y/ qui sont, comme il est reconnu pour le français, mieux caractérisées par la proximité de F3 avec F2 (/y/) ou F4 (/i/). Pour vérifier ce point statistiquement, une ANOVA à deux facteurs a été effectuée, les deux facteurs étant le « phonème » et la catégorie de « durée » (court – moyen – long). Les résultats montrent que la durée des voyelles a un effet significatif sur les valeurs de F1 [$F(49.25)\ p<0.0001$] et F2 [$F(9.85)\ p<0.0001$]. Cela peut être interprété comme des cibles non atteintes pour les voyelles plus courtes alors que les formants des voyelles longues sont plus proches des valeurs obtenues pour de la

parole lue par exemple. A titre d'information, les valeurs obtenues par Calliope restent plus extrêmes que les valeurs des voyelles longues de notre corpus.

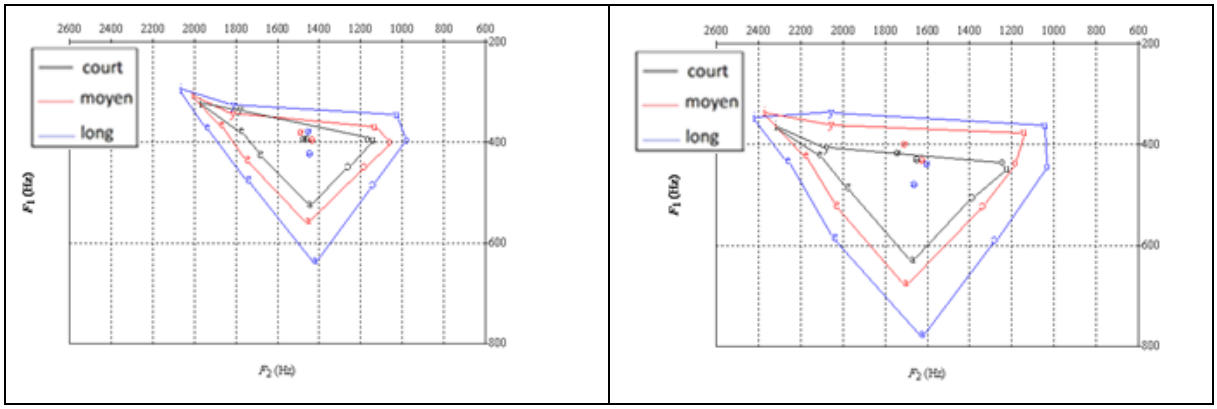


Figure 10 : valeurs moyennes mesurées pour F1 et F2 en fonction de la catégorie de durée. Par ordre ascendant de l'intérieur vers l'extérieur (noir [30 – 50 ms], rouge [60 - 80], bleu [90 - 110]). Les hommes sont représentés à gauche et les femmes à droite, d'après Gendrot et Adda-Decker (2010)

Femmes	i	y	e	ɛ	a	œ	ø	ɔ	o	u
$F1_{moy}$ [0.03-0.05]	363 (102)	407 (143)	420 (76)	485 (104)	627 (120)	430 (98)	415 (71)	506 (114)	434 (86)	449 (104)
$F1_{moy}$ [0.06-0.08]	336 (71)	361 (136)	421 (71)	520 (91)	674 (106)	432 (78)	400 (49)	522 (102)	436 (76)	377 (63)
$F1_{moy}$ [0.09-0.11]	347 (84)	336 (79)	432 (79)	585 (108)	776 (114)	480 (113)	440 (63)	590 (100)	444 (66)	363 (58)
$F2_{moy}$ [0.03-0.05]	2315 (200)	2074 (230)	2108 (187)	1977 (207)	1674 (256)	1650 (266)	1747 (178)	1385 (235)	1245 (203)	1225 (153)
$F2_{moy}$ [0.06-0.08]	2368 (178)	2057 (170)	2175 (170)	2030 (168)	1705 (205)	1625 (193)	1711 (121)	1340 (215)	1183 (175)	1141 (189)
$F2_{moy}$ [0.09-0.11]	2416 (171)	2056 (152)	2262 (152)	2041 (159)	1627 (200)	1663 (174)	1605 (113)	1281 (154)	1032 (178)	1043 (188)
$F3_{moy}$ [0.03-0.05]	3088 (257)	2793 (224)	2836 (180)	2798 (184)	2731 (194)	2731 (186)	2678 (160)	2724 (162)	2720 (211)	2744 (213)
$F3_{moy}$ [0.06-0.08]	3123 (237)	2735 (198)	2865 (165)	2807 (164)	2735 (186)	2700 (146)	2700 (145)	2750 (173)	2751 (187)	2723 (188)
$F3_{moy}$ [0.09-0.11]	3177 (209)	2683 (193)	2883 (180)	2794 (204)	2733 (227)	2681 (147)	2692 (139)	2768 (143)	2872 (190)	2767 (216)
$F4_{moy}$ [0.03-0.05]	4121 (313)	3861 (286)	4040 (348)	4012 (358)	3988 (355)	3868 (320)	3834 (298)	3908 (302)	3819 (275)	3894 (306)
$F4_{moy}$ [0.06-0.08]	4187 (288)	3748 (210)	4053 (333)	4017 (370)	3939 (357)	3835 (263)	3852 (212)	3933 (264)	3817 (268)	3812 (277)
$F4_{moy}$ [0.09-0.11]	4159 (298)	3748 (244)	4013 (378)	3907 (449)	3870 (394)	3863 (329)	3872 (294)	3969 (261)	3910 (218)	3921 (284)

Tableau 12 : valeurs moyennes de F1 et de F2 pour les femmes en fonction des intervalles de durée (entre parenthèses les écarts-types), d'après Gendrot et Adda-Decker (2010)

Hommes	i	y	e	ɛ	a	œ	ø	ɔ	o	u
$F1_{moy}$ [0.03-0.05]	323 (82)	334 (100)	376 (57)	424 (68)	523 (88)	393 (95)	394 (67)	449 (81)	394 (61)	396 (89)
$F1_{moy}$ [0.06-0.08]	306 (74)	344 (125)	365 (53)	434 (62)	555 (85)	395 (60)	380 (56)	449 (72)	397 (57)	369 (73)
$F1_{moy}$ [0.09-0.11]	295 (53)	327 (131)	369 (51)	475 (76)	635 (95)	422 (70)	377 (43)	485 (77)	396 (55)	346 (63)

$F2_{moy}$ [0.03-0.05]	1968 (183)	1780 (147)	1774 (165)	1685 (161)	1445 (208)	1455 (215)	1475 (208)	1262 (205)	1168 (205)	1140 (209)
$F2_{moy}$ [0.06-0.08]	2005 (182)	1815 (128)	1871 (151)	1747 (158)	1453 (163)	1429 (154)	1488 (157)	1182 (191)	1062 (183)	1134 (198)
$F2_{moy}$ [0.09-0.11]	2060 (192)	1811 (131)	1939 (148)	1740 (159)	1418 (134)	1442 (143)	1452 (125)	1140 (162)	979 (206)	1028 (215)
$F3_{moy}$ [0.03-0.05]	2720 (205)	2442 (191)	2522 (160)	2472 (162)	2437 (174)	2450 (176)	2403 (160)	2395 (164)	2390 (157)	2477 (171)
$F3_{moy}$ [0.06-0.08]	2800 (172)	2438 (161)	2546 (151)	2506 (122)	2431 (163)	2438 (146)	2423 (137)	2430 (166)	2468 (195)	2467 (174)
$F3_{moy}$ [0.09-0.11]	2859 (174)	2389 (184)	2583 (150)	2509 (152)	2454 (159)	2419 (150)	2381 (120)	2438 (160)	2514 (198)	2465 (194)
$F4_{moy}$ [0.03-0.05]	3454 (257)	3243 (222)	3414 (250)	3386 (264)	3402 (262)	3334 (241)	3220 (203)	3291 (220)	3294 (205)	3371 (222)
$F4_{moy}$ [0.06-0.08]	3507 (228)	3288 (181)	3450 (252)	3415 (289)	3444 (261)	3287 (214)	3200 (177)	3308 (217)	3370 (219)	3372 (219)
$F4_{moy}$ [0.09-0.11]	3524 (255)	3271 (199)	3446 (261)	3437 (250)	3447 (258)	3245 (213)	3227 (199)	3328 (193)	3400 (224)	3448 (260)

Tableau 13 : valeurs moyennes de F1 et de F2 pour les hommes en fonction des intervalles de durée (entre parenthèses les écarts-types), d'après Gendrot et Adda-Decker (2010)

Les valeurs présentées jusqu'ici ont été fournies en Hertz, ce qui indiquait que les valeurs n'étaient pas normalisées en fonction du sexe ou de l'identité du locuteur. Une normalisation par locuteur permet notamment de vérifier si les résultats ne sont pas dus à la sur-représentation de certains locuteurs dans un corpus (Adank, 2003 ; voir également Gendrot (2013) pour une revue de la littérature). Comme montré par les figures ci-dessous, les normalisations ne modifient pas les résultats observés pour les valeurs de formants selon la catégorie de durée des voyelles. Nous émettons l'hypothèse que pour des analyses aussi générales, la structure du système est invariante lorsqu'elles sont en quantité aussi importante et en distribution naturelle.

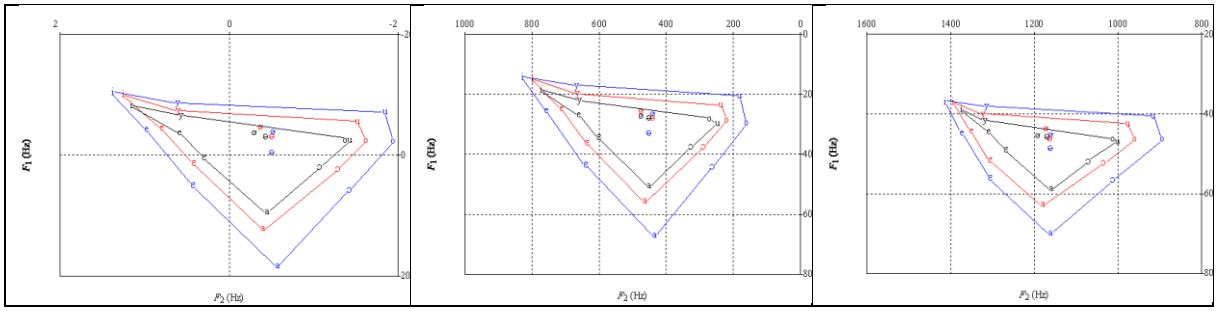


Figure 11 : valeurs moyennes de formants F1 et F2 selon les trois catégories de durée illustrées précédemment, en fonction de la normalisation utilisée. De gauche à droite : normalisation de Lobanov, Gerstmann et Bark, , d'après Gendrot (2010)

4.1.4.2 Contextes : contexte phonémique et morphologique

Après la présentation des formants des voyelles orales en fonction des catégories de durée, nous avons poursuivi nos investigations (*Gendrot et Adda-Decker, 2010)* avec le deuxième facteur de variation linguistique présenté en début de section : le contexte phonémique. La non-atteinte de cible observée sur la Figure 10 donne une première impression de centralisation vers l'intérieur du triangle vocalique,

puisque les voyelles périphériques se décalent vers une position centrale lorsqu'elles sont plus courtes. Mais le contexte segmental est un paramètre important dans les variations des caractéristiques spectrales des voyelles comme le montre la Figure 12. Cette figure montre principalement une différence au niveau des voyelles postérieures et centrales, avec un contexte dental qui voit augmenter les valeurs de F2 comparativement au contexte labial (et aux autres contextes bien que non illustrés ici). Cette augmentation du deuxième formant des voyelles postérieures est généralement interprétée comme une antériorisation, même si elle peut être combinée à d'autres causes (raccourcissement du conduit vocal dû à un arrondissement moindre ou à une remontée du larynx par exemple). Notons que la valeur moyenne des formants est ici fortement influencée par un contexte alvéolaire majoritaire.

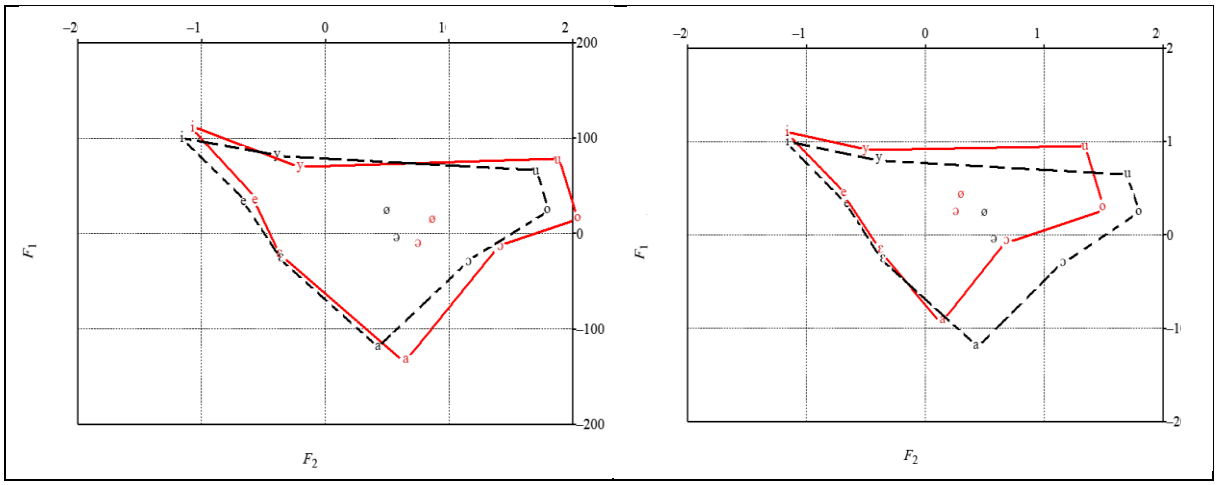


Figure 12 : valeurs de formants en contexte gauche-droite labial (gauche) et alvéolaire (droite) compares aux résultats moyens (en pointillé), d'après Gendrot et Adda-Decker (2010)

Si l'on considère que les voyelles les plus courtes sont plus sujettes à la coarticulation (Lindblom, 1963), c'est-à-dire prenant les caractéristiques spectrales du segment immédiatement à gauche et à droite de la voyelle, dans ce cas les voyelles les plus courtes seraient plus coarticulées et auraient une transition du deuxième formant pointant vers le locus correspondant au lieu d'articulation de la consonne avoisinantes (approximativement 600 Hz pour les consonnes labiales, 1800 Hz pour les consonnes dentales et 3000 Hz pour les consonnes palatales (Delattre et al., 1955). Pour la majorité des voyelles et des contextes, cette tendance est cohérente avec la centralisation observée sur la Figure 10. Pour quelques contextes spécifiques tels que /a/ en contexte labial, le /a/ court ne se décalera pas vers une position centrale de l'espace acoustique puisque le deuxième formant s'abaissera comparativement à sa cible acoustique (1444 Hz pour les hommes et 1677 Hz pour les femmes, i.e. les valeurs des voyelles longues indiquées dans le Tableau 10 et le Tableau 11). Au contraire, comme nous pouvons l'observer sur la **Erreur ! Source du renvoi introuvable.**, le /a/ court dans un contexte labial s'éloigne de l'espace vocalique et donc de son centre.

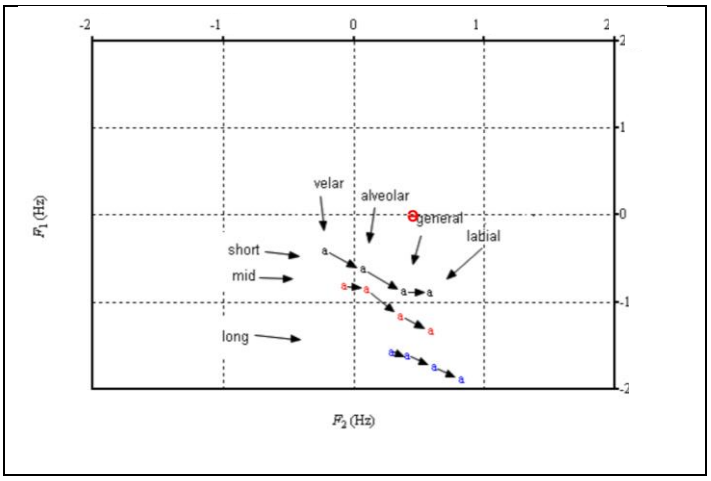


Figure 13 : Valeurs moyennes de F1 et F2 pour le /a/ en fonction de leur durée, et en fonction de leur contexte phonémique gauche-droite (dans l'ordre indiqué par les flèches, vélaire, alvéolaire, tous contextes, et labial) , d'après Gendrot et Adda-Decker (2010)

Cette plus grande coarticulation des voyelles courtes, plutôt qu'une centralisation, est confirmée dans notre corpus, mais les contextes montrant des phénomènes de non centralisation sont plus rares. Dans le cadre d'une comparaison entre l'influence de la durée et du contexte (Gendrot et Adda-Decker, 2010), nous avons pu mettre en évidence un effet cumulé de ces deux prédicteurs sans qu'aucun des deux ne se détache ostensiblement.

Ces résultats (*Gendrot et Adda-Decker, 2010)* s'inscrivent dans la lignée d'une littérature abondante et parfois divergente sur le sujet (Fourakis, 1991 ; Moon et Lindblom, 1994 ; Padgett et Tabain, 2005 ; Van Bergem, 1993), mais qui se basait sur de la parole lue et sur des contextes CVC où les deux consonnes à gauche et à droite de la voyelle sont identiques. Nous souhaitions ici obtenir un état des lieux plus complet et qui puisse mieux prendre en compte les interactions entre les voyelles, leur durée et leur contexte. Dans la parole non contrôlée, il est plus fréquent d'observer des contextes différents à gauche et à droite de la voyelle, et la question d'une éventuelle prévalence d'un contexte consonantique sur un autre se pose en termes d'influence. Rappelons également que les résultats présentés jusqu'ici moyennaient l'ensemble des contextes consonantiques qui par définition n'ont pas une distribution équilibrée comme indiqué dans le Tableau 1. Afin d'obtenir un nombre suffisant de contextes, et ce malgré les 30 heures de corpus étiquetés, nous avons choisi de regrouper les contextes en quatre catégories comme indiqué dans la section 4.1.2 : labial, alvéolaire, palato-vélaire et uvulaire, les phonèmes à l'intérieur de ces regroupements ayant une influence similaire sur les formants des voyelles (Stevens, 1989 ; Vaissière, 1985).

Nous avons pu montrer que le contexte uvulaire était le plus perturbateur, suivi par le contexte alvéolaire, le contexte vélaire, et pour finir le contexte labial (Gendrot et al., 2008). Ceci est résumé dans le Tableau 14, où les cases les plus foncées indiquent un effet coarticulant plus fort par opposition aux cases plus claires. Les locus des consonnes (Sussman et al., 1991) sont indiqués comme le point moyen vers lequel convergent les formants de chaque voyelle en fonction du contexte. Pour les cas où les consonnes environnantes ont un locus attirant le formant dans des directions contradictoires, comme dans la séquence P-a-T par exemple où le contexte labial a un locus en dessous du F2 de la voyelle /a/ alors que le contexte labial a un locus supérieur au F2, nous observons que le contexte le plus coarticulant va influencer la position du formant sur deux tiers de la durée de la voyelle environ, ce qui génère une augmentation moyenne du deuxième formant. Nous avons également pu observer que le locus des consonnes labiales est généralement plus élevé (~1200 Hz) que celui indiqué par

Delattre (<1000 Hz) sauf quand le contexte est exclusivement labial. Par contre, ce tableau ne permet de conclure quant à l'aspect anticipatoire ou rétrograde de la coarticulation en français.

	- P	- T	- K	- R
P-	F1: 400-500 F2: <1000	F1: 200-300 F2: 1800	F1: 300-400 F2: >2000 front v 1000 back v	F1: 800 F2: <1000
T-	F1: 200-300 F2: > 2000	F1: 200-300 F2: 1800-2000	F1: 200-300 F2: > 2000	F1: 800 F2: 1100-1200
K-	F1: 300-400 F2: 2000 front v 1000 /u/,/o/	F1: 300 F2: > 2000	F1: 300 F2: >2000 front v 1000 back v	F1: 800 F2: 1600 front v <1000 back v
R-	F1: 800 (but /u/) F2: <1000	F1: 800 (but /u/) F2: <1000	F1: 800 (but /u/) F2: <1000	F1: 800 (but /u/) F2: <1000

Tableau 14 : direction des formants des voyelles en fonction des contextes consonantiques gauche (de haut en bas) -et droite (de gauche à droite). La couleur des cases résume l'influence du contexte, du moins influent (clair) au plus influent (foncé), d'après Gendrot et Adda-Decker (2010)

4.1.4.3 L'analyse des voyelles quantiques : voyelles antérieures hautes

Les voyelles orales du français -et parmi celles-ci les voyelles /i/ et /y/- sont fréquemment considérées, par exemple par Jones, comme de bons prototypes des voyelles cardinales. /i/ et /y/ notamment sont mentionnées comme des représentantes idéales du processus de focalisation (Schwartz et al., 1993, 1997) puisque caractérisées par le rapprochement de deux de leurs formants (F3/F4 et F2/F3 respectivement). Dans ces exemples, le rapprochement de deux formants a trois conséquences :

- la création d'un pic spectral proéminent dans la région fréquentielle regroupant les deux formants (l'augmentation de l'amplitude de deux formants proches est de 6 dB pour une distance deux fois plus petite entre les deux formants (Fant, 1970).

- les formants proches sont également intégrés d'un point de vue perceptif en un pic simple (Chistovich et al., 1979).

- le deuxième formant effectif (F'2 ou 'F2prime'), prenant en compte la fréquence des formants supérieurs, lors d'une synthèse des voyelles cardinales à deux formants semble correspondre à la proéminence spectrale créée par le rapprochement de ces deux formants (Bladon et Fant, 1978).

Tabain et Perrier (2005) ont également montré au moyen d'une étude analysant en parallèle les réalisations articulatoires et acoustiques du /i/ en position pré-finale de constituant prosodique que les variations du troisième formant étaient plus larges que celles des deux premiers formants, les auteurs émettant l'hypothèse que le locuteur tente d'atteindre des cibles acoustiques plutôt qu'articulatoires. L'utilisation de F'2 dans nos précédentes études (Gendrot et Adda-Decker, 2007) n'ayant pas mis en évidence des variations plus larges, nous émettons l'hypothèse que nous pourrons observer des mouvements significatifs et plus importants des formants supérieurs F3 et F4. Notons également que l'opposition entre ces deux voyelles se fait essentiellement sur la base de l'arrondissement, un arrondissement plus important qui abaisse la hauteur du troisième formant, il est ainsi probable que l'allongement des voyelles /i/ et /y/ favoriserait un rapprochement des deux formants concernés pour les voyelles les plus longues.

Dans cette troisième sous-section, nous avons analysé (Gendrot et al., 2008) dans quelle mesure les fréquences des formants F3 et F4 sont sujettes à la variation en fonction de la durée vocalique. La motivation pour ces analyses des formants supérieurs tenait au fait que pour le français, l'arrondissement est un trait phonologique distinctif pour les voyelles antérieures. De même, nous avions observé que pour les voyelles /i/ et /y/ par exemple, les variations liées à la durée étaient réduites par rapport aux autres voyelles. La Figure 14 représente les déplacements des formants F2 vs. F3 (à gauche) et F3 vs. F4 (à droite) pour chaque catégorie de durée reliées entre elles par des flèches allant de la catégorie la plus courte à la catégorie la plus longue. Lorsque la durée mesurée est plus grande, F3 augmente considérablement pour la voyelle /i/, et c'est le mouvement le plus large observé parmi toutes les voyelles. Sur cette même figure, les variations de la voyelle /y/ sont également intéressantes puisque l'on peut observer qu'elles vont dans une direction opposée à celle des autres voyelles, i.e. les valeurs baissent quand la durée augmente. Le /e/ bien que proche acoustiquement du /y/ et particulièrement du /i/ n'est pas une voyelle focale et n'est pas caractérisé par ce rapprochement de formants.

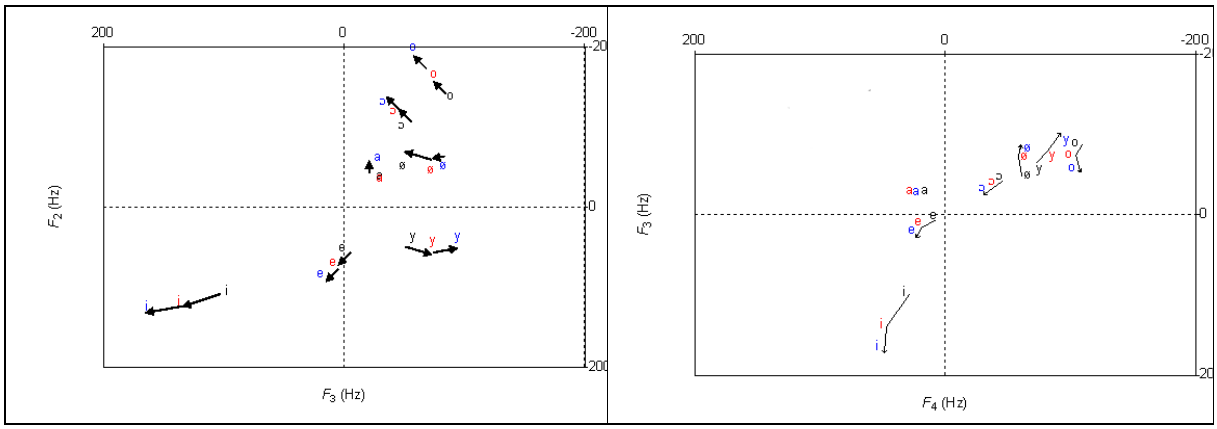


Figure 14 : valeurs moyennes de F2 vs. F3 (à gauche) et F3 vs. F4 pour les voyelles orales du français en fonction de leur durée répartie en trois catégories de durée (selon le sens de la flèche, en noir [30-50 ms], en rouge [60-80 ms] et en bleu [90-110 ms]). Normalisation de Gerstman, d'après Gendrot et al., 2008

Ces observations valent également pour /ø/ et /œ/, ce qui laisse à penser que l'arrondissement de ces voyelles, qui pousse à une baisse globale des formants, et plus spécifiquement F3, n'est pas étranger à ce phénomène. F3 baisse lorsque la durée augmente, ce qui est contraire à ce que nous observons pour les autres voyelles. Nous l'interprétons ici comme un arrondissement supplémentaire de ces voyelles (/y/, /ø/ et /œ/) lorsqu'elles sont longues, ce qui explique les résultats différents observés pour ces voyelles dans la section 4.1.4.1. Ces voyelles n'observent pas un mouvement vers le centre acoustique aussi que les autres voyelles, car leur variation se retrouve sur le troisième formant essentiellement (Gendrot et al., 2008).

Ces exemples confirment la nécessité de prendre en compte le 3ème formant (au minimum) pour rendre la variabilité des voyelles (françaises) /i/ et /y/, mais également /ø/ et /œ/. Notons que les voyelles postérieures ne sont pas concernées ici, pour deux raisons selon nous : (i) elles n'ont pas de contrepartie étirée en français, et (ii) leurs formants supérieurs sont caractérisés par une amplitude trop faible pour jouer un rôle perceptif. Ces résultats renforcent l'idée que l'hyper-articulation des voyelles longues ne se réalise pas de façon identique pour toutes les voyelles : elle se réalise selon un renforcement des traits distinctifs. Il est habituel de considérer que les voyelles hyper-articulées occupent un espace plus important dans l'espace vocalique en s'éloignant du centre vocalique (proche

du schwa). C'est exact pour les voyelles ouvertes, mais ça l'est moins pour les voyelles fermées /i/,/y/ et /u/ réalisées par une ou plusieurs résonances de Helmholtz, et caractérisées par une focalisation des formants F3/F4, F2/F3 et F1/F2 respectivement. Une hyper-articulation de ces voyelles aboutira à un rapprochement supplémentaire de ces formants focalisés, ce qui pourra générer des problèmes de détection de formants pour les algorithmes puisque les deux formants focalisés sont difficiles à dissocier l'un de l'autre (Gendrot et al., 2008). Notons qu'une fermeture accrue de ces voyelles fermées mènerait inexorablement à une spirantisation (Torreira et Ernestus, 2011).

Par la suite, nous avons comparé ces résultats avec ceux d'autres langues (Gendrot et al., 2007). Dans un premier temps, notre objectif était de vérifier si la distance entre les formants F3 et F4 pour /i/ et F2 et F3 pour /y/ (quand il est présent) est moindre en français que dans sept autres langues que nous avions à notre disposition dans des corpus similaires (allemand, anglais américain, arabe, espagnol, italien, mandarin, portugais ; le système d'alignement utilisé est le même que pour les corpus précédents, de même pour les méthodes d'analyse) (Gendrot et Adda-Decker, 2007). Le cas échéant, ce pourrait être une validation de leur caractère souvent reconnu comme « cardinal ». Nous avons effectué pour ce faire une comparaison avec les sept autres langues en mesurant l'écart entre F3 et F4 pour le /i/. Concernant le /y/ seulement présent en allemand et en mandarin, nous effectuerons une comparaison de l'écart entre F2 et F3. Dans le cas de l'anglais et de l'allemand ayant une distinction entre un /i/ dit « relâché » (/ɪ/) et un /i/ dit «tendu», seules les variantes «tendues», plus proches de celles du français, ont été considérées. Pour le mandarin, les différents tons appliqués à chaque voyelle sont fusionnés.

	F1	F2	F3	F4	F4 – F3
allemand	319 (70)	1991 (222)	2610 (239)	3621 (248)	1012 (269)
anglais	352 (61)	2044 (186)	2503 (199)	3442 (225)	939 (244)
arabe	398 (130)	2102 (169)	2678 (141)	3364 (295)	686 (258)
espagnol	375 (57)	2126 (155)	2784 (149)	3634 (126)	851 (226)
français	**302 (87)**	**2024 (158)**	**2848 (228)**	**3494 (258)**	**646 (230)**
italien	347 (61)	2065 (231)	2693 (236)	3589(400)	895 (301)
mandarin	360 (109)	2132 (358)	2836 (290)	3644 (265)	809 (304)
portugais	344 (67)	1906 (185)	2503 (277)	3576 (277)	1075 (329)

Tableau 15 : valeurs moyennes des formants F1, F2, F3 et F4 pour /i/ incluant leurs écarts-types respectifs entre parenthèses, ainsi que l'écart en Hertz entre F3 et F4, d'après Gendrot et Adda-Decker, 2007

	F1	F2	F3	F3 – F2	F4
allemand	348 (90)	1598 (190)	2357 (197)	759 (229)	3451 (222)

français	**325 (124)**	**1833 (154)**	**2455 (211)**	**622 (205)**	**3271 (195)**
mandarin	348 (65)	2136 (182)	2650 (201)	514 (214)	3507 (145)

Tableau 16 : valeurs moyennes des formants F2 et F3 pour /y/ incluant leurs écarts-types respectifs entre parenthèses, ainsi que l'écart en Hertz entre F2 et F3, d'après Gendrot et Adda-Decker, 2007

Les tableaux ci-dessus indiquent que pour la voyelle /i/, l'écart F4 - F3 est le plus faible en comparaison des sept autres langues analysées. Seul l'écart observé pour l'arabe s'approche de celui du français. Dans une optique pédagogique, il semblerait alors trompeur d'utiliser le même symbole de l'API (par exemple le /i/ anglais, le /i/ allemand et le /i/ français) pour une même voyelle dans différentes langues puisqu'ils n'ont pas les mêmes caractéristiques spectrales. Le /i/ anglais par exemple est réalisé avec un schéma F1/F2 similaire à celui du français mais avec un F3 plus bas suggérant notamment une voyelle peu étirée en comparaison du français. Le /i/ anglais s'éloigne de fait des caractéristiques focales décrites en introduction. Certaines tentatives ont ainsi été effectuées (Vaissière, 2007) pour affiner la notation phonétique de ces sons, en particulier /i/ et /y/.

En ce qui concerne /y/, l'écart F3-F2 est considérablement moins important pour le français que l'allemand. Quant à la voyelle /y/ pour le mandarin, la hauteur importante (2100 Hz) du 2ème formant laisse à penser qu'il s'approche du /y/ suédois, non caractérisé par un changement d'affiliation de cavité comme c'est généralement le cas pour le français (Schwartz et al., 1993 ; Vaissière, 2007). Cependant, le faible nombre d'occurrences recueillies (environ 150 contre plus de 1000 pour l'allemand) nous invite à considérer ces résultats avec prudence. Dans une étude précédente (Gendrot et Adda-Decker, 2005), nous avions observé pour l'allemand que les variations des voyelles antérieures fermées telles que le /i/ et le /y/ étaient aussi larges que celles des autres voyelles, alors que l'espace acoustique est plus « fourni » dans cette zone que le français, l'allemand ayant des contreparties « relâchées » aux voyelles /i/ et /y/ notamment. Selon ces résultats, cette différence peut être rapportée au fait que l'allemand n'est pas caractérisé par un rapprochement de F2 et F3 pour /y/ ou de F3 et F4 pour /i/ : ces voyelles ne sont pas focales en allemand.

Pour finir, nous avons effectué (Gendrot et al., 2008) quelques mesures sur les amplitudes des formants pour vérifier si dans le cas de voyelles plus longues, un rapprochement des formants F3/F4 pour /i/ ou F2/F3 pour /y/ génère une augmentation de l'amplitude des formants concernés. L'amplitude des formants a été mesurée de façon semi-automatique avec PRAAT (pour plus de détails, Gendrot et al., 2008) en mesurant la valeur la plus élevée de l'enveloppe spectrale obtenue par lissage cepstral autour de chaque formant, avec une fourchette de 100 Hz pour F1, 150 Hz pour F2 et 200 Hz pour F3. Il aurait été utile de prendre un certain nombre de précautions quant aux mesures d'amplitudes de formants, notamment par des mesures relatives, puisque nous n'avons aucun renseignement sur les conditions d'enregistrement. Cependant les très faibles barres d'erreur mesurées ici laissent à penser que ces mesures sont cohérentes. D'après les mesures brutes d'amplitudes de formants sur la Figure 15, nous pouvons observer globalement pour l'ensemble des voyelles, que les amplitudes des différents formants augmentent pour les voyelles de durée intermédiaire, puis baissent pour les voyelles les plus longues. Pour la voyelle /i/ par contre, l'amplitude du 4ème formant s'accroît à mesure que la durée de la voyelle augmente. Une tendance semblable peut-être observée pour l'amplitude du 3ème formant, bien que moins nette. De même pour la voyelle /y/, l'amplitude du 2ème formant ne décroît pas pour les durées les plus longues (mais pas pour l'amplitude du 3ème formant cependant).

Ces résultats (Gendrot et al., 2008) suggèrent un renforcement acoustique sensible de l'amplitude du 3ème formant et du 4ème formant à mesure que la durée de la voyelle /i/ augmente. Ces tendances sont moins nettes pour /y/. Des mesures relatives de différences entre les amplitudes des formants ont été effectuées pour estimer l'amplitude combinée des formants pour ces voyelles. Des mesures A1-A3 et A1-A4 indiquent des valeurs significativement plus faibles pour l'ensemble des voyelles plus longues - à l'exception des voyelles postérieures /o/ et /u/ caractérisées par une faible amplitude de F3 et F4 - suggérant ainsi une pente spectrale relevée. Des mesures A2-A3, A2-A4 et A3-A4 indiquent des valeurs significativement plus faibles et négatives pour les /i/ plus longs et à l'inverse, plus élevées pour les /y/ plus longs. Ces résultats suggèrent une proéminence spectrale F3/F4 pour /i/ et F2/F3 pour /y/. Ces mesures pourraient cependant être approfondies : la pondération de ces valeurs d'après leur ordre et leur position dans le spectre (Fant, 1970) pourrait être effectuée afin de prendre en compte leur poids perceptif relatif.

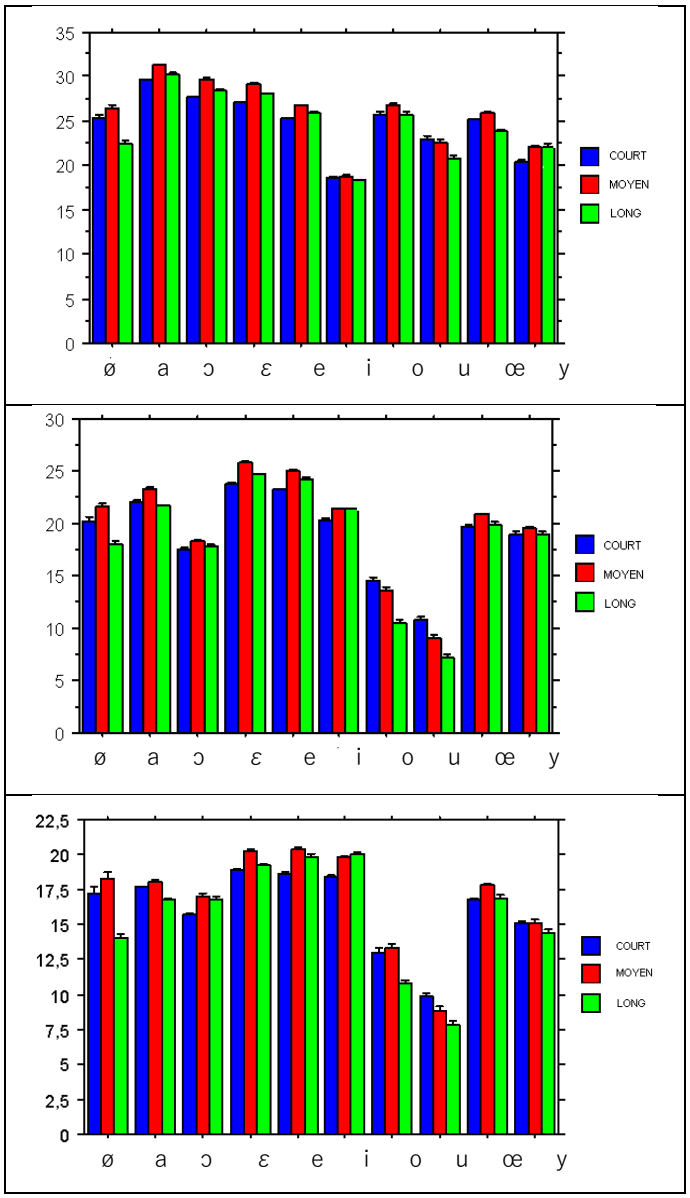


Figure 15 : amplitude des 2ème, 3ème et 4ème formants des voyelles orales du français en fonction de leur durée, d'après Gendrot et al. (2008)

4.1.5 Une réduction vocalique applicable à d'autres langues ?

Nous avons également effectué des comparaisons plus générales sur la réduction vocalique sur ces sept autres langues, en nous concentrant une nouvelle fois sur la relation entre la durée physique des voyelles et la fréquence de leurs formants, et plus particulièrement en envisageant l'espace acoustique global occupé par l'ensemble des voyelles et la taille de leur inventaire vocalique (cf. Tableau 17).

Nous nous sommes également intéressés à la variation des voyelles quantales /a/, /i/ et /u/ qui dans le cadre de la théorie quantique (Stevens, 1989) pourraient être plus stables que les autres voyelles de par la position rapprochée de leurs formants.

langue	Voyelles périphériques	Autres voyelles	Total.
Espagnol (Eur)	i u o a e		5
Italien	i u o a e		5
Arabe (litt)	i u a	ɪ ʊ æ	6
Chinois Mand.	i y u o a ɛ		6
Portugais (Eur)	i u o ɔ a ɛ e	ə ɯ	9
Français	i y u o ɔ a ɛ e	œ ø	10
Anglais (US)	i: u: o ɔ ɑ ɜ: ɛ	ɪ ʊ ə æ	11
Allemand	i: y: u: e: ø: ɛ: o: ɔ a:	ɪ ʏ ʊ a ø œ	15

Tableau 17 : nombre de voyelles dans l'inventaire vocalique de chaque langue analysée. Les voyelles périphériques permettant de mesurer l'espace occupé sont indiquées dans une colonne spécifique, d'après Gendrot et Adda-Decker (2007)

Les tons du mandarin n'ont pas été considérés dans cette étude et sont fusionnés pour chaque timbre vocalique : les considérer à part aurait multiplié le nombre de catégories en réduisant le nombre d'occurrence pour chacune d'entre elles. D'autre part, une analyse préliminaire nous a confortés dans l'idée que les variations formantiques entre chaque réalisation tonale restaient faibles pour les aspects abordés ici.

Etant donné les différences entre nos corpus, nous avons effectué (Gendrot et Adda-Decker, 2007) une procédure de normalisation des voyelles, en évitant les procédures plus classiques (dites extrinsèques) telles que Lobanov, Nearey, Gerstmann, qui peuvent avoir un effet sur les distances entre voyelles puisque les relations entre voyelles sont prises en compte (voir Gendrot, 2013 pour une revue). Nous avons eu recours à la mesure de normalisation de Syrdal et Gopal (1986) qui se résume selon les étapes suivantes : (1) les données de formants sont converties en Bark selon la formule de Traunmüller (Traunmüller et Eriksson, 1995). (2) c'est une normalisation dite intrinsèque puisqu'elle a recours uniquement aux formants de chaque voyelle individuellement. Elle prend en compte les différences entre la f0 et les formants adjacents : F1 – f0 en lieu et place du premier formant pour traiter de l'aperture de la voyelle ; et F3-F2 pour la dimension antéro-postérieure et qui serait plus fiable que F2 selon les auteurs de cette normalisation. Selon Adank (2003), cette normalisation est efficace pour réduire les différences anatomiques. Elle est également, selon nous, intéressante dans la mesure où elle intègre le troisième formant sans avoir à ajouter une dimension supplémentaire dans la représentation (elle s'approche en ce sens du F2 prime).

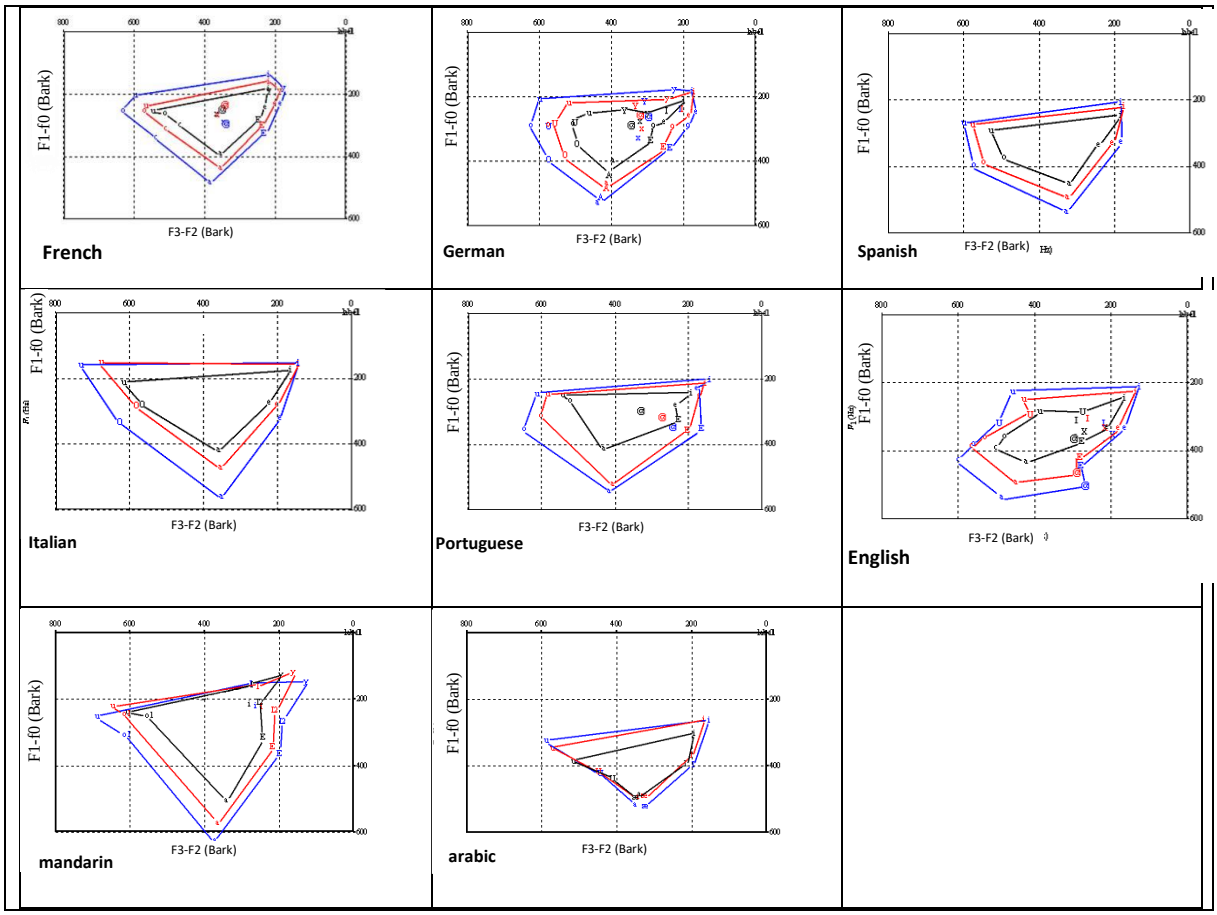


Figure 16 : valeurs moyennes de F1-f0 et de F3-F2 pour les 8 langues en fonction de la durée mesurée des voyelles. De l'intérieur vers l'extérieur (noir [30 – 50 ms], rouge [60 - 80], bleu [90 - 110]), d'après Gendrot et Adda-Decker (2007)

Au-delà des mesures de dispersion déjà présentées, nous avons également eu recours à une mesure de l'espace vocalique (Gendrot et Adda-Decker, 2007 ; Gendrot et Adda-Decker, 2010) en calculant le périmètre constitué par les voyelles périphériques (les voyelles sont considérées comme périphériques quand elles favorisent un espace maximum tout en minimisant le contour de ce même espace). Les résultats de ces analyses montrent une tendance claire à la réduction spectrale des voyelles courtes pour les huit langues considérées, avec toutefois une magnitude moins importante pour l'arabe qui seul possède le trait phonologique de quantité. Ces résultats vont dans le sens d'une contrainte physiologique de la réduction spectrale. Nous n'avons pas pu montrer un effet de la taille de l'inventaire vocalique sur l'espace vocalique global comme cela est prédit par la théorie de la dispersion adaptative (voir Al-Tamimi et Ferragne, 2005 ; Meunier et al., 2003, pour des résultats divergents). Ce résultat négatif pourrait soutenir l'idée que les langues utilisent des critères additionnels tels que la nasalité, la quantité, la diphtongaison ou la qualité de voix pour ajouter de nouvelles voyelles dans un espace acoustique restreint. La mesure de la dispersion doit nécessairement jouer un rôle comme détaillé par Audibert et al. (2015) et présenté dans la section suivante. Sur la base de mesures de dispersion (illustrées en section 4.2), nous avons pu montrer que seule la voyelle quantale /i/ affiche plus de stabilité formantique que les autres voyelles, ce qui n'est pas le cas pour les deux autres voyelles quantales /a/ et /u/. Le /a/ est une voyelle très sensible à la coarticulation et il n'est pas très surprenant d'observer de fait une forte variabilité. La variabilité du /u/ pourrait s'expliquer en considérant que le /u/ est produit par une constriction plus centrale qu'elle ne l'est présentée dans les ouvrages de phonétique articulatoire (Vaissière, 2009) et que la stabilité de

son deuxième formant serait en fait due à un arrondissement des lèvres qui n'est pas phonologique et de fait non contraint. La stabilité plus importante de /i/ pourrait s'expliquer par la non prise en compte du troisième formant dans les mesures de dispersion, alors que pour le français par exemple, c'est essentiellement la mesure du troisième formant qui permet de mesurer la variation du /i/ (Gendrot et Adda-Decker, 2008).

4.1.6 Styles de parole

Les résultats que nous avons présentés jusqu'à présent traitaient exclusivement de données issues de parole journalistique. Afin de vérifier si ces résultats sont spécifiques à ce style de parole, nous allons utiliser dans cette section le corpus de parole spontanée NCCFr (Nijmegen Corpus of Casual French). Ce corpus est composé de 22 conversations en binômes d'environ 1 heure et demie chacune, soit approximativement 45 minutes de parole pour chacun des 44 locuteurs, transcrite manuellement et segmentée automatiquement par le LIMSI (Torreira et al., 2010). La première question de recherche concerne les phénomènes de réduction importants que nous avons pu observer pour la parole journalistique et nous cherchons ici à savoir si ces phénomènes peuvent être accrus en parole spontanée. Le cas échéant, cette réduction supplémentaire par rapport à la parole journalistique pourrait-elle être due uniquement à des différences de durée phonémique, et par extension de débit ? Des premières mesures montrent que le corpus de parole spontanée contient un nombre beaucoup plus important de voyelles courtes. Pourrait-on malgré tout observer des différences de réduction pour des durées vocaliques comparables dans les deux corpus ? Ce résultat montrerait que ce n'est pas le simple effet de durée qui est en cause mais bien un style spontané, plus hypo-articulé que la parole journalistique (Gendrot et al., 2012 ; Gendrot et al., 2015).

L'approche présentée ici a été effectuée depuis un découpage en séquences, dont le critère de sélection a été la présence de pauses (d'un minimum de 50 ms) telles que détectées par le système d'alignement son-texte. Le Tableau 18 présente un résumé de ces séquences dans le corpus de parole journalistique et le corpus de parole spontanée.

	nombre	Durée moyenne en (s)	Ecart-type	Débit moyen (phon/s)
p. journalistique	562 935	2.71	1.43	13.6
p. spontanée	516 933	1.68	1.15	15.3

Tableau 18 : caractéristiques principales des séquences dans les deux corpus (ESTER et NCCFr), d'après Gendrot et al. (2015)

L'objectif est d'analyser à l'intérieur de ces séquences les phénomènes d'allongements et de réduction. Plus tard, dans la section 4.2.5, nous analyserons également les variations de f0 (incluant les phénomènes de déclinaison). Les variations de durée sont souvent laissées de côté dans la littérature, notamment dans les modèles prosodiques par exemple, principalement à cause des variations importantes de durée intrinsèque des différents phonèmes. Or il pourrait apparaître que les phénomènes d'allongement caractérisant les fins d'unités prosodiques sont prédominants par rapport aux phénomènes mélodiques, et ce particulièrement dans le cas de la parole spontanée. Pour ce faire,

nous avons utilisé une mesure que nous avons utilisée à plusieurs reprises dans nos travaux : une mesure de durée normalisée en effectuant le ratio de la durée des phonèmes mesurés dans leur contexte phonémique : nous avons nommé cette mesure « ralentissement » (Gendrot et al., 2012 ; Gendrot et al., 2015).

La somme de la durée de la voyelle d'intérêt dans son contexte précédent et de la durée de la même voyelle dans son contexte suivant divisée par deux. Il eût été plus efficace de considérer des triphones mais un trop faible nombre d'occurrences d'une quantité importante de triphones nous a obligé a procéder en deux moitiés. Cette valeur moyenne de durée est ainsi considérée comme une référence afin d'en calculer le ratio par rapport à la voyelle sous étude.

$$(\mathrm{C_{prec}})\ v\ (\mathrm{C_{suiv}}) = \frac{\text{durée}\ (\mathrm{C_{prec}})\ v\ + v\ \text{durée}\ (\mathrm{C_{suiv}})}{2}$$

Comme mesuré par Nooteboom (1997) pour la longueur des phonèmes à l'intérieur des mots, et plus classiquement décrit par la loi de Menzerath, parmi les séquences que nous avons analysées, plus la séquence mesurée est longue, plus le nombre de phonèmes contenus dans cette séquence est important, et plus la durée de ces phonèmes est faible (cf. Figure 17).

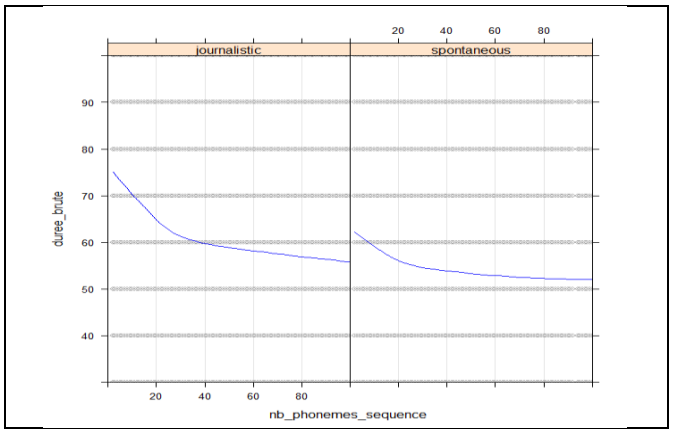


Figure 17 : mesure de durée phonémique en fonction du nombre de phonèmes dans la séquence. A gauche parole journalistique et à droite parole spontanée, d'après Gendrot et al. (2015)

Pour la Figure 18 ci-dessous, la durée de chaque phonème (normalisée par rapport aux valeurs de durée de référence) est affichée en fonction de sa position au sein de la séquence (en pourcentage de durée). Pour les deux corpus, nous observons un allongement qui commence à partir de 60 % de la durée de la séquence, mais significativement moins net en parole spontanée. L'allongement de début de séquence est observé en parole journalistique seulement. Ces résultats sont observés quelle que soit la durée de la séquence.

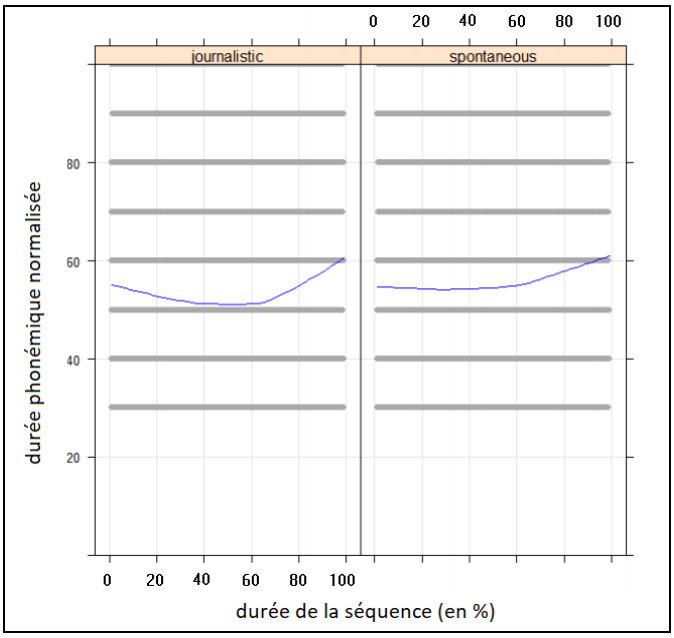


Figure 18 : mesure de durée phonémique normalisée en fonction de la position dans la séquence. A gauche parole journalistique et à droite parole spontanée, d'après Gendrot et al. (2012)

4.1.6.1 Comparaison de la réduction spectrale

Après avoir observé des valeurs de débit plus élevées, et des durées vocaliques plus courtes en parole spontanée, nous pouvons visualiser ci-dessous l'espace vocalique fournissant un indice de la réduction vocalique. L'espace vocalique étant plus petit, la réduction vocalique est plus importante en parole spontanée (Figure 19). En considérant des catégories de durée comparables (de 30 ms sur la Figure 20 gauche, ou 40 ms sur la Figure 20 droite), l'espace vocalique est toujours plus petit pour la parole spontanée, ce qui indique que la durée induite par un débit plus rapide n'est pas le seul facteur impliqué dans la réduction spectrale. La fusion des sons /u/ et /o/ ou bien des sons /y/ et /e/ apparait comme récurrente et pourrait marquer une évolution possible de ces sons dans une perspective diachronique.

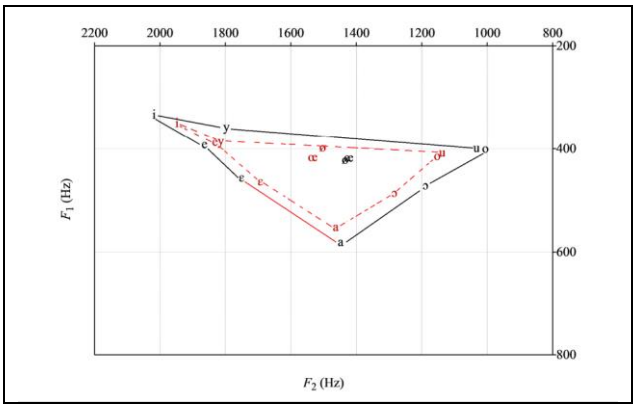


Figure 19 : espace vocalique pour les locuteurs masculins en parole journalistique (traits pleins) vs. parole spontanée (pointillés) pour toutes les voyelles sans distinction de durée, d'après Gendrot et al. (2015)

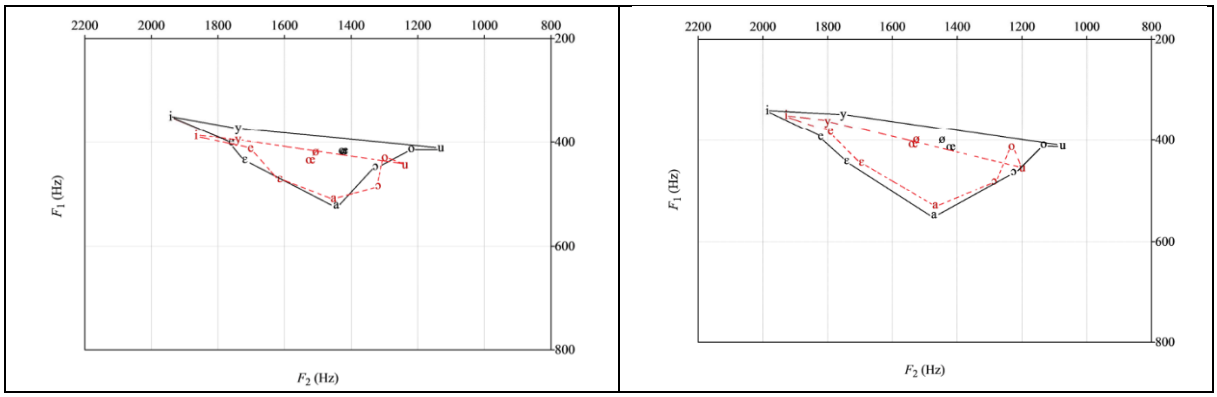


Figure 20 : espace vocalique pour les locuteurs masculins en parole journalistique (traits pleins) vs. parole spontanée (pointillés). A gauche, voyelles de 30 ms ; à droite, voyelles de 40 ms, d'après Gendrot et al. (2012)

La quantification de la variation d'un phonème dans un système (Audibert et al., 2015 ; Gendrot et al., 2008) a fait l'objet de plusieurs publications puisque la variation d'un phonème doit être comprise par opposition à d'autres phonèmes avec lesquels il s'oppose, ou bien dans le système phonémique tout entier. L'évaluation de la réduction telle que nous la présentons pourrait être considérée comme simpliste puisque ne prenant en compte que l'aire de l'espace vocalique de façon indirecte. Dans un travail effectué en collaboration avec Nicolas Audibert, d'autres métriques que celle de la distance par rapport au centre ont été testées, incluant l'aire de l'espace vocalique occupé par les voyelles périphériques, le ratio d'écart sur F1 ou F2, l'aire de dispersion de chaque voyelle, ou la perte de contraste entre catégories vocaliques. Cette dernière métrique intitulée 'ContrastLoss' est calculée comme le taux de mauvaises classifications de la voyelle dans sa classe dans un modèle à prédiction linéaire (voir Harmegnies et Poch-Olivé, 1992).

Dans cette étude (Audibert et al., 2015), nous avons également ajouté le style de parole « lecture » (par le biais du corpus BREF (Lamel et al., 1991), aux corpus de parole journalistique et de parole spontanée. Ceci nous a permis de considérer le style et la durée phonétique comme deux variables qui ont pu être comparées afin de vérifier leur poids dans les phénomènes de réduction vocalique. Les deux variables croisées dans cette étude sont le style (lu, journalistique, spontané) et la durée phonétique des voyelles (court, moyen, long). Un modèle linéaire à effets mixtes a été construit avec le style et la classe de durée comme effets fixes et le locuteur comme effet aléatoire afin de prédire la variation sur les métriques acoustiques. Les tests ont révélé que ces métriques montrent toutes un effet significatif à $p<10^{-5}$ sur les variables style et durée. Les voyelles courtes sont plus réduites que les voyelles moyennes, elles-mêmes plus réduites que les voyelles longues. De même le style de parole spontanée implique des réalisations plus réduites que la parole journalistique, elle-même impliquant plus de réductions que la parole lue. Afin de comparer l'influence de la durée et du style sur les valeurs des métriques, des tailles d'effets estimées par des valeurs de $\chi 2$ dans les tests de rapport de vraisemblance ont été calculées et indiquent un effet plus large de la durée pour toutes les métriques sauf pour la métrique 'perte de contraste' qui montre des effets de réduction plus importants du style comme l'indique la Figure 21 à droite comparée par exemple à la métrique de distance (à gauche) par rapport au centre acoustique.

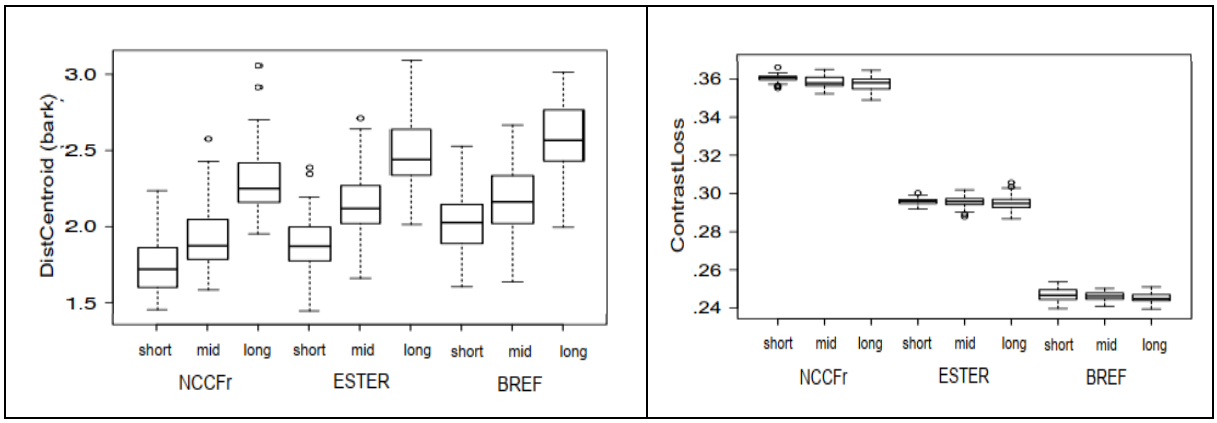


Figure 21 : boites à moustaches des trois corpus NCCFr, ESTER, BREF (spontané/journalistique/lu) et des trois classes de durée (court, moyen et long) pour deux des métriques utilisées : DistCentroid et ContrastLoss, d'après Audibert et al. (2015)

Ces résultats (Audibert et al., 2015) montrent un effet cumulé de la durée et du style de parole dans la réduction vocalique. Cependant, en fonction de la métrique utilisée pour évaluer la réduction vocalique, il est possible d'observer un effet plus important de l'un ou de l'autre. Pour conclure, il parait évident que ces deux facteurs se combinent dans leur rôle d'hypo- ou hyper-articulation.

4.1.7 Résumé des analyses vocaliques automatiques

Avant de nous orienter vers les aspects prosodiques des grands corpus et leur traitement automatique, nous profitons de cette transition pour effectuer une rapide discussion rétrospective de cette section. Après avoir présenté des valeurs formantiques de référence du français en fonction du contexte consonantique, nous avons montré des phénomènes de réduction acoustique pour toutes les voyelles en fonction de leur durée phonétique, du style de parole, mais également en fonction du contexte consonantique. La majorité des contextes consonantiques tend à une centralisation acoustique des voyelles, et la durée phonétique des voyelles accentue cette tendance. Cette réduction s'observe également dans plusieurs langues avec des contraintes phonologiques différentes. Nous avons montré que des mesures effectuées de façon automatique sur des corpus alignés automatiquement restent cohérentes si l'on prend quelques précautions méthodologiques.

Il est malgré tout nécessaire de poursuivre ces investigations et de chercher d'autres sources impliquées dans les phénomènes de réduction observés. Des travaux plus récents ont montré notamment que l'information lexicale et la structure prosodique jouaient un rôle dans l'hyper- et hypo-articulation des segments (Aylett et Turk, 2006 ; Jaeger, 2010). Une analyse de facteurs prosodiques tels que la position de la syllabe dans le mot, ou dans différents types de constituants pourra nous aider à mieux cerner ces phénomènes, ce que nous présentons dans la section suivante. Les informations lexicales seront également abordées dans la section 4.3.1, et nous aborderons ces résultats à la lumière des travaux les plus récents dans la discussion générale.

4.2 Structuration prosodique de l'information

Le français n'est pas une langue à accent lexical, et est typiquement considéré comme une langue à rythme syllabique (« syllable-timed ») avec un allongement final en fin de groupes des sens (ou groupes rythmiques). On pouvait s'attendre à ce que des phénomènes de réduction vocalique liés à la durée soient plus faibles que pour une langue à accent lexical, tel que cela était suggéré par Delattre (1965). Les phénomènes d'accentuation seraient ainsi reportés à des niveaux supérieurs au mot. Après avoir analysé la variation formantique des voyelles en fonction de leur durée, notre objectif était de localiser ces sources de variation de durée, c'est-à-dire les positions dans la parole qui induisent les phénomènes d'hypo- et d'hyper-articulation. Deux facteurs pourraient expliquer un allongement vocalique en français : (i) La présence d'accents de focalisation/emphase qui, en français, se positionnent plus vraisemblablement sur les syllabes initiales de mots et qui impliquent – en plus d'une intensité accrue et d'une f0 augmentée – un allongement vocalique. (ii) La proximité de frontières est une cause fréquente d'allongements vocaliques et peut également être corrélée à la présence de pauses. Dans nos investigations, nous nous sommes intéressés à ce deuxième facteur de variation. Dans un premier temps, les voyelles de syllabes finales de mots ont été analysées puisque potentiellement plus longues, et comparées aux voyelles en syllabe initiale de mot (généralement non allongées, sauf en cas de présence d'un accent d'insistance) et aux syllabes internes de mot (ni initiales ni finales, et théoriquement les plus courtes). Dans un deuxième temps, nous avons effectué les mêmes analyses du point de vue de la présence de pauses, puisque celles-ci génèrent un ralentissement des phonèmes dans leur entourage (Gendrot et al., 2006). Pour finir, sur la base d'un algorithme de regroupement lexical basé sur la transcription orthographique, nous avons identifié des unités s'approchant des syntagmes accentuels et nous avons pu analyser les frontières (initiales et finales) de ces syntagmes (Gendrot et al., 2011 ; Gendrot et al., 2012). Ces travaux s'inscrivent dans la lignée de travaux sur la prosodie articulatoire (Fougeron, 2001 ; Tabain et Perrier, 2005) qui ont montré que pour des positions initiales et finales de constituants prosodiques formant une hiérarchie (allant de la syllabe, au mot, au syntagme accentuel, et pour finir au syntagme intonatif), on observe une hyper-articulation progressive caractérisée par un renforcement articulatoire tel que le contact lingual ou le débit d'air nasal.

4.2.1 Variations morphologiques

Les frontières de mots ont été obtenues sur la base de la transcription manuelle, puis de l'alignement effectué automatiquement. Les voyelles décrites comme initiales de mots sont considérées comme des initiales absolues, comme par exemple dans les mots 'arme' [a̲ʁm] et les positions décrites comme finales de mot sont en finale absolue, comme par exemple dans le mot 'bras' [bʁa̲]. Nous aborderons dans la discussion de cette section la possibilité de prendre en compte la position en pénultième comme pour le mot 'trame' [tʁa̲m] et la 2ème position du mot comme pour le mot 'larme' [la̲ʁm]. Les voyelles initiales de mots sont fréquemment produites avec une liaison en français ("deux [dø] amis [ami]" ==> [døzami]). Dans le processus de segmentation automatique, les liaisons sont annotées si elles sont prises en compte dans le dictionnaire de prononciation. Si l'appartenance phonologique de la consonne de liaison au premier ou au deuxième mot fait encore débat, elle est considérée dans le système automatique que nous avons utilisé comme appartenant au premier mot (et donc [a] sera considéré comme initiale de mot). Bien que ce phénomène puisse être pris en compte pour une étude future, nous avons décidé de laisser la segmentation automatique telle quelle. Remarquons pour finir

que la phonotactique du français implique que /ɔ/ ne peut pas se trouver en position finale et que, par conséquent¸ aucun résultat pour cette voyelle ne sera présenté en position finale absolue.

La Figure 22 révèle que les voyelles en syllabe finale de mot ont des valeurs plus extrêmes que les syllabes initiales de mot sur les axes F1/F2, à l'exception des voyelles antérieures fermées /i/ et /e/. Une normalisation (Lobanov) est utilisée pour faciliter la présentation des résultats cumulés hommes/femmes ici.

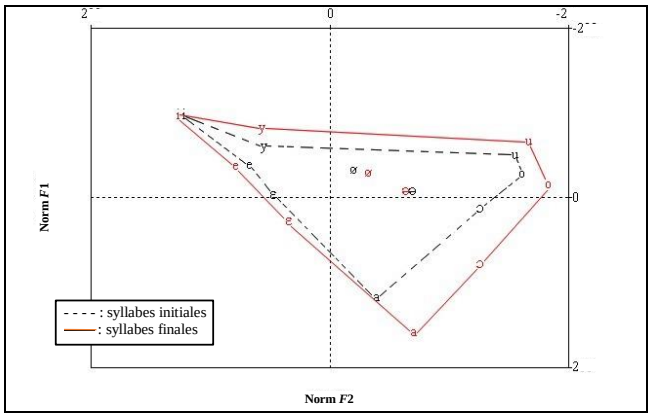


Figure 22 : valeurs moyennes de F1 et F2 pour les voyelles en syllabe initiale et finale (Normalisation lobanov, toutes durées confondues), d'après Gendrot et al. (2006)

Le Tableau 19 et le Tableau 20 détaillent la répartition des voyelles en fonction de leur position dans le mot, et en fonction de leur catégorie de durée. Afin de mieux interpréter ces résultats, nous avons également illustré une mise en regard de la distribution des voyelles en fonction de leurs valeurs de f0. Des catégories de f0 équilibrées (uniformes) ont ainsi été déterminées sur le même principe que les catégories de durée :

	f0 basse	f0 moyenne	f0 haute
Hommes	<= 110 Hz	110 <f0 <= 140	> 140 Hz
Femmes	<= 160 Hz	160 <f0 <= 210	> 210 Hz

Nous pouvons observer sur ces tables que les voyelles longues sont également celles qui ont des valeurs de f0 plus élevées, plus particulièrement en syllabe finale de mots. Les voyelles longues sont également plus fréquentes lorsque positionnées en syllabe finale de mots. Ce tableau résume simplement les caractéristiques prosodiques morphologiques du français où les voyelles en syllabe initiale et intermédiaire de mot sont plus courtes avec une f0 plus basse que les voyelles en syllabe finale de mot. Une augmentation de la f0 sur la voyelle pourra entrainer un relèvement léger de l'ensemble des formants (Gendrot, 2005). Cette relation entre la f0 et la durée sera utile lors de l'interprétation syntaxique des phénomènes d'hyper-articulation en section 4.2.3.2.

f0 / durée	bas	moyen	haut	total (durée %)
Court	**14**	**14**	9	37
Moyen	11	**14**	**14**	**39**
long	5	7	**12**	24
total(f0 %)	30	35	34	100

Tableau 19 : répartition (en %) des voyelles en syllabe initiale de mots en fonction de leurs catégories de durée et de f0 (les nombres en gras représentent les catégories majoritaires), d'après Gendrot et al. (2006)

f0 / durée	bas	moyen	haut	total (durée %)
court	10	**11.5**	8.5	30
moyen	8	11	**14**	33
long	7.5	10	**19**	**37**
total(f0 %)	25	33	42	100

Tableau 20 : répartition (en %) des voyelles en syllabe finale de mots en fonction de leurs catégories de durée et de f0 (les nombres en gras représentent les catégories majoritaires), d'après Gendrot et al. (2006)

La Figure 23a et la Figure 23b montrent les variations mesurées pour les voyelles en syllabe initiale et finales respectivement, tout en mettant en évidence les différentes catégories de durées utilisées précédemment dans Gendrot et Adda-Decker (2005), à savoir [30–50 ms] pour les voyelles courtes, [60–80 ms] pour les voyelles à durée intermédiaire et [90-110] pour les voyelles longues. Ces figures indiquent à première vue des tendances similaires, également identiques à celles relevées par la Figure 10. Nous pouvons cependant signaler que les voyelles ouvertes /a/ et /ɛ/ mesurées en syllabe finale de mots ont des valeurs de F1 significativement plus élevées que ces mêmes voyelles mesurées en syllabe initiale de mots. Ces différences observées sur l'axe F1 ont également été notées par Gendrot (2005) sur des triangles vocaliques effectués sur la base des catégories de f0 mentionnées ci-dessus et non plus de catégories de durée.

Les résultats diffèrent également concernant la position de la syllabe dans le mot pour les voyelles fermées /i/, /y/, /u/ et /o/. En effet, les variations formantiques pour ces voyelles sur l'axe F1 sont non significatives ou peu importantes lorsque mesurées en syllabe finale de mots, alors que l'on peut remarquer un net détachement du 1er formant des voyelles fermées courtes en syllabe initiale (avec des valeurs plus élevées) par rapport à leurs contreparties plus longues. Nous suggérons ici que les variations observées dans la distribution de ces voyelles, ainsi que les variations formantiques mesurées correspondent à la répartition naturelle de ces voyelles. La position allongeante de fin de mot implique donc des voyelles plus périphériques, à l'exception des voyelles antérieures fermées /i/

et /e/. Les variations formantiques de ces voyelles avaient été notées comme plus faibles précédemment.

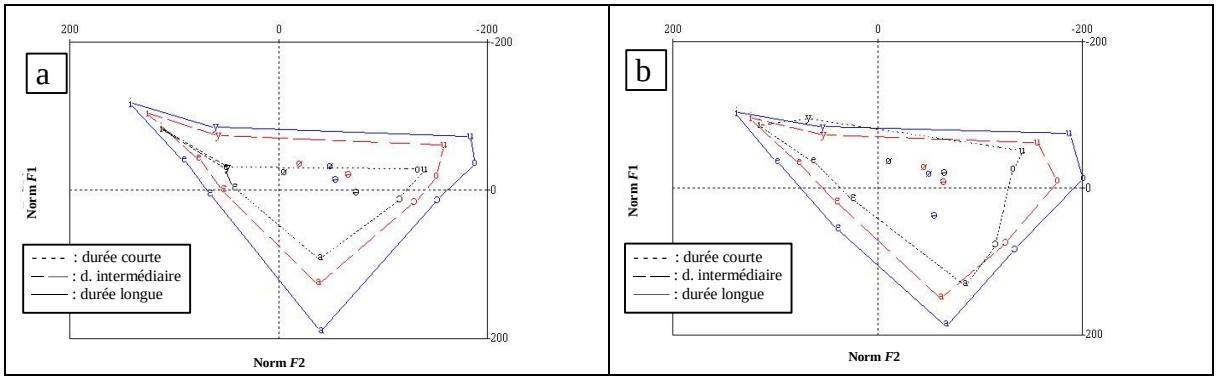


Figure 23 : valeurs moyennes de F1 et F2 pour les voyelles en fonction de leur durée (normalisation lobanov). a : voyelles en syllabe initiale de mots b : voyelles en syllabe finale de mots

4.2.2 Variations prosodiques : présence de la pause

Nous avons également réalisé des mesures identiques en fonction de la présence de pauses immédiatement avant et après les voyelles analysées. Les pauses ont été déterminées dans la transcription par l'algorithme d'alignement suivant le même principe que les phonèmes, avec une durée minimale de 50 ms. Le Tableau 21 résume ainsi les catégories obtenues au moyen d'exemples, ainsi que le nombre d'occurrences recueillies.

	pause précédant V.	pause suivant V.
pause	**[pause]** **a**bri	matel**a**s **[pause]**
occurrences	avec : 920 sans : 22000	avec : 1600 sans : 21300

Tableau 21 : résumé des différentes catégories sélectionnées, d'après Gendrot et al. (2006)

Rappelons ici qu'il est particulièrement délicat de se baser sur la syntaxe de la transcription pour déterminer des éventuelles frontières prosodiques comme cela est fait pour l'analyse de phrases lues. En effet, il est fréquent par exemple de constater la présence de pauses entre un article et le nom qui lui est associé, ce qui va à l'encontre de bon nombre de prédictions, et est souvent considéré comme une (pause de) mise en valeur du mot suivant cette pause. Nous avons ainsi décidé de déduire des frontières prosodiques grâce à la détection de pauses. En effet, la présence d'une pause en français est fréquemment corrélée à une fin de groupe intonatif en français qui est également marquée par un fort allongement vocalique. Comme cela est illustré par la Figure 24a et la Figure 24b, les voyelles précédées ou suivies d'une pause sont caractérisées par des valeurs plus extrêmes, occupant ainsi un espace acoustique plus large. A nouveau, ces variations sont plus faibles pour les voyelles antérieures fermées.

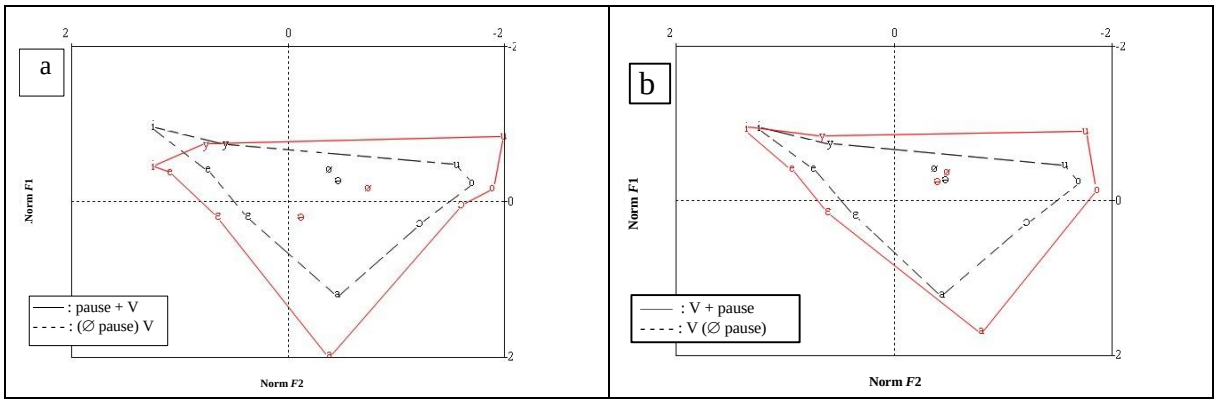


Figure 24 : valeurs moyennes de F1 et F2 pour les voyelles en fonction de l'absence/présence de la pause à proximité de la voyelle analysée. a : pause précédant la voyelle ; b : pause suivant la voyelle, d'après Gendrot et al. (2006)

4.2.3 Variations prosodiques : au niveau du syntagme

Nous avons pu mesurer l'impact de la position syllabique dans le mot et de la présence d'une pause sur la réalisation acoustique des voyelles orales. La position de la voyelle dans le syntagme a également une influence significative. Les constituants prosodiques sont produits dans la parole à des fins de démarcation pour l'auditeur puisqu'ils signalent des occurrences de frontières (Vaissière et Michaud, 2006). Ceux-ci sont également considérés comme étant organisés au sein d'une hiérarchie prosodique : chaque constituant est imbriqué dans un constituant de plus haut niveau (Nespor et Vogel, 2007). Dans ce courant de recherche, il a été montré que plus le niveau du constituant prosodique est élevé dans la hiérarchie, plus les phonèmes aux frontières de ces constituants sont longs et hyper-articulés (pour le français, Fougeron, 2001 ; Tabain, 2003). Ces positions ont plus rarement été étudiées d'un point de vue spectral. Pour le français, Fougeron (2001) à partir de l'électropalatographie (EPG) et Tabain (2003, 2005) grâce à l'articulographie ont travaillé respectivement sur les positions initiales et finales. Leurs résultats montrent qu'un phonème en position initiale ou finale de constituant est influencé par sa position dans la hiérarchie prosodique. Plus le niveau est élevé, plus le phonème est hyper-articulé (voir aussi Cho, 2005 pour l'anglais américain, Fuchs et al., 2013 pour l'allemand).

Afin de réussir à reproduire ces études, nous avons ajouté le niveau du syntagme comme constituant prosodique (Gendrot et al., 2016), et nous l'avons comparé aux positions mesurées dans les deux sections précédentes (les frontières de mot et les frontières de constituant précédé/suivi d'une pause). La nouveauté réside dans l'utilisation d'un algorithme qui, sur la base de la transcription orthographique, va constituer des syntagmes dont nous considérons qu'ils sont très proches des syntagmes accentuels analysés dans les travaux sus-cités. Nous avions pour objectif d'évaluer si les voyelles positionnées à des frontières de constituants prosodiques de plus haut niveau sont hyper-articulées comparées à leurs homologues de constituants prosodiques de plus bas niveau. Pour ce faire, nous avons évalué l'espace acoustique utilisé par l'ensemble des voyelles périphériques pour chacun des niveaux prosodiques (en analysant les positions initiales et finales séparément). Nous avons également repris la mesure de dispersion utilisée précédemment. Selon les hypothèses avancées précédemment, nous nous attendons à un élargissement de l'espace vocalique à mesure que l'on remonte dans la hiérarchie prosodique, i.e. en partant de la syllabe, puis du mot, en passant par le groupe accentuel et pour finir le groupe intonatif.

Afin de quantifier ces différences, nous avons calculé la dispersion (par distance euclidienne) pour chaque niveau prosodique depuis le centre acoustique mesuré sur l'intégralité des données (F1 à 450 Hz et F2 à 1450 Hz, calculé dans Gendrot et Adda-Decker (2007) pour le français, et inspiré de Bradlow et al. (1996) comme illustré dans la Figure 25).

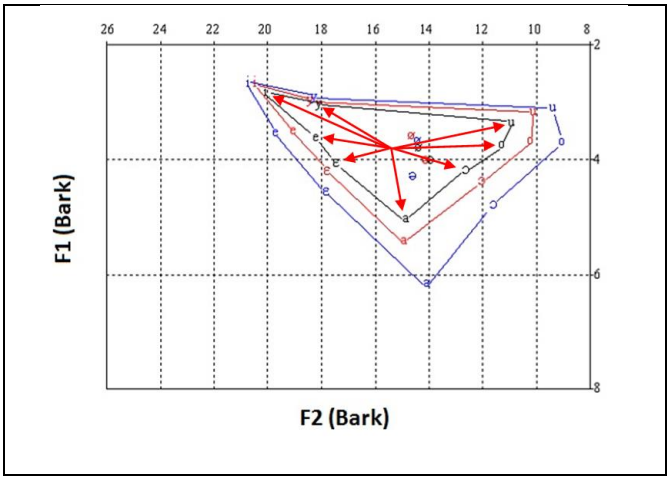


Figure 25 : illustration de la mesure de dispersion vocalique

Si la voyelle s'éloigne du centre acoustique d'un niveau prosodique à l'autre de façon significative, alors elle est considérée comme hyper-articulée. Nous gardons à l'esprit que cette mesure peut s'avérer imprécise dans la mesure où elle est liée au degré de centralisation de la voyelle qui, comme précisé ci-dessus, n'est qu'une conséquence secondaire de la coarticulation. Cependant, lorsque toutes les voyelles s'éloignent les unes des autres afin de favoriser leur distinction, elles accroissent nécessairement l'espace vocalique (voir la théorie de la dispersion adaptative de Lindblom (1990) pour une interprétation de ce phénomène dans les préférences des systèmes vocaliques des langues du monde).

Nous décrivons ci-dessous (Gendrot et Gerdes, 2011 ; Gendrot et Gerdes, 2012)) les choix effectués pour détecter automatiquement chaque catégorie, ou plus précisément les deux catégories restantes puisque les frontières de mots ainsi que les frontières de groupes intonatifs (via la présence de pause) ont été définis dans la section précédente. Manquent encore le niveau syllabique et le syntagme (accentuel).

Les syllabes sont déterminées à partir de la segmentation. Des règles de syllabation inspirées de Pallier (1994) et Adda-Decker et al. (2005) ont été utilisées à partir du flux continu de phonèmes, i.e. sans prendre en compte les frontières de mots. Par exemple, pour la séquence de deux mots 'bon ami' [bɔnami] est segmentée en 3 syllabes : 'bo', 'na' et 'mi' (à moins qu'elles ne soient séparées par une pause, auquel cas on aurait [bɔn.ami]. Les pauses sont considérées comme des délimiteurs et les syllabes, selon ce principe, ne peuvent pas contenir de pauses. Contrairement aux autres catégories, les syllabes ne seront pas analysées à leurs positions initiales et finales. En effet, il était difficile - sinon impossible - de collecter des voyelles en position initiale de syllabe tout en étant syllabe intermédiaire de mot. Les seules syllabes de ce type sont les syllabes de type 'V' contenues par exemple dans le mot 'aéroport' qui représentent 2 % des syllabes dans le corpus étudié ici. Nous avons donc décidé de

prendre en compte les syllabes qui ne sont ni finales ni initiales de mots, sans filtrer le type de syllabe. Par conséquent, les voyelles considérées ici comme internes de mots sont de façon prédominante (à 78 %) des voyelles en position finale de syllabe.

Le troisième niveau analysé ici est le syntagme accentuel ou groupe accentuel (voir exemple ci-dessous). Afin d'obtenir ce niveau, un chunking syntaxique a été effectué sur la base d'un étiquetage grammatical automatique combiné à quelques règles de regroupement mises en place grâce au chunker du Natural Language Toolkit (http://nltk.org/index.php/Main_Page) :

1. Dans un premier temps, chaque mot a été étiqueté selon toutes les catégories disponibles dans le dictionnaire français des formes fléchies (Lefff : Clément et al., 2004), légèrement modifié pour nos besoins. Par exemple, nous avons supprimé les catégories de mots rares qui se confondaient avec des catégories de mots plus fréquents (par exemple pour l'adjectif 'sûr' qui entrainait des étiquetages erronés.

2. Ensuite le chunker du *Natural Language Toolkit* a été utilisé pour générer deux types de segments :

    a. Les noms, les prépositions et les verbes sont regroupés avec leur entourage le plus proche (clitiques, déterminants, prépositions, adjectifs, etc.)

    b. Toutes les séquences de mots non définies par la règle précédente ont été regroupées

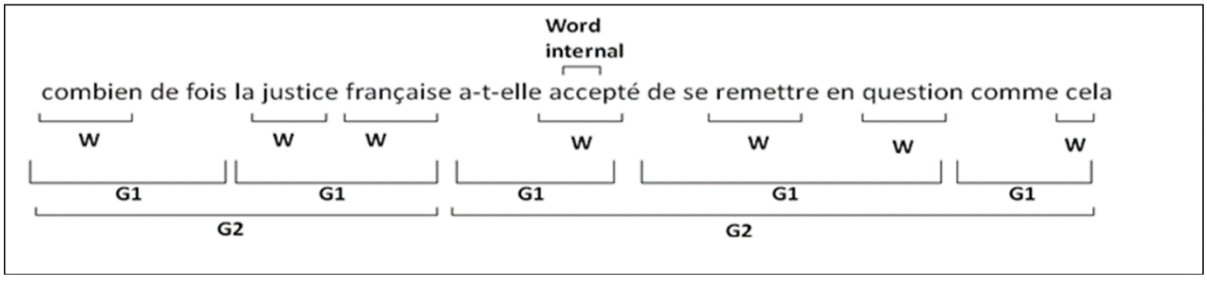
3. Dans une étape finale, trois règles de regroupement ont été appliquées :

    a. Regroupement de tout segment terminant sur un auxiliaire ou modal avec le segment suivant

    b. Regroupement de tout segment verbal avec le segment suivant si la combinaison fait moins de 7 syllabes

    c. Regroupement de tout autre suite de segments qui fait moins de 7 syllabes

Un exemple du résultat de cet algorithme est illustré ci-dessous



Les segments découpés par cet algorithme peuvent avoir plus de 7 syllabes, si les règles précédentes le permettent, par exemple « avec qui j'ai pu m'entretenir », qui forme un groupe très naturel et difficile à découper. La règle des 7 syllabes (Wioland, 1985) sera discutée ultérieurement dans cette étude. Par ce "chunking", nous tentons de nous approcher de la réalisation du syntagme accentuel. Nous sommes conscients que tous ces syntagmes ne seront pas "accentués", c'est-à-dire qu'ils ne seront pas tous caractérisés par un allongement final et/ou un contour mélodique montant. Cependant, utiliser des informations prosodiques pour s'en assurer aurait introduit un caractère circulaire dans notre étude puisque les voyelles les plus longues sont elles-mêmes hyper-articulées. Notre méthode vise donc à évaluer la réalisation spectrale de syntagmes accentuels à un niveau

syntaxique (sous-jacent), plutôt qu'en considérant des syntagmes accentuels d'après leurs caractéristiques prosodiques.

La quatrième et dernière catégorie prosodique analysée est le syntagme intonatif, détecté automatiquement sur la base des pauses comme détaillé en 4.2.2.. Nous nous sommes arrêtés sur ce choix puisqu'il a été montré que la présence de pauses est un facteur important pour signaler la réalisation d'un syntagme intonatif (Jun et Fougeron, 2000). Une détection de la forme du contour final de f0 (montant/descendant) a également été effectuée dans le but de ne prendre en compte que les syntagmes ayant un contour montant. Cette méthode permet ainsi de les distinguer d'une position finale d'énoncé ayant un contour descendant. Aucune précaution de cet ordre n'a pu être effectuée pour les positions initiales. Nous sommes encore une fois conscients que la détection automatique de cette catégorie peut engendrer un certain nombre de détections erronées. Une explication semblable à celle évoquée pour les syntagmes accentuels sera proposée ici : l'objectif de cette étude est d'obtenir quatre catégories prosodiques, aussi proches que possible de celles mentionnées dans la littérature, sans se baser sur des caractéristiques prosodiques d'allongement. Une analyse supplémentaire avec des valeurs de durée et de f0 permettra aussi de tester les quatre catégories prosodiques, confirmant ainsi la fiabilité de ces catégorisations. La très large quantité de données utilisée pourrait également aider à compenser les éventuelles erreurs de catégorisation.

L'utilisation de grands corpus de parole continue apporte un certain nombre d'avantages tel qu'un nombre très important d'occurrences produites de façon naturelle. Cependant, par rapport à un corpus de parole contrôlée, un certain nombre de paramètres comme par exemple le contexte segmental et morpho-syntaxique ne peuvent être contrôlés, et alors que certains contextes sont extrêmement fréquents, d'autres au contraire sont plus rares.

Pour contourner ces problèmes, nous avons eu recours à la mesure normalisée de ralentissement introduite précédemment (Gendrot et al., 2012 ; voir section 4.1.6) qui permet de prendre en compte le contexte segmental pour la durée (tandis que les valeurs de f0 sont moins influencées par le contexte segmental, et donc une telle mesure apparaît moins indispensable). Les autres variables tels que contexte consonantique pour les mesures de formants, la catégorie grammaticale du mot portant le phonème analysé sont pris en compte au moyen d'un modèle mixte à effets aléatoires, où les mots et les contextes consonantiques regroupés en quatre catégories (comme précédemment : labial, dental, palato-vélaire, uvulaire) sont insérés comme variables aléatoires. Les résultats qui nous intéressent ici sont non pas l'influence des différents paramètres évoqués ci-dessus sur la réalisation des voyelles, mais l'influence des constituants prosodiques une fois que le poids de ces différents paramètres a été pris en compte.

Comme le montre la Figure 26, les voyelles occupent un espace acoustique de plus en plus important en remontant la hiérarchie prosodique. Les mesures de dispersion présentées par la Figure 27 ci-dessous révèlent que les valeurs augmentent globalement avec le niveau de la hiérarchie prosodique. Seules les voyelles /i/ et /y/ semblent moins varier.

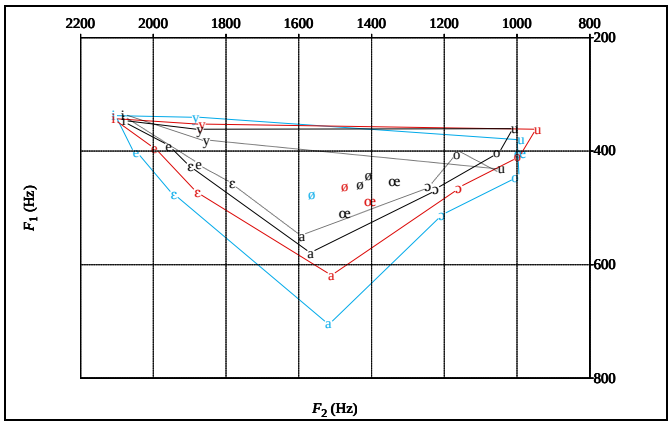


Figure 26 : comparaison des quatre catégories prosodiques en position initiale. De l'intérieur vers l'extérieur : syllabe, mot, syntagme accentuel et syntagme intonatif, d'après Gendrot et al. (2015)

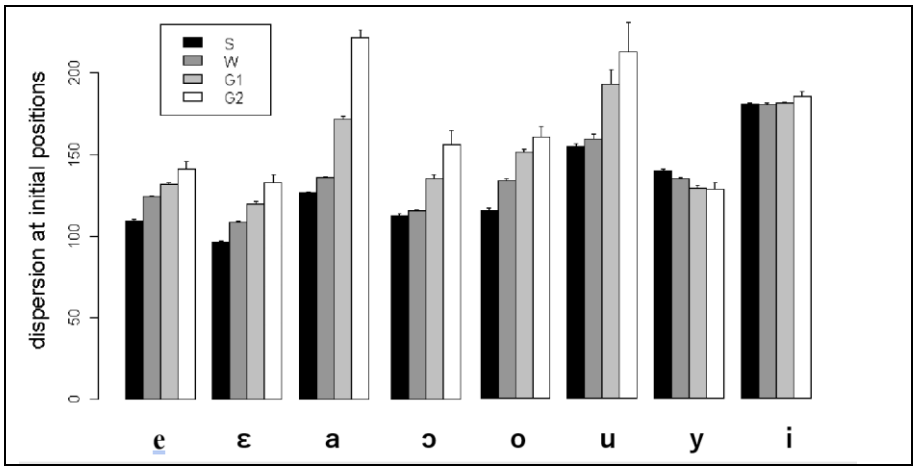


Figure 27 : dispersion en position initiale en fonction de la hiérarchie prosodique (S: syllabe, W: mot, AP: syntagme accentuel, et IP: syntagme intonatif, d'après Gendrot et al. (2015)

L'analyse statistique des mesures de dispersion a permis d'établir que seul /a/ affiche des différences significatives pour chaque niveau. Les voyelles /ɛ/, /e/, /o/ montrent une dispersion significative entre trois niveaux sur les quatre. Pour /ɔ/ et /u/, seuls deux niveaux peuvent être distingués significativement, dans les deux cas, les niveaux syllabe/mot vs. syntagme accentuel/intonatif. Pour la voyelle /i/, seul le dernier niveau (groupe intonatif) a une valeur de dispersion plus élevée que les autres. /y/ révèle une tendance inverse à celle observée pour les autres voyelles. Nous avons montré dans une étude précédente (Gendrot et al., 2008) que /i/ et /y/ révèlent moins de variation dans le plan F1/F2 que les autres voyelles, car les variations pour ces voyelles en français sont ciblées sur F3 et F4. Pour /i/, F4 baisse significativement à mesure que l'on monte dans la hiérarchie prosodique (mais pas F3). Ces variations sur F3 et F4 permettent à /i/ et /y/ d'être plus focales en rapprochant les formants F3/F4 et F2/F3 respectivement (Gendrot et al., 2008 ; Schwartz et al., 1997). En termes articulatoires, /i/ serait plus étiré à mesure qu'il monte dans la hiérarchie prosodique alors que /y/ serait plus arrondi, ce qui accroit leurs caractéristiques articulatoires.

Une rapide analyse des valeurs de durée et de f0 (non présentées ici) montre que les deux paramètres augmentent avec le niveau de la hiérarchie prosodique, ce qui confirme que les phonèmes sont hyper-articulés, non seulement au niveau spectral mais également au niveau prosodique. Ces résultats

prosodiques confirment également la fiabilité de notre détection automatique de catégories prosodiques.

4.2.3.1 Positions finales

Comme observé pour les positions initiales en Figure 26, la Figure 28 montre que les voyelles occupent un espace acoustique de plus en plus important en remontant la hiérarchie prosodique pour les positions finales (Gendrot et al, 2015). Il semble encore une fois que les variations observées pour /i/ sont plus faibles que celles observées pour les autres voyelles, tandis que /y/ est caractérisé par des variations erratiques. On remarque toutefois qu'en position initiale, c'est essentiellement F1 qui est impacté alors qu'en position finale, on a une expansion combinée de F1 et de F2. Les mêmes modèles statistiques que ceux présentés pour les positions initiales ont été effectués, et montrent que les voyelles /e/, /a/ et /o/ révèlent des différences significatives pour tous les niveaux. /ɛ/, /i/ et /u/ montrent quant à elles une variation de dispersion significative pour trois niveaux sur les quatre. Nous pouvons noter que les valeurs F du test statistique sont plus importantes avec plus de différences significatives que pour les positions initiales, ce qui suggère une hyper-articulation plus importante en position finale. Les variations de F3 et F4 ont également été analysées pour les voyelles /i/ et /y/ comme dans la section précédente : comme observé pour les positions initiales, pour /y/, à mesure que l'on monte dans la hiérarchie prosodique, F3 baisse significativement (mais pas F4). Pour /i/, F4 baisse significativement à mesure que l'on monte dans la hiérarchie prosodique (mais pas F3).

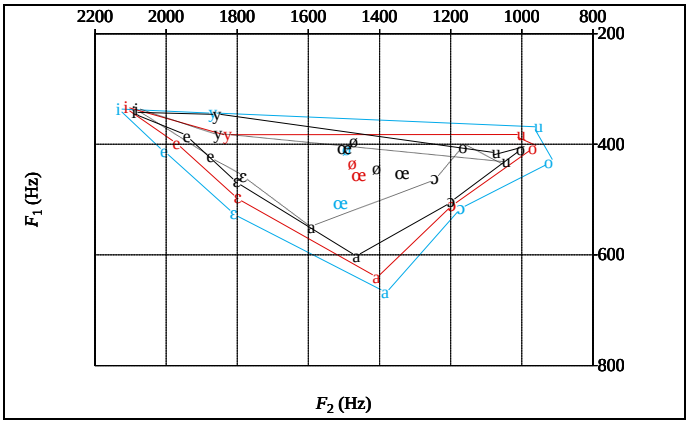


Figure 28 : comparaison des quatre catégories prosodiques en position finale. De l'intérieur vers l'extérieur : syllabe, mot, syntagme accentuel et syntagme intonatif, d'après Gendrot et al. (2015)

Ici encore, les valeurs de durée et de f0 révèlent que ces deux paramètres augmentent avec le niveau de hiérarchie prosodique. Comme noté pour les mesures de dispersion, l'amplitude de la variation entre les niveaux est plus importante que celle observée pour les positions initiales.

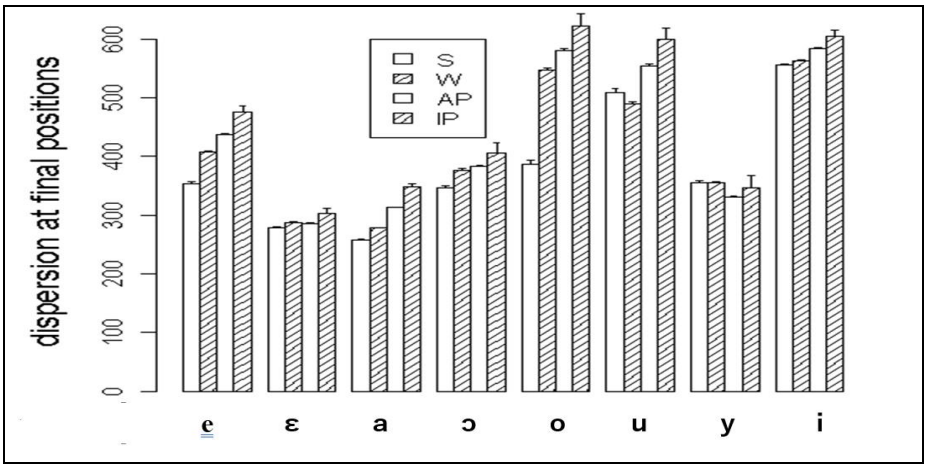


Figure 29 : dispersion en position finale en fonction de la hiérarchie prosodique (S: syllabe, W: mot, AP: syntagme accentuel, et IP: syntagme intonatif, d'après Gendrot et al. (2015)

4.2.3.2 Discussion de la partie hiérarchie prosodique

Comme nous en avions fait l'hypothèse, nous observons une hiérarchie prosodique (de la syllabe au mot, puis du syntagme accentuel jusqu'au syntagme intonatif) sur la base de mesures spectrales et de mesures prosodiques (de durée et de f0). Cependant, comme signalé par les études précédentes dans ce domaine (Fougeron, 2001 ; Tabain, 2003, pour le français), tous les niveaux ne peuvent pas être distingués de façon systématique. Une explication possible concerne la détection automatique des catégories prosodiques. Nous avons pu remarquer, principalement pour les positions initiales, que les différences non significatives portaient sur les deux plus hautes catégories de la hiérarchie prosodique (respectivement le syntagme accentuel et intonatif). Le syntagme intonatif est la catégorie la moins représentée en termes de fréquence d'occurrence, favorisant ainsi une erreur standard plus élevée : les tendances peuvent aller dans la direction attendue avec des valeurs pour le syntagme intonatif plus élevées que pour le syntagme accentuel mais sans atteindre systématiquement le seuil de significativité (2 sur 8 pour les positions initiales ; 5 sur 7 pour les positions finales). La première question qui se pose concerne la certitude que les catégories automatiquement détectées et intitulées 'syntagme accentuel' et 'syntagme intonatif' correspondent réellement à ce que l'on attend. Les mesures prosodiques également effectuées semblent le confirmer. La question suivante concerne les paramètres à analyser pour tester ces catégories prosodiques. L'hyper-articulation observée concerne-t-elle strictement les valeurs de durée et de f0 qui influencent à leur tour la réalisation spectrale de la voyelle, ou bien la réalisation spectrale pourrait être - au moins en partie - indépendante et liée à un renforcement effectué dans l'organisation de la parole en syntagmes (Gendrot et al., 2015).

Nous avons remarqué que les valeurs de durée et de f0 augmentaient parallèlement aux valeurs de dispersion, en remontant la hiérarchie prosodique. Il semblait dans un premier temps que ces paramètres étaient liés puisque la f0 et la durée sont connus pour marquer la présence de frontières. Des mesures de corrélations ont été effectuées mais comme précisé par Keating pour l'anglais américain (Keating et al., 2004) elles se sont révélées faibles, ce qui pourrait suggérer que des stratégies tantôt spectrales ou tantôt prosodiques pourraient exister pour marquer la présence de frontières (en position initiale : r=0.21 entre les mesures de dispersion et de durée et r=0.26 entre les mesures de dispersion et de f0 ; en position finale : r=0.18 entre les mesures de dispersion et de durée et r=0.27 entre les mesures de dispersion et de f0). Si les paramètres de durée et de f0 qui sont habituellement considérés comme des marqueurs de catégories prosodiques, ne sont pas fortement

corrélés aux mesures de dispersion, cela suggère qu'il pourrait y avoir des phénomènes de compensation utilisées par les locuteurs entre les variations spectrales et les variations prosodiques pour marquer les frontières de constituants prosodiques. La relative indépendance des paramètres prosodiques et spectraux est confirmée par le point suivant : de manière à ne retenir que les groupes intonatifs et laisser de côté les fins d'énoncés, nous avons mesuré les contours de f0 sur les voyelles analysées et n'avons conservé pour la suite des analyses que les contours montants pour la catégorie des groupes intonatifs. Une comparaison des catégories ainsi filtrées avec les catégories restantes a montré que les valeurs de dispersion étaient équivalentes alors que les valeurs de durée étaient plus élevées pour les voyelles ayant un contour montant. Ceci confirme que l'hyper-articulation peut être indépendante des valeurs de durée, mais aussi - comme le suggère Fougeron (2001) pour le français - qu'il y a peu sinon pas de différences articulatoires entre le groupe intonatif et le niveau énoncé (Gendrot et al., 2015).

Ces résultats nous amènent à un nouveau questionnement à propos du choix des catégories prosodiques à détecter. Y-a-t-il un nombre limité de catégories prosodiques ou pourrait-on détecter un continuum graduel de catégories ? La longueur (en phonèmes, mais aussi en millisecondes) de chaque catégorie augmente en montant la hiérarchie prosodique (syllabe, mot, groupe accentuel, groupe intonatif). Si la longueur de la catégorie était en fait un facteur favorisant l'hyper-articulation, nous pourrions observer un continuum de catégories prosodiques plutôt qu'un nombre fini. Pour ce faire, nous avons utilisé une nouvelle fois l'algorithme de chunking en modifiant le nombre maximum de syllabes. La figure 18 ci-dessous nous montre la durée du dernier phonème en fonction de la catégorie prosodique (syllabe, mot, groupe accentuel, groupe intonatif), mais en faisant varier la nombre de syllabes maximalement autorisé par l'algorithme de chunking pour le groupe accentuel (de 5 à 10 syllabes). On s'aperçoit que la variation à l'intérieur de la catégorie groupe accentuel est très faible, ce qui corrobore l'existence d'un groupe accentuel sous-jacent non dépendant de sa durée.

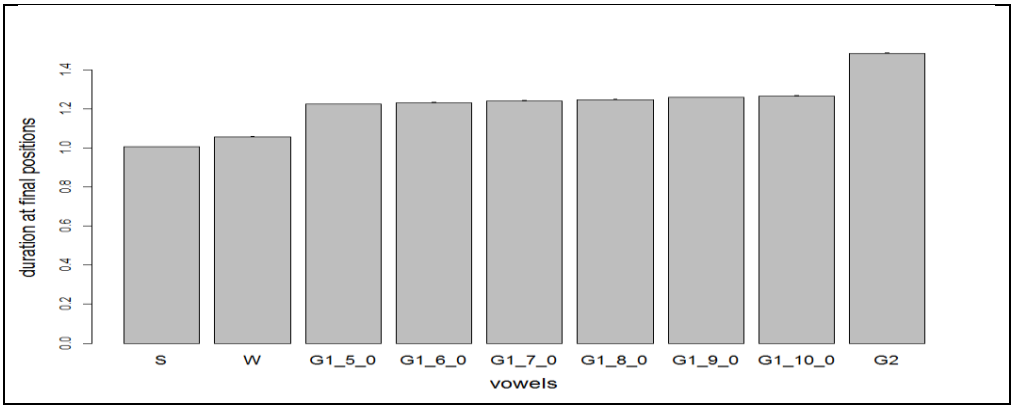


Figure 30 : valeurs de f0 en position finale en fonction de la hiérarchie prosodique (S: syllabe, W: mot, AP: syntagme accentuel, et IP: syntagme intonatif (d'après Gendrot et Gerdes, communication personnelle)

Dans cette étude, aucune sélection n'a été faite sur les catégories grammaticales des mots portant les voyelles analysées, mis à part la prise en compte du mot en tant que variable aléatoire dans nos modèles statistiques. Or, la distribution des catégories grammaticales n'est pas équilibrée, principalement à cause des débuts de groupes accentuels ou de groupes intonatifs plus fréquemment composés de prépositions, conjonctions ou déterminants, alors que pour les deux autres catégories

prosodiques (syllabe et mot), les noms et les adjectifs sont les plus représentés. Il pourrait être attendu que les mots grammaticaux soient moins hyper-articulés, puisqu'ils sont plus fréquents et qu'ils portent moins d'information cruciale dans la parole (une abondante littérature existe sur la relation entre contenu informatif des mots, ainsi que leur fréquence lexicale, et la réalisation des sons contenus dans ces mots ; voir la théorie de l'hypo- et hyper-articulation de Lindblom (1990) et Wright (2004) par exemple). On aurait donc pu s'attendre à des débuts de groupes accentuels et de groupes intonatifs avec une hyper-articulation moindre que les débuts de mots par exemple. Les différences notées entre les variations sur les positions initiales et les positions finales pourraient corroborer ce fait. Cependant, il convient de se rappeler que les positions initiales de chaque catégorie prosodique correspondent également à la fin d'une catégorie de même niveau. Comme proposé par Byrd et Saltzman (2003), ces frontières sont des moments de ralentissement articulatoire ('pi-gesture') qui favorisent l'hyper-articulation. Il n'est donc pas si surprenant d'observer les mêmes résultats et phénomènes d'hyper-articulation en position initiale et finale de tous nos niveaux prosodiques, et ce quelle que soit la catégorie grammaticale impliquée. Si le locuteur choisit de signaler une frontière prosodique dans sa production, on peut émettre l'hypothèse qu'il le fera quelle que soit la catégorie grammaticale des mots autour de cette frontière (Gendrot et al., 2015).

Nous avons également testé (Gendrot et al., 2015) l'analyse de voyelles en deuxième position de début ou de fin de constituant prosodique (respectivement en deuxième position ou en position pré-finale), et l'empan semble plus important sur les positions finales que sur les positions initiales. L'hyper-articulation observée sur la voyelle pré-finale est très proche de celle observée sur la voyelle finale, alors que l'hyper-articulation de la voyelle en position initiale stricte est significativement plus importante que l'hyper-articulation de la voyelle en deuxième position (post-initiale). Encore une fois, les différences notées entre les variations sur les positions initiales et les positions finales pourraient corroborer ce fait.

Le dernier point de cette discussion (Gendrot et al., 2015)concerne l'impact de ces résultats sur la nature du renforcement phonétique dans la parole dans les positions mises en valeur par les locuteurs : s'agit-il d'expansion de la sonorité (Beckman, 1992) ou d'hyper-articulation des voyelles (De Jong, 1995) ? Ces deux hypothèses peuvent être confrontées en observant les valeurs de F1 pour les voyelles fermées. En effet, dans le cas de l'expansion de sonorité, un renforcement des voyelles basses aura pour effet une augmentation de la valeur de F1, alors que pour l'hypothèse de l'hyper-articulation, un renforcement des voyelles basses aura pour effet un abaissement de la valeur de F1. L'hypothèse de l'hyper-articulation prédit que les traits des voyelles renforcés seront accrus, par exemple une voyelle basse sera encore plus basse, une voyelle arrondie sera encore plus étirée, etc. Pour les autres voyelles, ces deux hypothèses auront des prédictions semblables avec une augmentation des valeurs de F1 pour les voyelles renforcées. Rappelons que pour les voyelles fermées, F1 n'est pas strictement lié à l'aperture car il s'agit d'une résonance de Helmholtz, qui s'abaisse en fonction d'une série de manœuvres articulatoires telles que l'abaissement du larynx, l'avancement de la langue, l'arrondissement des lèvres ou l'aperture. La Figure 26 et la Figure 28 montrent clairement un cas d'hyper-articulation : à mesure que les voyelles basses gagnent en longueur, leur premier formant s'abaisse. Les résultats observés sur F3 et F4 pour /i/ et /y/ corroborent également cette hypothèse de l'hyper-articulation des voyelles renforcées, puisque /i/ une fois renforcé voit son troisième formant augmenter pour favoriser l'étirement, alors que /y/ voit son troisième formant baisser pour favoriser l'arrondissement.

4.2.4 Comparaisons multilingues (2/2)

Les travaux effectués sur la variation spectrale en fonction de la prosodie posent la question de la spécificité de ces résultats obtenus sur le français. Les résultats sont-ils dépendants du système phonémique et/ou prosodique du français ou bien peuvent-ils être confirmés sur d'autres langues ? Les mesures effectuées dans la section 4.1.5 sur la réduction spectrale en fonction de la durée montraient des résultats similaires entre les sept langues analysées si l'on s'en tient aux aspects généraux, mais nous avons pu montrer la spécificité du troisième formant pour le français. Dans le but d'affiner nos recherches, nous avons voulu analyser les réalisations spectrales des voyelles en fonction des spécificités prosodiques de l'espagnol et de l'allemand. Pour nous, l'objectif était de mieux comprendre les sources de l'hyper-articulation, celle-ci étant due à une meilleure réalisation pour des raisons accentuelles, ou bien à un marquage prosodique pour que les frontières de mots et de constituants soient mieux perçues.

4.2.4.1 Position morphologique en français et en espagnol

Dans cette section, nous comparons deux langues romanes, le français et l'espagnol (Gendrot et al. 2017). L'espagnol est une langue à accent lexical et tous les mots pleins ont au minimum un accent primaire sur l'une des trois dernières syllabes (Navarro Tomás, 1944 ; Quilis, 1981). On peut observer des mots oxytoniques tels que *camaleón (« caméléon »),* des mots paroxytoniques tels que *cabeza (« tête »)* et pour finir des mots pré-paroxytoniques tels que *América (« Amérique »)*. Le français n'a pas d'accent lexical tel que décrit pour l'espagnol ; il est plus volontiers porté par le groupe prosodique et a un rôle démarcatif dans la parole (Jun et Fougeron, 2000). L'accent est positionné sur la fin des groupes accentuels selon la terminologie de Jun et Fougeron, également nommés groupes phonologiques par Post (2000) ou mots prosodiques par Hirst et Di Cristo (1998), etc. En espagnol, le domaine prosodique est le mot prosodique, comparable au groupe accentuel pour le français (Hualde, 2012). Cependant, ces deux groupes n'ont pas la même fonction dans les deux langues. Alors qu'elle est démarcative en français et aide à la discrimination des frontières de syntagmes, elle est culminative en espagnol et aide à la perception des mots. Dans ce cadre, si la dernière syllabe des groupes accentuels est théoriquement accentuée en français, elle ne l'est pas en espagnol. Cette différence nous a permis de tester les syllabes finales de mots dans les deux langues et d'en observer les spécificités, ce qui nous permettra de dissocier les aspects dus à l'accent et ceux dus à la prosodie.

Un corpus de parole de journalistique a été utilisé dans les deux langues. Pour l'espagnol 13 heures de parole ont été analysées, et pour le français, 13 heures de parole ont été aléatoirement choisies parmi le corpus ESTER détaillé précédemment et les mêmes méthodes ont été utilisées. Seuls les mots dissyllabiques et trisyllabiques ont été analysés. La variation de dispersion acoustique (i.e. la distance euclidienne) en fonction de la durée des voyelles est comparable pour les deux langues (cf. Figure 31), même si elle plus importante en espagnol pour les voyelles antérieures fermées, rejoignant en cela les résultats obtenus dans la section 4.1.4.3.

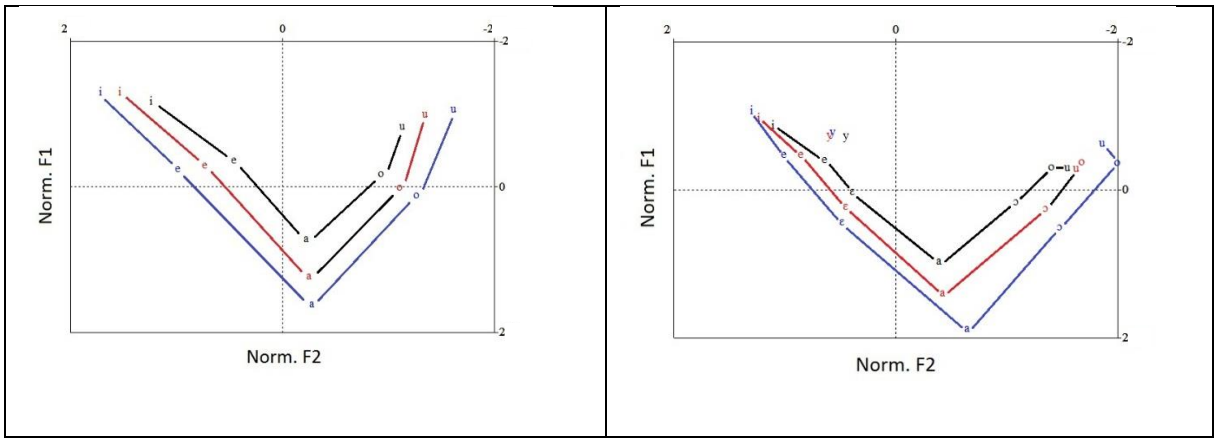


Figure 31 : formants F1 et F2 des voyelles de l'espagnol (gauche) et du français (droite) en function de la durée vocalique (noir pour les voyelles courtes rouge pour les voyelles intermédiaires, et bleu pour les voyelles longues). Les valeurs de formants sont normalisées à l'aide de la normalisation de Lobanov, d'après Gendrot et al. (2017)

Dans un deuxième temps, nous avons analysé la réalisation des voyelles en fonction de l'accentuation et de la position dans le mot. Pour les mots dissyllabiques en français (non illustré), les voyelles en syllabe finale sont caractérisées par une dispersion acoustique plus importante ($p<0.0001$), une durée plus élevée ($p<0.0001$) et une f0 plus élevée ($p<0.0005$) que les voyelles en syllabe initiale. Pour l'espagnol (Figure 32), nous avons également ajouté l'accent comme variable et avons comparé les syllabes initiales non accentuées aux syllabes finales accentuées, se rapprochant ainsi du français. La Figure 32 montre que si les syllabes finales ont une durée et une f0 plus élevées (respectivement $p<0.0001$ et $p<0.006$), leur dispersion acoustique n'est pas significativement différente ($p=0.24$).

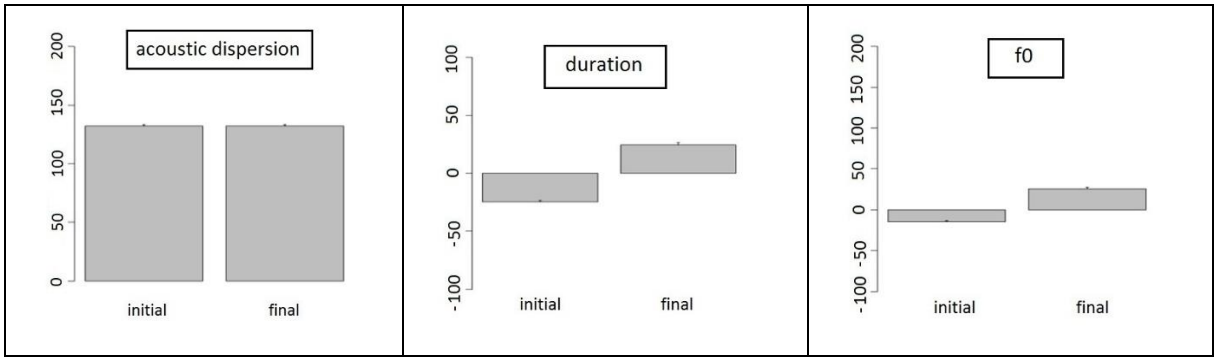


Figure 32 : dispersion acoustique, durée, et f0 normalisées des voyelles en syllabe initiale non accentuée vs. syllabe finale accentuée dans les mots dissyllabiques en espagnol, d'après Gendrot et al. (2017)

A l'inverse en espagnol, toujours pour les mots dissyllabiques, les voyelles en syllabe finale non accentuée ont une durée plus longue ($p<0.01$) mais une f0 et une dispersion acoustique plus basse (respectivement $p<0.005$ et $p<0.001$) que les voyelles en syllabe initiale accentuée. Une analyse croisée de ces deux facteurs a confirmé ces résultats, ce qui montre que la dispersion acoustique et la durée peuvent être décorrélées, comme proposé en section 4.2.3.2. Les informations spectrales pourraient donc être utilisées pour marquer l'accentuation lexicale, alors que les informations prosodiques seraient utilisées pour délimiter les mots. Or, ces informations sont impossibles à analyser séparément en français de par leur statut cumulé. Une analyse des voyelles pré-pausales a quant à elle pu montrer une dispersion acoustique, ainsi que des valeurs de durée et de f0, plus importantes dans les deux

langues par rapport à des voyelles non pré-pausales. Ce résultat peut s'interpréter par l'observation dans les deux langues d'un renforcement des voyelles dans une position prosodique de haut niveau tel que celui généré par la pause, généralement la fin de groupe intonatif. Nous avons pu montrer dans cette étude qu'un renforcement vocalique pouvait être observé dans les deux langues lors d'une augmentation de la durée, mais ce sont surtout les variations prosodiques de durée, et non les variations accentuelles, qui génèrent ce renforcement (Gendrot et al., 2017).

4.2.4.2 Réduction vocalique en fonction de la position prosodique en allemand et en français

Dans la lignée de ces travaux, nous avons appliqué l'analyse de la variation spectrale en fonction des constituants prosodiques à l'allemand (Gendrot et al., 2012). L'allemand étant une langue germanique avec un accent lexical, elle permet une nouvelle fois de vérifier la spécificité du français de par son absence d'accent lexical. Le niveau du groupe accentuel pourrait particulièrement être concurrencé en termes d'hyper-articulation par le niveau lexical comme cela a été le cas pour l'espagnol. Nous avons eu recours au corpus ESTER pour le français, et à 20h de parole journalistique issus de journaux télévisés obtenus sur la chaine ARTE, transcrits, alignés et analysés sur les mêmes principes que précédemment. Les détections des catégories prosodiques pour le français sont les mêmes que celles présentées en section 4.2.3. De même que pour l'espagnol, les niveaux de la syllabe, du mot, et du syntagme intonatif ont été obtenus sur la base de la segmentation en phonèmes. Quant au découpage en syntagmes, un nouvel algorithme a été développé pour calquer celui créé pour le français. Nous avons choisi une limite de sept syllabes par syntagme comme pour le français pour partir d'un point de référence identique. Cette limite pourrait être discutable puisque les mots lexicaux sont globalement plus longs en allemand qu'en français, mais le principe de la comparaison toutes choses égales par ailleurs a été préféré.

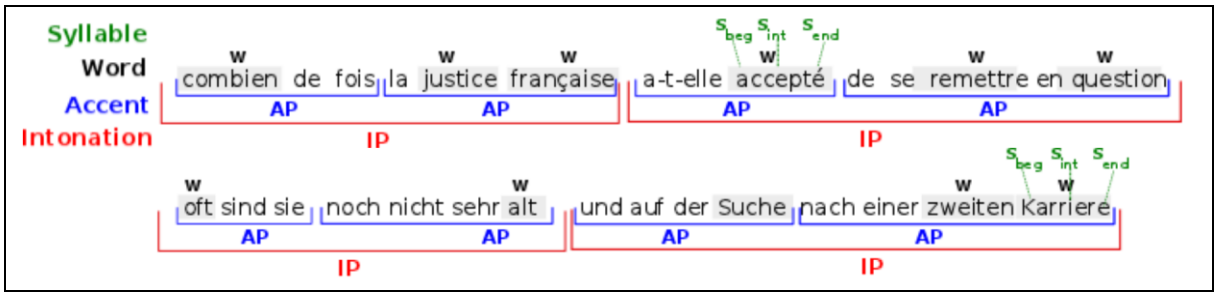


Figure 33 : illustration des quatre niveaux prosodiques pour le français et l'allemand : syllabe (S), mot (W), groupe accentuel (AP) et groupe intonatif (GI). Seuls les mots lexicaux sont pris en considération, d'après Gendrot et al. (2012)

Comme indiqué par la Figure 34 et la Figure 35, les valeurs de durée et de distance depuis le centre acoustique augmentent globalement pour chaque niveau de la hiérarchie prosodique dans les deux langues.

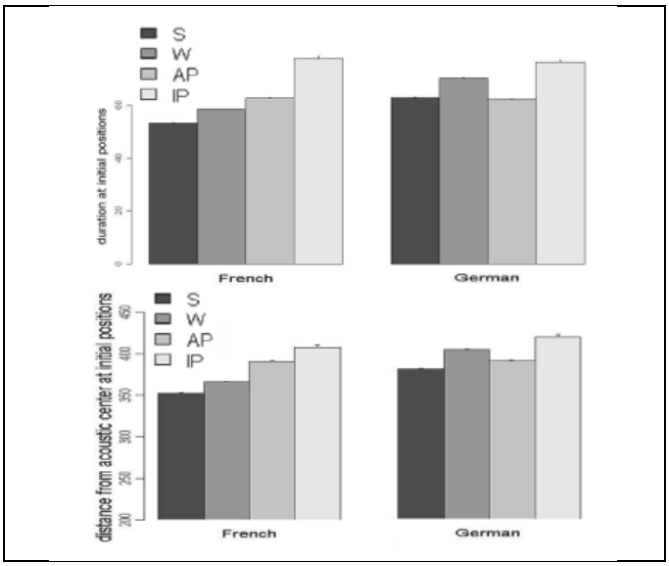


Figure 34 : durée et distance du centre acoustique en position initiale de constituant prosodique (S : syllabe ; W : mot ; AP : groupe accentuel ; IP : groupe intonatif), d'après Gendrot et al. (2012)

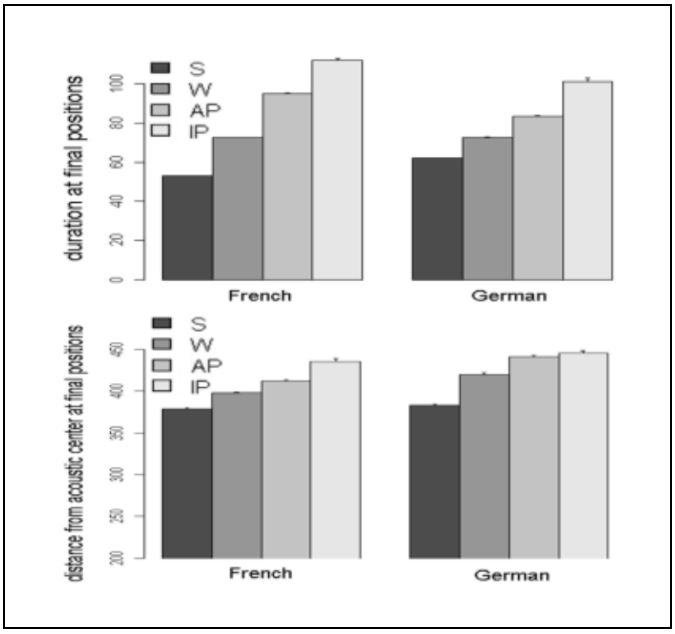


Figure 35 : durée et distance du centre acoustique en position finale de constituant prosodique (S : syllabe ; W : mot ; AP : groupe accentuel ; IP : groupe intonatif), d'après Gendrot et al. (2012)

L'allemand cependant se différencie du français par la catégorie prosodique « mot » qui a des valeurs plus élevées en position initiale, rompant ainsi la hiérarchie progressive entre les quatre catégories. L'accentuation lexicale en allemand, majoritairement sur la syllabe initiale implique également un allongement et un renforcement de la voyelle placée dans cette syllabe. La Figure 36 montre effectivement le renforcement (sous la forme de distance acoustique par rapport au centre de l'espace vocalique) en fonction de la position de la syllabe dans le mot dans les deux langues. La syllabe initiale en français peut être marquée par accent d'intensité (rythmique) notamment dans le cadre de la parole journalistique, mais cette caractéristique est beaucoup plus marquée en allemand.

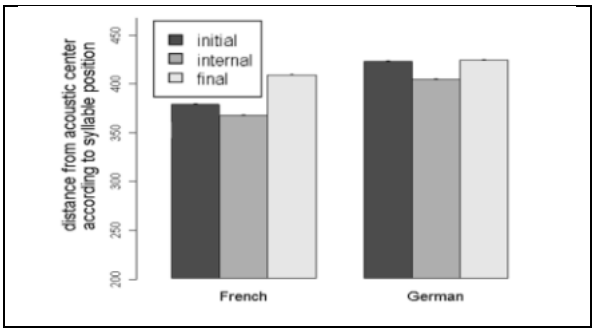


Figure 36 : distance acoustique par rapport au centre de l'espace vocalique en fonction de la position de la syllabe dans le mot en français et en allemand, d'après Gendrot et al. (2012)

4.2.4.3 Mesures de déclinaison interlangues

Dans cette courte section, je présente un travail qui a été effectué en collaboration avec Carolin Schmid lors d'un stage qu'elle a effectué en France (Gendrot et Schmid, 2011 ; Schmid et al., 2012). Le but de notre travail commun était de comparer la ligne de déclinaison en allemand et en français. La déclinaison de la f0 est définie comme la tendance globale de la fréquence fondamentale à baisser au cours d'une séquence, entre une ligne supérieure reliant ses pics locaux et une ligne inférieure reliant ses vallées locales qui baissent également. Un resetting de la f0 a lieu au début de chaque nouvelle séquence (cf. 't Hart et al., 1990, à voir dans la Figure 37).

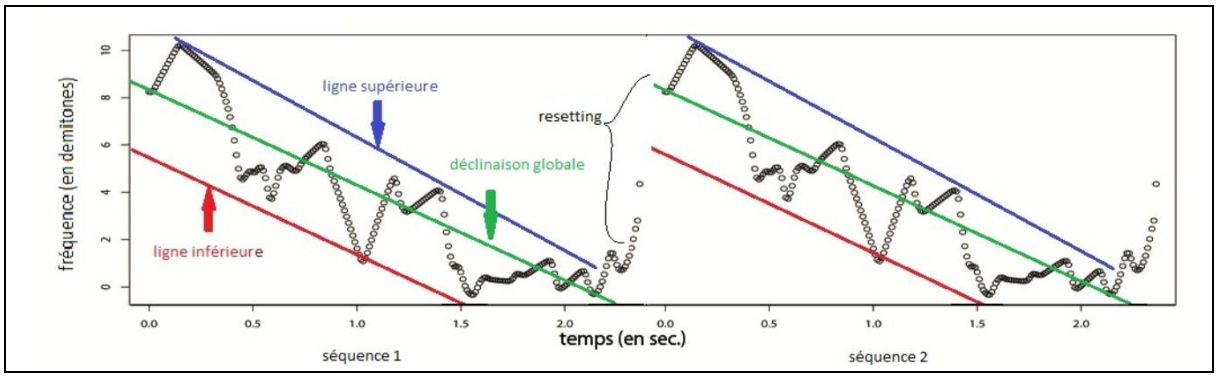


Figure 37 : tendance de mouvement de la f0 : baisse globale au cours d'une séquence, mouvements descendants et montants au niveau local entre deux lignes descendantes globalement (ligne supérieur et inférieur), resetting de la f0 entre deux séquences, d'après Schmid et al. (2012)

La tendance globale de la f0 à décliner peut être liée à plusieurs facteurs : à la pression sous-glottique (Lieberman, 1967), à la traction de la trachée (Maeda, 1976) et aux mouvements des muscles laryngés (Ohala et Ewan, 1973). Pourtant certaines incertitudes subsistent : l'aspect de la déclinaison est-il dépendant de la langue ou est-il contrôlé par le locuteur ? Les difficultés à définir plus précisément la nature de la déclinaison sont liées au fait qu'il est délicat d'observer la déclinaison pure. La courbe globale de la f0 est constituée de mouvements de différents niveaux prosodiques (Fujisaki, 1988), ce qui fait que la déclinaison est souvent masquée par d'autres facteurs comme des composants de l'accentuation ou des phrases (montée de continuation, resetting, montée ou descente finale) ou des facteurs microprosodiques. Nous avons tenté dans une petite étude de tester cet aspect au sein de grands corpus. L'objectif était double : dans un premier temps, relever le challenge de détecter des séquences de parole et leur ligne de déclinaison de façon automatique, et deuxièmement vérifier si la

quantité de données de grands corpus de parole non contrôlée permettait d'ajouter un regard nouveau à la littérature. Afin de mesurer automatiquement la ligne de déclinaison, nous avons procédé de la façon suivante. Des valeurs de f0 ont été extraites toutes les 10 ms sur les positions des segments voisés. Nous avons également calculé la valeur du 'resetting' qui représente la différence (en st) entre la première valeur de la f0 d'une séquence et la dernière valeur de la f0 de sa séquence précédente. Les courbes de la f0 ont été optimisées en les interpolant afin d'obtenir un contour continu (interrompu précédemment par les segments non voisés), puis en les filtrant par un filtre passe-bas, et en convertissant les valeurs en Hertz en valeurs en demi-tons (st) par la formule suivante : $st = 12 \times \log x_2 \frac{\text{f0}}{\text{f0} - 5^{\text{ème}} \text{ quantile}}$

Nous avons choisi comme fréquence de base pour chaque séquence le 5ième quantile de la fréquence moyenne de toutes les séquences d'un même locuteur. Pour mesurer la déclinaison nous avons d'abord calculé la ligne de régression globale par modélisation des moindres carrés ('ordinary least square modelling') pour la séquence entière (à voir dans la Figure 38, à gauche) ainsi que pour la partie médiane de la phrase (après avoir retiré les premières et dernières 500 ms des séquences). Ensuite les positions des pics et des vallées du contour de la f0 ont été mesurées par l'algorithme 'convex-hull' afin d'obtenir la ligne supérieure et la ligne inférieure du contour unique de la f0 d'une séquence (voir Figure 38, à droite). Afin de retrouver une éventuelle tendance propre à chaque langue, les pics et les vallées d'une séquence ont été moyennés respectivement à 6 points relatifs : les valeurs figurant à un temps jusqu'à 10 % de la durée de la séquence, celles figurant entre 10 et 30 %, entre 30 et 50 %, entre 50 et 70 %, entre 70 et 90 % et entre 90 et 100 %.

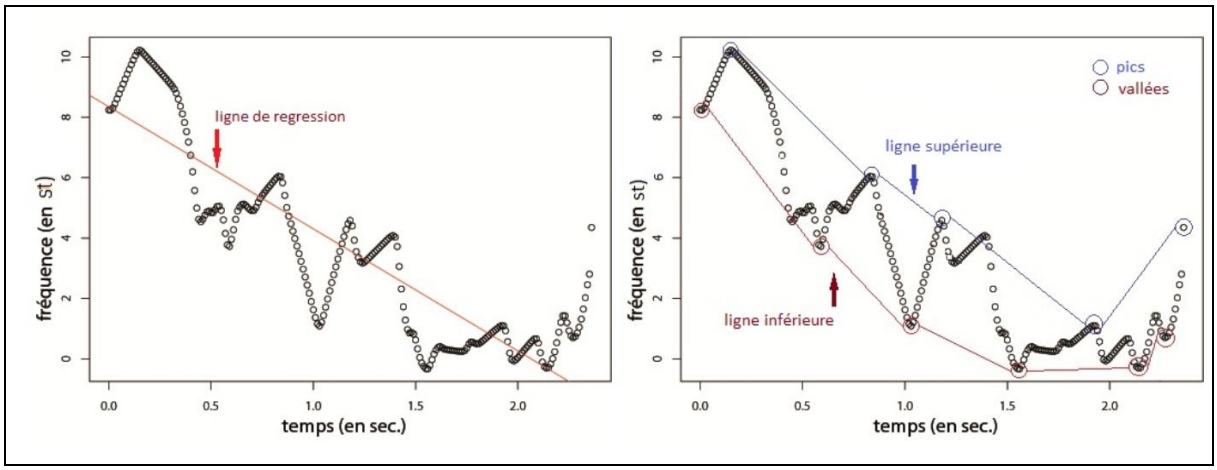


Figure 38 : à gauche : la ligne de régression calculée sur le contour unique des valeurs de la f0 d'une séquence. à droite : les ligne supérieures et inférieures reliant les pics et les vallées du contour unique des valeurs de la f0 d'une séquence, détectées à partir de l'enveloppe convexe, d'après Schmid et al. (2012)

Le corpus français consistait initialement en 46 630 séquences d'une durée moyenne de 1.75 secondes et en allemand de 33 394 séquences avec une durée moyenne de 1.4 secondes. Nous avons décidé de baser nos analyses également sur les séquences d'une durée d'entre 1 et 4 secondes et avec une pente négative. Ce choix se justifiait en outre par nos propres observations : les séquences d'une durée inférieure à 1s montrent des valeurs de régression extrêmes et les séquences d'une durée supérieure à 4s peuvent éventuellement résulter d'une erreur de segmentation en séquences. Nous avons pu constater des similitudes entre les 2 langues en ce qui concerne le degré des pentes négatives ainsi que des facteurs qui l'influencent. En allemand le degré moyen de la pente sur l'intégralité de chaque

séquence s'élève à -2.5 st/s et en français à -2.4 st/s pour le contour global de f0 sur la séquence complète. Le degré de la pente sur la partie médiane de chaque séquence est de -2.3 st/s en allemand et de -2.4 st/s en français. La corrélation entre la durée de la séquence et le degré de la pente est de $r^2 = 0.4$ pour les deux langues : plus la séquence est courte, plus sa pente négative est raide. Entre la valeur de l'ordonnée à l'origine ('intercept') et le degré de la pente, le coefficient de corrélation s'élève à $r^2 = 0.6$ pour les deux langues : plus la valeur de l'ordonnée à l'origine est haute, plus la pente négative est raide.

Une corrélation entre la valeur du resetting au début de la séquence et le degré de la pente ne se montre que pour des valeurs positives du resetting, i.e. à partir de 0 st/s (r2 = 0.2 en allemand et r2 = 0.3 en français) : plus le resetting est important, plus la pente négative est raide. Les valeurs négatives du resetting suggèrent la présence de séquences entre lesquelles la f0 continue à baisser et pour lesquelles aucune corrélation avec le degré de la pente ne peut être constatée dans les deux langues (r2 < 0. 02). Si l'on considère le resetting de la f0 comme marqueur de frontière définissant la séquence, ce résultat montre que notre approche de la notion de la phrase aurait pu être améliorée en se basant à la fois sur les silences et sur la présence d'un resetting positif pour définir les séquences de la parole.

Nous avons également pu observer (Schmid et al.,2012) des caractéristiques propres à chaque langue, et ce particulièrement pour les mouvements des lignes supérieures et inférieures en allemand et en français (Figure 39). En allemand, la ligne supérieure (régression moyenne : -2.5 st/s) est plus raide que la ligne inférieure (2.1 st/s). En français c'est au contraire la ligne inférieure (-2.4 st/s) qui est plus raide en moyenne que la ligne supérieure (-2.1 st/s). Les valeurs de f0 montrent un plus grand registre en français, avec une distance moyenne de 15 st/s entre la plus basse et la plus haute valeur, contre 13 st/s pour l'allemand.

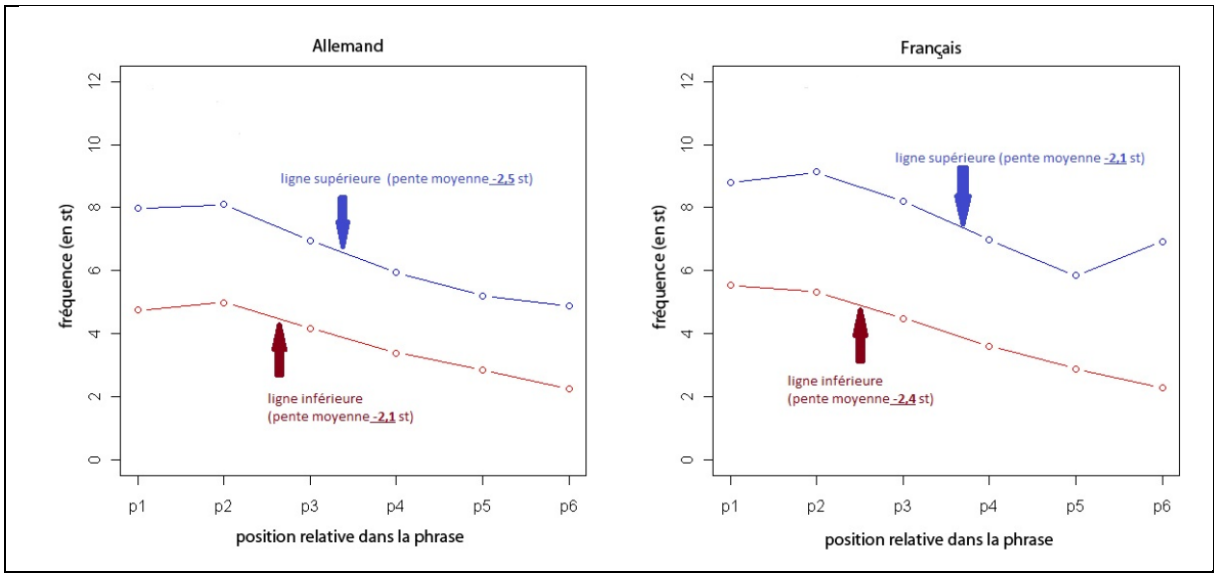


Figure 39 : comparaison des lignes inférieures et supérieures de f0 à l'aide des moyennes des pics et des vallées à des positions relatives dans les séquences en allemand et en français, d'après Schmid et al. (2012)

Si les lignes inférieures ont tendance à baisser dans chacune des langues, la ligne supérieure en français montre une montée finale avec une valeur moyenne de 1.6 st/s ($p < 0.0001$) tandis qu'en allemand cette ligne montre une descente finale avec une valeur moyenne de -0.5 st/s. Cette différence fait référence aux systèmes prosodiques des deux langues déjà évoqués précédemment, le français étant

une langue à frontières (et utilisant des montées finales pour marquer ses frontières), alors que l'allemand est une langue à accent lexical et les montées de f0 étant situées à l'intérieur des séquences. Cette différence dans l'accentuation peut être mesurée par la corrélation entre la durée de la séquence et le nombre des pics de f0. Dans les séquences de même durée se trouvent toujours plus de pics en allemand qu'en français.

Pour conclure cette section, les pentes de ces deux langues présentes de nombreuses similarités en ce qui concerne les lignes de régression du contour global : une pente moyenne d'environ -2.5 st/s comparable à celle trouvée par Yuan et Liberman (2010) pour l'Anglais. Pourtant les mouvements des lignes inférieures et supérieures du contour de f0 semblent fournir des informations différentes. La ligne supérieure apparaît influencée par des mouvements locaux de f0 et correspond au système phono-prosodique de la langue alors que la ligne inférieure serait plus liée aux aspects physiologiques et donc plus universelle (Schmid et al.,2012).

4.2.5 Différences de styles de parole (2/2)

Nous avons étendu ce travail (Gendrot et al., 2012 ; Gendrot et Schmid, 2011 ; Schmid et al., 2012) sur la ligne de déclinaison à deux autres corpus présentés en section 4.1.6 et permettant la comparaison de styles de parole : le corpus NCCFr (parole spontanée) et le corpus ESTER (parole journalistique). Dans les deux corpus, la pente moyenne est fortement dépendante de la longueur de la séquence comme le montre la Figure 40, plus la phrase est longue et plus la pente est mesurée comme faible, et ce particulièrement en parole spontanée (corrélation de Pearson : r^2=0.4 en parole journalistique contre r^2=0.31 en parole spontanée). La valeur moyenne de la pente de la ligne de déclinaison est plus faible en parole spontanée (-2.48 demi-tons/seconde pour la parole journalistique contre 2.25 pour la parole spontanée, différence significative à p<0.0001 pour un test-t).

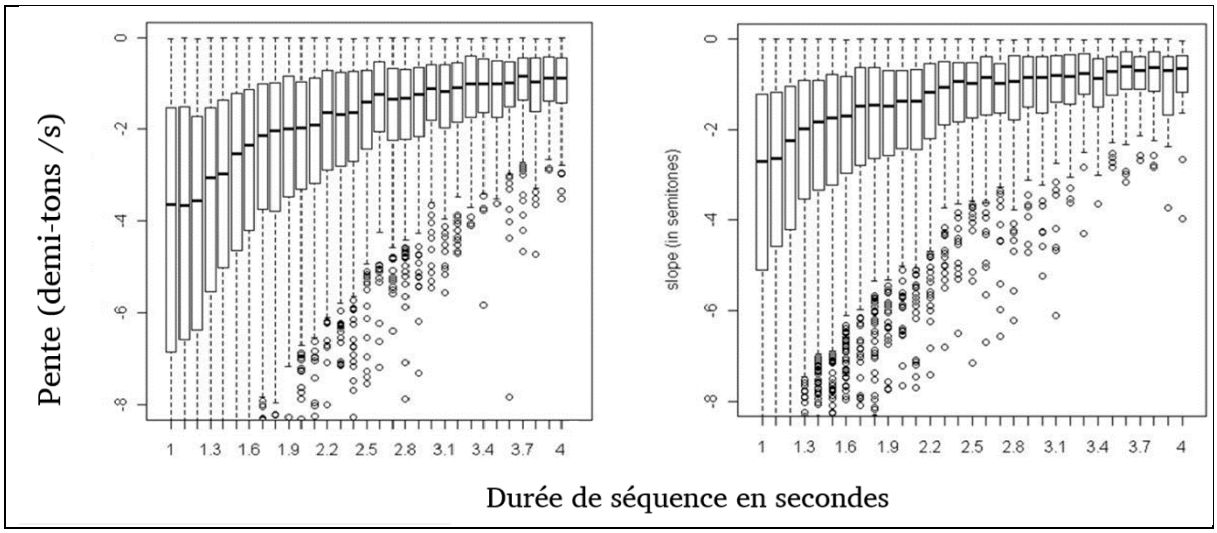


Figure 40 : distribution des pentes moyennes (et écart-type) de la f0 en fonction de la durée de la séquence. A gauche parole journalistique et à droite parole spontanée, d'après Gendrot et al. (2012)

Comme dans la section précédente, la Figure 41 montre le contour de f0 séparé en une ligne supérieure (pics de f0) et une ligne de base (vallées). Quelle que soit la durée des séquences, nous pouvons

observer que les montées de continuation sont faibles voire absentes pour le corpus de parole spontanée (Figure 41 et Figure 42). Pour les phrases inférieures à 2 secondes, les montées de f0 initiales dont le maximum se situe à environ 15 % du début de la séquence sont présentes bien que moins amples en parole spontanée. En analysant les séquences de durée croissante (de 1 à 2 secondes, puis 2 à 3 secondes, etc.), nous pouvons remarquer que la f0 (ligne supérieure et ligne de base) voit sa ligne de déclinaison relevée (plus plate) à partir de la moitié de la séquence pour les séquences de plus de 3 à 4 secondes. Nous émettons l'hypothèse que le planning des séquences étant moins prévisible en parole spontanée qu'en parole journalistique, pour les séquences plus longues (au-delà de 3 secondes) il est difficile pour le locuteur d'anticiper la ligne de déclinaison et nous observons un redressement de la ligne de déclinaison sur la 2ème moitié de la séquence. Ces résultats montrent également que la ligne supérieure et la ligne inférieure se comportent différemment, où la ligne inférieure correspondrait plus volontiers à un niveau physiologique, similaire dans les deux styles, alors que la ligne supérieure serait plus dépendante du style (Gendrot et al., 2012).

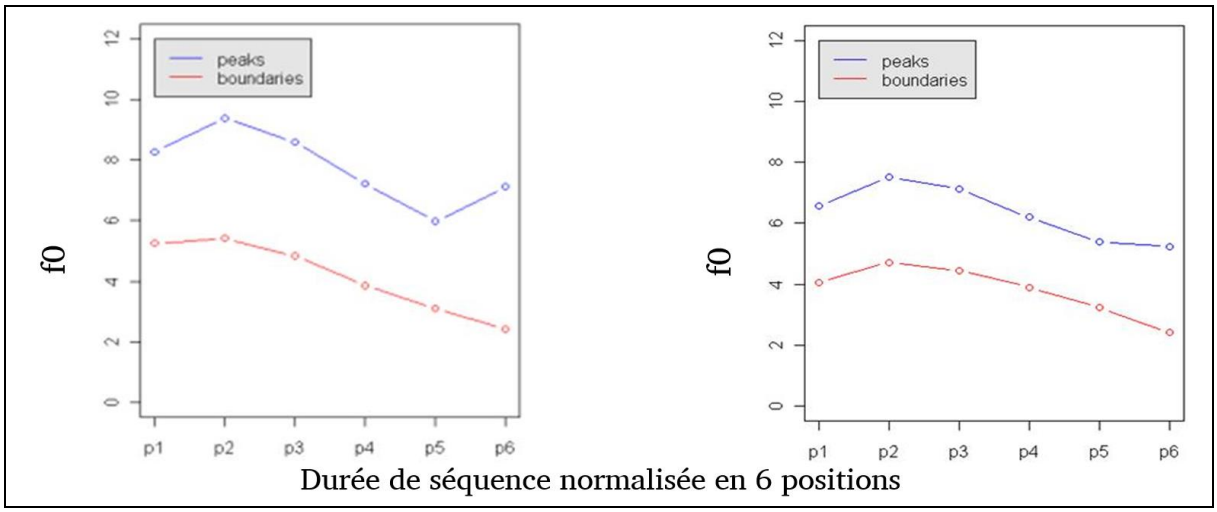


Figure 41 : contour de f0 normalisé en ligne supérieure et ligne de base. Temps normalisé. A gauche parole journalistique et à droite parole spontanée. Séquences de 2 secondes, d'après Gendrot et al. (2012)

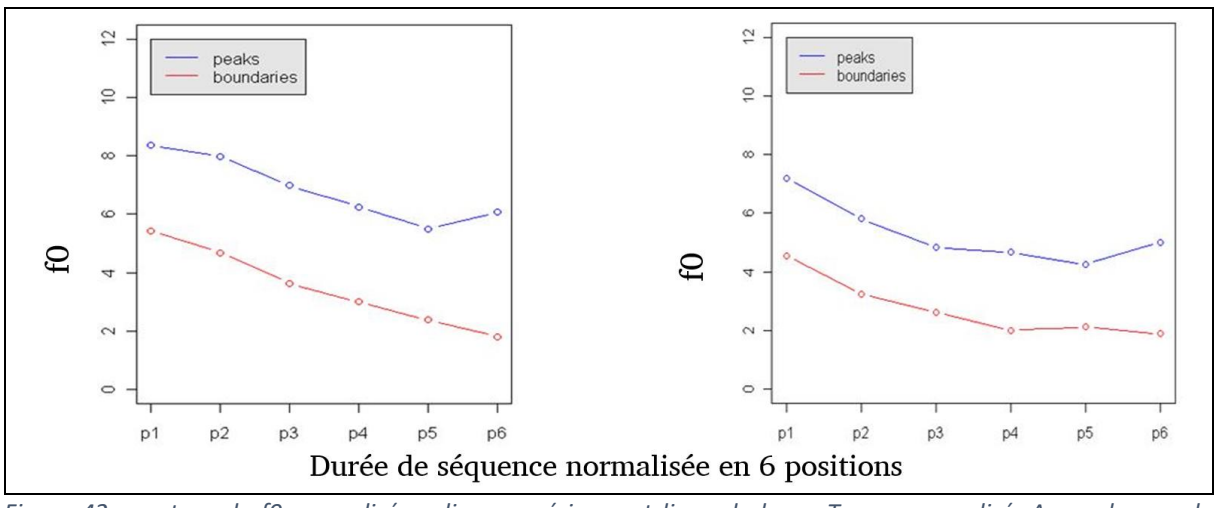


Figure 42 : contour de f0 normalisé en ligne supérieure et ligne de base. Temps normalisé. A gauche parole journalistique et à droite parole spontanée. Séquences de 5 secondes, d'après Gendrot et al. (2012)

4.2.6 Confrontation de la théorie et de la pratique

Dans des études plus centrées sur la prosodie relativement à la structure syntaxique (et informationnelle) des énoncés, en collaboration avec Mathieu Avanzi et Lisa Brunetti, nous avons voulu vérifier si la réalisation acoustique en parole non contrôlée se conformait aux prédictions théoriques et à ce qui avait pu être observé dans la littérature sur des phrases lues. Des corpus de parole spontanée (NCCFr et CFPP) ont été utilisés afin d'identifier au moyen d'un concordancier les différentes occurrences recherchées. Parmi les quelques dizaines d'heures de corpus récoltées, il est fréquent de n'avoir au final qu'une ou deux centaines de séquences à analyser, ce qui peut ensuite être vérifié manuellement.

Dans une première étude (Avanzi et al., 2010), nous avons pu montrer que la réorganisation de l'information par la dislocation, i.e. l'antéposition de groupes syntaxiques dans un énoncé sous la forme « Det + Nom, il est », n'était pas caractérisée par une mise en relief acoustique significativement différente de groupes syntaxique de type « Det + Nom + Verbe », contrairement à ce qui était établi dans un cadre syntaxique théorique et sur de la parole lue.

Dans une seconde étude (Avanzi et al., 2012), trois types de clitiques réintroduits ont été analysés afin de vérifier si leur statut pragmatique pouvait avoir une influence sur leur réalisation. Les clitiques se répartissaient selon les catégories suivantes : (i) topique continu qui montre que le thème de la discussion fait partie du topique en cours, (ii) topique référent qui peut être déduit de la phrase précédente, (iii) topique réintroduit qui a été déjà abordé précédemment mais mentionné à nouveau après un délai. Un découpage plus fin a également été réalisé au sein des deux premiers topiques. Les résultats obtenus nous ont amené à considérer le marquage prosodique des fonctions discursives comme moins systématique que ceux généralement supposés en observant des phrases isolées ou en faisant des analyses de type qualitatif. En effet, seul le topique réintroduit se distingue des deux autres, et aucun des découpages plus fins au sein des deux premiers topiques ne s'est distingué des deux autres. L'utilisation de corpus spontanés et écologiques a ainsi le mérite d'inviter à revisiter certaines théories syntaxiques et pragmatiques à la lumière de ces résultats.

4.2.7 Résumé des travaux sur l'influence de la prosodie

Les travaux présentés dans cette section ont mis en évidence l'importance de la prosodie sur la réalisation acoustique des voyelles dans des corpus de parole non lue. La position dans le mot, le syntagme accentuel et le syntagme intonatif sont trois facteurs de variation récurrents que l'on observe en français, en allemand et en espagnol. La comparaison entre trois langues aux systèmes accentuels différents nous a permis de séparer la structure accentuelle et la structure prosodique, pouvant être mise en avant respectivement soit par des informations spectrales (formants) de façon prépondérante, soit par des paramètres prosodiques (f0 et durée). Dans ce dernier cas, l'hyper-articulation spectrale serait alors une conséquence de l'allongement prosodique. Des méthodes innovantes afin d'analyser automatiquement la structure prosodique dans des grands corpus ont été présentées, incluant un découpage en syntagmes basé sur la transcription orthographique.

Dans la section suivante, je présente des travaux qui soulèvent des questions d'ordre phonologique sur des segments individuels. Ces travaux ont également été novateurs dans le sens

où ils ont proposé des modèles de prédiction avec un nombre de facteurs jamais proposés jusqu'alors, une interaction entre des mesures articulatoires invasives et des mesures acoustiques, et la présentation d'une tendance diachronique globale malgré une variabilité inter-locuteurs forte.

4.3 Trois études segmentales spécifiques : le schwa, le /R/ et la fusion e/ɛ

Dans cette nouvelle section, je présente les résultats d'autres travaux effectués en collaboration. Ces travaux avaient pour but de répondre à des questionnements linguistiques plus précis. Dans le premier cas, la question du schwa est abordée et permet de détailler la part phonétique et la part phonologique de la réduction spectrale que nous avons présentée précédemment. Dans le deuxième cas, le phonème /R/ est analysé quant à son statut phonologique et à sa variation supposée « libre ». Pour finir, une analyse d'un processus en cours en français : la fusion du /e/ et du /ɛ/. Ces exemples nous permettent de comprendre les apports et les limites de la phonologie de corpus.

4.3.1 Le schwa : une élision phonétique ou phonologique

Les processus de réduction que nous avons présentés sont des processus phonétiques dans la mesure où la réduction spectrale ne participe pas d'un processus cognitif qui pourrait être pris en compte dans la représentation sous-jacente du mot. Par exemple, dans le mot « gouvernement », la réduction du /u/ pourrait aboutir à une réalisation de ce mot proche de [gəvɛʁnəmɑ̃]. Cependant, cette réduction n'est pas un phénomène conscient au point que ces deux réalisations fassent référence à deux entités distinctes. De même, si la réduction phonétique est généralement considérée comme étant la non-atteinte des cibles acoustiques et articulatoires, il est fréquent que la variation de prononciation des mots par rapport à leur prononciation canonique aboutisse à une élision de certains phonèmes (Ernestus (2000) pour le néerlandais par exemple). La disparition d'un phonème pourrait alors être considérée comme l'aboutissement d'un phénomène graduel de réduction phonétique. La réduction phonologique vocalique quant à elle réfère à la neutralisation d'un contraste vocalique comme dans l'exemple 'explanation' vs. 'explain' en anglais. Il s'agit d'une substitution catégorielle des sons et non d'une réduction graduelle qui ne dépend pas du débit ou du registre. Il existe un phonème en français qui permet de mettre à l'épreuve cette hypothèse de la réduction phonétique contre la réduction phonologique. En effet, les phénomènes de réduction phonétique coexistent avec un processus d'élision segmentale : l'alternance d'un /ə/ avec 0 dans un certain nombre de mots tels que « fenêtre » ([fənɛtʁ] ou [fnɛtʁ]). Bien que la littérature n'ait pu aboutir à un consensus quant à son statut (voir Côté et Morrison, 2007 pour une revue de cette littérature), il est entendu qu'il ne s'agit pas d'un processus phonétique mais d'une alternance phonologique. La question restait donc ouverte concernant la variabilité acoustique de ces deux variantes. Pourtant il est raisonnable de s'attendre à ce que le schwa soit sujet à de la réduction temporelle et spectrale comme les autres segments, et que cette réduction puisse aboutir à son élision complète. Nous avons effectué une analyse quantifiée sur plus de 3000 occurrences de schwas dans ESTER, qui nous a permis d'établir des critères objectifs entre la présence et l'absence du schwa, et l'étendue de sa variation (Bürki, Fougeron, et al., 2011). En effet, le seuil à partir duquel un phonème serait considéré comme élidé n'est pas aussi évident que l'on pourrait conclure dans un premier temps (e.g., Pitt et al., 2005 et ce point pourra être appuyé par une

expérience de perception visant à demander à des auditeurs naïfs quelle variante ils entendent et les mettre en parallèle des critères acoustiques que nous aurons déterminés.

Nous avons une nouvelle fois utilisé le corpus ESTER et la sélection des occurrences a été effectuée comme suit : plusieurs bases de données lexicales avec leur prononciation ont été concaténées (voir Burki et al., 2008 pour plus de détails). Les informations n'étant pas uniformes d'une base à l'autre, tous les mots pouvant contenir un schwa (optionnel ou obligatoire) ont été retenus pour le filtrage suivant. Les schwas présents dans des mots composés, à des frontières de mots ou de clitiques ou avant un suffixe dérivatif ont été retirés et seuls les mots apparaissant au minimum cinq fois dans le corpus ont été conservés.

Le système d'alignement automatique de l'IRISA a été utilisé comme référence pour décider des mots dont les schwas pouvaient être considérés comme optionnels et donc retenir notre attention. Si le dictionnaire de prononciations autorise ces variantes et que l'alignement propose dans le corpus des occurrences d'un mot avec schwa et des occurrences de ce même mot sans schwa, nous considérons alors que le mot contient un schwa optionnel. Un juge a ensuite écouté toutes les occurrences afin d'éliminer les erreurs du système d'alignement (mauvais items lexicaux), les productions non natives, les séquences bruitées, pour un total de séquences écartées de 8.6 %. Nous avons ainsi obtenu 4294 occurrences correspondant à 191 mots différents produits par 361 locuteurs dont la segmentation a été ensuite corrigée manuellement. Afin de déterminer si le schwa est présent ou non, nous avons cherché visuellement la présence du premier et deuxième formant qui le cas échant permettait de déterminer la présence et ainsi la durée du schwa. Sur cette base, 3098 occurrences de variantes avec schwas ont été sélectionnés pour une analyse acoustique. Une analyse des deux premiers formants a été effectuée selon les mêmes paramètres que pour les sections précédentes.

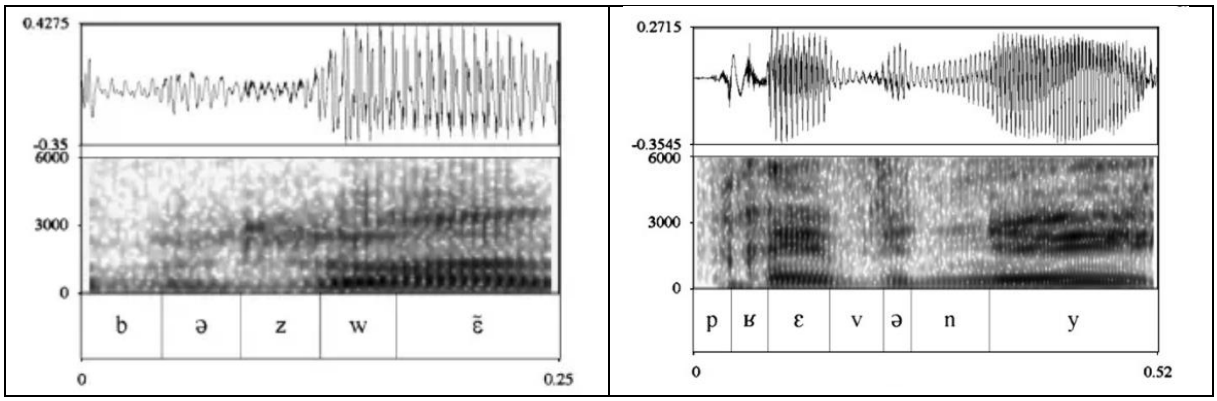


Figure 43 a et b : spectrogrammes et ondes sonores de deux mots contenant un schwa un schwa optionnel, d'après Bürki et al. (2010)

4.3.1.1 Analyse acoustique de la réduction du schwa

Les schwas ont été mesurés en moyenne avec une durée de 51 ms (écart-type : 18 ms, médiane : 50 ms) ce qui confirme la littérature selon laquelle le schwa est une voyelle particulièrement courte comparativement à d'autres, par exemple 67 et 65 ms pour les voyelles /ø/ et /œ/ par exemple (Fougeron et al., 2007). Rappelons également que ces données présentes ont été extraites d'un corpus de parole journalistique pour lequel la réduction phonétique est à priori moindre que dans un corpus de parole spontanée. Comme le montre la Figure 44, la distribution des durées du schwa est proche de la normale, et 14 % des schwas ont une durée inférieure ou égale à 30 ms. Selon Meunier et Espesser (2011), ces voyelles qui peuvent être qualifiées d'extra-courtes représentent 30 % (incluant

d'autres voyelles) dans leur corpus, cette différence pouvant s'expliquer par le style de parole spontané.

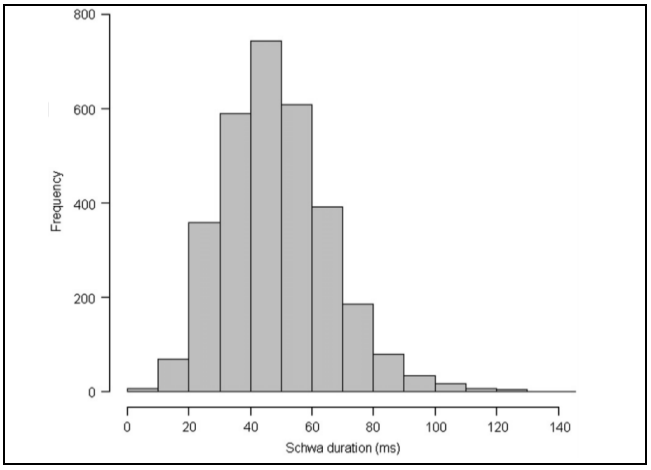


Figure 44 : distribution des durées des 3098 schwas identifiés avec une structure formantique, d'après Bürki et al. (2010)

Afin d'analyser la qualité acoustique du schwa en fonction de sa durée, nous avons appliqué la méthode développée précédemment en regroupant les durées en catégories (ici deux catégories seulement, les courtes inférieures à 50 ms et les longues supérieures ou égales à 50 ms afin d'inclure suffisamment de contextes consonantiques dans chaque catégorie). Les contextes consonantiques ont été regroupés en quatre grandes classes : labial, dental-alvéolaire, palato-vélaire et uvulaire. Deux modèles mixtes linéaires ont ensuite été estimés, le premier avec F1 comme variable dépendante et le deuxième avec F2 comme variable dépendante. Dans ces deux modèles, le locuteur et le mot ont été insérés comme variables aléatoires, la durée du schwa (court vs. long), les contextes précédent et suivant comme prédicteurs. Si les voyelles n'atteignent pas leur cible acoustique quand elles sont plus courtes, elles devraient se diriger vers le locus des consonnes adjacentes. On s'attend donc à ce que les valeurs moyennes observées montent pour les schwas courts dans les contextes où la valeur du locus de la consonne est élevée (pour les consonnes alvéolaires, palato-vélaires, et uvulaires donc les locus sont respectivement à 1800, 3000 et 1400 Hz respectivement). Pour le contexte labial, les valeurs de F2 ne devraient pas être affectées par la consonne labiale à proprement parler puisque la langue n'est pas impliquée, mais plus vraisemblablement par l'autre consonne constituant le contexte, si elle n'est pas labiale elle-même. Dans notre corpus, les schwas étaient dans un contexte asymétrique dans 1292 cas sur 1660, avec un contexte labial d'un côté et un autre contexte de l'autre côté de la voyelle. Dans ces cas, la montée de F2 est due à l'influence de la consonne non labiale (Gendrot et Adda-Decker, 2007).

La valeur de F1 a été mesurée à 409 Hz en moyenne (405 Hz pour les schwas courts et 412 Hz pour les schwas longs), mais la durée n'a pas montré d'effet principal pour la variation de F1 qui reste une voyelle moyenne en termes d'aperture. La valeur moyenne de F2 a été mesurée à 1452 Hz (1484 Hz pour les schwas courts et 1426 Hz pour les schwas longs). Le modèle statistique a révélé trois effets principaux : la durée ($\beta = 37.1$, $t = 4.7$, $p < 0.0001$), le contexte suivant ($F(3,2937)=22.9$, $p=0.0001$) et précédent ($F(3, 2937)=109.0$, $p=0.0001$), sans interaction significative. Comme nous en avions émis l'hypothèse, F2 est plus élevé pour les voyelles plus courtes quel que soit le contexte consonantique, ces changements étant illustrés par la Figure 45.

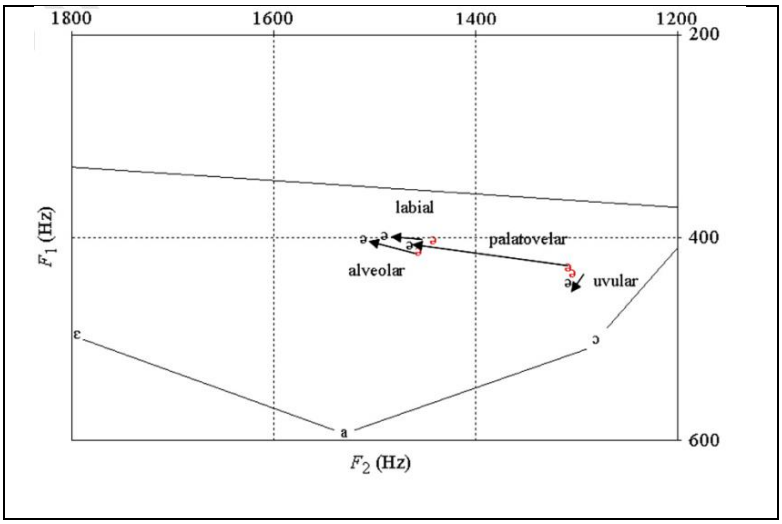


Figure 45 : valeurs de F1 et F2 des schwas en fonction du contexte consonantique suivant (labial, dental-alvéolaire, palato-vélaire, vélaire et uvulaire) et de la durée de la voyelle (court-longue). La flèche va de la catégorie longue vers la courte, d'après Bürki et al. (2010)

4.3.1.2 Perception de l'élision du schwa

Nous avons ensuite , (Bürki et al., 2010) entrepris d'analyser la catégorisation perceptive des variantes avec schwa et des variantes sans schwa. Le premier objectif était de déceler si la présence d'un schwa est aussi catégorielle qu'il est généralement admis. Dans nos corrections manuelles des étiquetages, nous avons pu nous apercevoir qu'il existait des items pour lesquels la catégorisation était ambiguë, et nous avons pu trouver dans la littérature des études qui ont mentionné des problèmes similaires pour l'élision de phonèmes (Pitt et al., 2005 par exemple). Dans un deuxième temps, nous devions examiner les caractéristiques des segments ambigus et tenter de déterminer les variables influençant ce jugement. La durée était un facteur attendu bien sûr, mais le contexte consonantique (lié à la phonotactique, et à l'orthographe) pouvait avoir un impact non négligeable. Pour tester ces deux objectifs, un juge a fait une première écoute des 4294 items. En se basant seulement sur l'audio, il a pu classer les mots en trois catégories : « clairement avec schwa », « clairement sans schwa » ou « ambigu ». A l'issue de cette étape, 67 % des items ont été classés avec schwa, 25 % sans schwa et 8 % (330 occurrences) étaient ambigus. Un sous-ensemble composé de 24 items de chaque catégorie a été sélectionné en uniformisant l'intelligibilité, la qualité de l'enregistrement, la non répétition de mots dans le sous-ensemble et la position du schwa à l'intérieur du mot (15 en position initiale et 9 en position médiane pour chaque catégorie). 22 locuteurs français de l'université de Genève ont participé à ce test, ils se sont vus proposer chacun des 72 items trois fois, chaque liste ayant été randomisée. Pour chaque item, deux transcriptions orthographiques, une avec le « e », et l'autre avec un « ' » apparaissaient sur l'écran d'ordinateur pendant 2 secondes. Puis, 750 ms après la disparition de ces transcriptions, le stimulus a été présenté auditivement aux informateurs qui avaient ensuite à appuyer sur un bouton étiqueté « avec e » ou « sans e ».

La classification des stimuli par le premier juge a été analysée de la façon suivante. Le groupe de schwas ambigus consistait en 240 mots avec un schwa initial et 90 mots avec un schwa médian. Pour 136 de ces mots, aucun intervalle acoustique n'avait été détecté correspondant au schwa en fonction des critères formantiques que nous avions définis. 194 items présentaient un intervalle acoustique d'une durée moyenne de 28 ms (écart-type = 13). Un modèle mixte linéaire généralisé a été utilisé avec le locuteur et le mot comme variables aléatoires et la catégorisation ambiguë vs. non ambiguë comme

réponse. Les résultats ont montré que la présence du schwa dans le mot est jugé comme plus ambiguë si le schwa se trouve dans la première syllabe ($F(1, 4282)=10.4$, $p=0.01$) et si une des consonnes entourant le schwa est une obstruante sourde vs. obstruante sonore ou sonante $F(2, 4282)=13.7$, $p=0.0001$ pour la consonne suivante ; $F(2, 4282)=5.8$, $p=0.01$ pour la consonne précédente). Les contraintes phonotactiques (est-ce que le cluster formé par les consonnes entourant le schwa existe en français ?) n'ont pas eu d'effet significatif.

Concernant la classification par les 22 juges, nous avons obtenu 66 jugements pour chaque item. Afin de déterminer si un item a reçu des jugements unanimes ou hétérogènes (i.e. si la présence du schwa est jugée comme ambiguë), nous avons réalisé un test du Chi² pour chaque item. Nous avons comparé la distribution des réponses des participants à deux distributions théoriques contenant seulement des réponses unanimes (une distribution avec 100 % de réponses « avec schwa » et une distribution avec 100 % de réponses « sans schwa »). Les réponses ont montré que 26 des 72 items n'étaient pas différents d'une distribution théorique de réponses unanimement « avec schwa » et pouvaient ainsi être classées unanimement comme variantes « avec schwa ». 25 items ne différaient pas d'une distribution théorique de réponses uniquement « sans schwa ». Les 21 items restants différaient des distributions théoriques et ont donc été considérées comme ambigus en termes d'absence/présence de schwa. La classification de ces items par les 22 juges ne correspond pas exactement aux catégories définies par le premier juge. Trois items classés comme « clairement sans schwa » par le premier juge ont été classés comme ambigus par les 22 juges, et six items initialement classés comme ambigus ont été jugés comme des variantes clairement « avec schwa » pour trois d'entre eux, et des variantes clairement « sans schwa » pour les trois restantes. Dans l'ensemble les accords entre la classification du premier juge et celle des 22 juges étaient élevés et significatifs (Spearman $rho=0.91$, $S=5584.4$, $p<0.0001$), mais ces résultats confirment que les auditeurs peuvent différer dans leurs jugements. Pour tous les items classés comme ambigus sur la base des jugements des 22 juges, 40 % des participants en moyenne n'ont pas donné trois réponses identiques aux trois répétitions des stimuli.

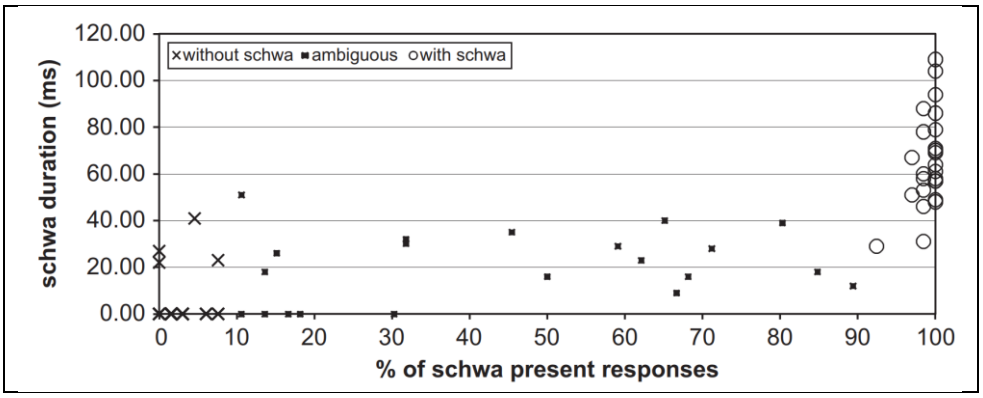


Figure 46 : durées des schwas (en ms) en fonction du pourcentage des réponses « avec schwa ». Les croix, cercles vides et carrés remplis sont utilisés pour représenter les items selon les catégories définies perceptivement comme « avec schwa », « sans schwa » et « ambigus » respectivement, d'après Bürki et al. (2010)

Dans la Figure 46, la distribution des items est affichée en fonction du pourcentage de réponses « avec schwa » et de la durée acoustique du schwa. Trois catégories sont mises en avant basées sur les réponses des 22 juges : « avec schwa » (cercles), « sans schwa » (croix) et « ambigus » (carrés remplis). Sans surprise, une corrélation positive a est observée entre la durée du schwa et le pourcentage de réponses « avec schwa » ($r=0.8$, $p<0.0001$). Les items avec un schwa dont la durée est supérieure à

51 ms appartiennent toujours à la catégorie « avec schwa », et pour la plupart des items de la catégorie « sans schwa », aucun intervalle voisé avec une structure formantique n'avait été observé. Cependant, la présence d'un intervalle voisé avec une structure formantique et la durée de cet intervalle ne suffisent pas à expliquer la catégorisation perceptive. Par exemple, la Figure 43a montre le mot « besoin » avec un schwa de 41 ms mais qui a reçu seulement 4.5 % de réponses « avec schwa ». Tous les items « ambigus » montrent des valeurs sous ce seuil mais nous n'avons pu observer aucune corrélation entre le pourcentage des réponses « avec schwa » et la durée du schwa quand celle-ci est inférieure à 51 ms ($r=0.2$, ns). A titre d'exemple, dans la Figure 43b, le mot « prévenu » contient un intervalle vocalique à structure formantique de 30 ms (inférieure à celui du mot « besoin ») mais pour lequel les réponses « avec schwa » sont de 32 % (très supérieures à celles du mot « besoin »). Comme attendu, la durée a une influence majeure sur les réponses des participants, mais ce n'est pas la seule variable à influencer les réponses. Un autre modèle mixte généralisé a été effectué avec les réponses des participants comme variable dépendante (avec schwa » vs. « sans schwa ») et le locuteur et le mot comme variables aléatoires. Les prédicteurs suivants ont été insérés dans le modèle de manière séquentielle, et conservés lorsque significatifs (à $p<0.05$) : durée du schwa (linéaire et quadratique), nombre de répétitions, débit de parole, fréquence lexicale (issue de la base Lexique 3), sonorité de la consonne suivante, sonorité de la consonne précédente, légalité phonotactique du cluster formé par les consonnes entourant le schwa, durée du mot, position du mot dans le schwa, fréquence estimée de la variante sans schwa (Racine, 2007). Quand deux prédicteurs étaient corrélés, ils ont été orthogonalisés de manière à éviter la colinéarité au sein du modèle. L'orthogonalisation a consisté à modéliser une variable A à partir de la variable B à l'aide d'un modèle linéaire, les résidus de ce modèle ont été utilisés comme prédicteurs plutôt que les valeurs brutes de la variable A. Le modèle final a indiqué que la probabilité d'obtenir les réponses « avec schwa » augmente avec la durée ($\beta=0.17$; $F(1, 4739)=276.2$; $p<0.0001$) et décroit avec le débit de parole (orthogonalisée avec la durée ; $\beta=0.028$; $F(1,4739)=41.8$; $p<0.0001$). La probabilité d'obtenir d'une réponse « avec schwa » est également influencée par le contexte consonantique suivant ($F(2, 4739)=3.6$; $p<0.05$) et précédent ($F(2, 4739)=47.4$; $p<0.0001$). Cette probabilité est aussi plus élevée quand la consonne précédente ou suivante est une sonante plutôt qu'une obstruante sourde. Ce modèle a également révélé des interactions entre le contexte et la durée du schwa : quand la consonne adjacente est une obstruante sourde, l'absence d'un intervalle acoustique attribuable à un schwa ne conduit jamais à une réponse « avec schwa ». Mais dès qu'un intervalle avec du voisement et une structure formantique apparait si petite soit sa durée, la probabilité d'une réponse « avec schwa » monte drastiquement. Pour un contexte constitué d'obstruantes voisées ou de sonantes, une durée plus longue du schwa est nécessaire pour augmenter dans les mêmes proportions la probabilité d'une réponse « avec schwa ». Les auditeurs considèrent probablement plus volontiers les indices de voisement et de structure formantique comme attribuables à la consonne. Nous avons également observé une interaction entre la durée du schwa et la durée du mot qui montre que pour les mots plus courts, un schwa plus long est nécessaire pour obtenir une réponse « avec schwa ». Il est en effet reconnu que les segments sont plus longs à l'intérieur des mots plus courts (Lehiste, 1972) et les auditeurs adaptent très logiquement leur interprétation des indices acoustiques à la durée des segments environnants. L'inclusion d'une variable aléatoire pour les 22 juges et pour les mots contenant le schwa a augmenté significativement la précision du modèle selon des tests du rapport des vraisemblances ce qui confirme que les auditeurs ont des degrés de sensibilité différents, notamment du fait du débit de parole.

Contrairement à ce que la littérature propose pour le schwa comme étant une alternance catégorielle et non ambiguë et s'opposant en cela aux autres phonèmes, nous avons pu montrer que le schwa en français subit des phénomènes de réduction phonétique semblables aux autres phonèmes , (Bürki et al., 2010). Dans notre étude perceptive, nous avons également établi que la détection d'un schwa d'un

point de vue perceptif pouvait être très ambiguë, et que l'identification de ce dernier pouvait reposer sur des facteurs autres que sa seule structure acoustique. L'élision d'un schwa pourrait donc être autant causée par une réduction extrême que par un processus phonologique catégoriel. En se positionnant dans le cadre de la phonologie articulatoire (Browman et Goldstein, 1992), il est possible de considérer des règles phonologiques catégorielles (comme l'assimilation) comme des processus phonétiques graduels. Ainsi Barnes et Kavitskaya (2002) ont suggéré que dans le cas d'une variante sans schwa, le schwa serait occulté par un chevauchement extrême des gestes consonantiques adjacents. Mais l'élision du schwa est également un processus qui se produit dans un débit de parole normal et pour lequel des phénomènes d'assimilation existent entre les consonnes précédant et les consonnes suivant le schwa (par exemple « je sais pas », [ʃsɛpa] voire [ʃʃɛpa] ce qui va à l'encontre d'une élision comme processus graduel ultime d'une réduction, ou alors ce processus aura été intégré cognitivement au préalable. Bien sûr, comme nous le signalerons dans la discussion générale de ce texte, la méthode appliquée ici aura nécessairement eu des implications sur les résultats, les critères retenus pour la segmentation, la présentation des stimuli et de leur contexte au premier juge, puis aux 22 juges, etc. Nous avons donc poursuivi cette étude par une autre analyse sur grands corpus afin de comparer les facteurs impliqués dans les variantes avec schwa et les variantes sans schwa.

4.3.1.3 Analyse multi-facteurs de la réduction

Nous avons examiné l'alternance du schwa en analysant les variables qui conditionnent d'un point de vue strictement acoustique la présence vs. l'absence du schwa, et la durée du schwa lorsque celui-ci est présent (toujours au sein du corpus ESTER). Notre étude (Bürki, Ernestus, et al., 2011) avait deux objectifs : le premier était de déterminer quelles variables influencent l'alternance du schwa et sa durée. Certaines études (Hansen, 1994 ; Malécot, 1976) ont déjà été effectuées sur des corpus de parole enregistrée, mais à l'exception de Racine et Grosjean (2002), les variables considérées quant à la réalisation du schwa étaient prises une par une, alors que Racine et Grosjean ont montré que si plusieurs variables, dont on a montré qu'elles étaient significatives individuellement, sont combinées dans une régression, l'influence d'une ou plusieurs d'entre elles pouvait disparaitre. Ainsi, la connaissance de ces variables pourrait renseigner des théories linguistiques. Par exemple les théories selon lesquelles les réalisations des mots sans schwa résultent d'un processus d'élision phonologique (Dell, 1985) ont besoin d'arguments pour expliquer comment ce processus phonologique peut être sensible à différentes variables.

Le second objectif de ce travail était de vérifier si une analyse de corpus pouvait, au-delà d'une simple description de la distribution des variantes de prononciation, nous renseigner sur la nature catégorielle ou graduelle de la production de variantes, et sur la phase de traitement impliquée. Ces questions sont plus généralement traitées à l'aide d'expériences comportementales (Bagou et al., 2009 ; Bürki et al., 2010) dont l'avantage est de contrôler parfaitement tous les stimuli et les paramètres et les affectant, mais qui induisent invariablement des conditions non naturelles. Dans ce cadre, l'application des résultats à de la parole spontanée restera toujours hypothétique. La combinaison des deux, qui implique de vérifier sur de la parole non contrôlée les résultats d'expérimentations psycholinguistiques est une approche qui se devait d'être approfondie dans le but d'une application plus systématique.

Nous nous sommes concentrés sur la question psycholinguistique de la phase du processus de production impliquée dans l'alternance entre la variante avec schwa et la variante sans schwa. Les modèles de production de la parole distinguent plusieurs phases : le schwa pourrait être présent dans la représentation phonologique du mot et disparaître lors de l'implémentation phonétique par un paramétrage minimal de la durée ne laissant pas de place physique pour la production du schwa (Levelt, 1989). De même, dans le cadre de la phonologie articulatoire (Browman et Goldstein, 1992),

le schwa pourrait être absent du signal à cause du chevauchement des gestes articulatoires des consonnes environnantes. Dans cette considération phonétique, l'absence du schwa est le résultat naturel du raccourcissement du schwa, et dans cette optique l'alternance du schwa et sa durée sont susceptibles d'être influencés par les mêmes variables (Bürki et al., 2011).

Dans un autre cadre théorique, l'alternance du schwa pourrait être la conséquence de règles phonologiques qui suppriment le schwa d'une unique représentation phonologique stockée dans le lexique mental (Dell, 1985), ou qui ajoutent le schwa (Côté et Morrison, 2007) pendant l'encodage phonologique. Si l'alternance du schwa est phonologique par nature, elle n'est pas l'aboutissement naturel d'un processus phonétique de réduction et devrait donc être affectée par des variables différentes de celles qui affectent la durée de la voyelle. Malheureusement les études aboutissant à ces conclusions sont basées sur de l'introspection ou des données contrôlées et créées ad-hoc. Pour finir, l'alternance du schwa pourrait se produire parce que le lexique mental contient des représentations phonologiques des deux variantes (avec et sans schwa), ce qui correspondrait aux théories basées sur les exemplaires (Bybee, 2001 ; Pierrehumbert, 2001). Bürki et ses collègues (2010) ont montré que les locuteurs produisent les variétés avec et sans schwa plus rapidement ou plus lentement en fonction de leur fréquence lexicale, ce qui suggère que les deux variantes sont stockées dans le lexique mental avec des fréquences lexicales différentes. Ce résultat va dans le sens d'une hypothèse phonologique et on s'attendrait que la durée du schwa et son alternance soient gouvernés par des variables différentes.

Pour ce faire, nous avons réutilisé les 4294 occurrences produites par 361 locuteurs au sein du corpus ESTER et présentées dans la section précédente. Pour rappel, sur ces 4294 items, 1198 ont été réalisés sans schwa, et 3096 avec un schwa.

Nous présentons ensuite les 17 variables testées pour l'alternance du schwa, ainsi que sa durée quand celui-ci est présent et qui sont résumées dans le Tableau 22. Ces variables correspondent aux variables testées dans la littérature et dont l'effet significatif a été établi au moins deux fois.

variable	Description
Variables segmentales	- voisement de la consonne précédente et suivante - mode d'articulation de la consonne précédente et suivante - lieu d'articulation de la consonne précédente et suivante - le mot commence par « re »
Variables phonotactiques	- nombre de consonnes dans la séquence autour du schwa - respect du principe de sonorité - existence d'une séquence consonantique - fréquence de la séquence consonantique
Variables morphologiques / grammaticales	- composition morphologique - classe grammaticale
Variables prosodiques	- position du mot dans la phrase - position du schwa dans le mot - longueur du mot (en phonèmes et en syllabes)
Variable débit de parole	- (nombre de syllabes dans les deux mots précédents + mot-cible + deux mots suivants) / durée de cette séquence
Variable locuteur	- variable aléatoire
Variable lexicale	- fréquence de la variante phonologique (avec ou sans schwa)
Variable prédictibilité du mot	- fréquence lexicale (Lexique et ESTER) - information mutuelle précédente et suivante

Tableau 22 : récapitulatif des variables utilisées dans cette étude, d'après Bürki et al. (2011)

Nous expliquons ci-dessous les variables dont la description dans la colonne droite du Tableau 22 qui ne se suffirait pas à elles-mêmes, en commençant par les variables segmentales (Bürki et al., 2011).

Suivant la littérature, le voisement, le lieu et le mode d'articulation de la consonne précédente et de la consonne suivante peuvent avoir une influence sur l'élision du schwa, soit au total six variables catégorielles dont la distribution est précisée dans le Tableau 23. Nous avons également inclus une variable catégorielle indiquant si le mot commence par « re ».

	Voicing		Manner of articulation				Place of articulation		
	Voiced	Voiceless	Fricative	Plosive	Liquid	Nasal	Front	Mid	Posterior
Following consonant	2905	1389	705	1323	1031	1235	1527	1975	792
Previous consonant	2523	1771	1513	1479	1117	185	822	2447	1025

Tableau 23 : distribution des occurrences en fonction des propriétés de la consonne précédant ou suivant le schwa, d'après Bürki et al. (2011)

Pour ce qui concerne les variables phonotactiques, puisque l'élision du schwa aboutit à une séquence consonantique, les propriétés segmentales et phonotactiques de cette séquence ont souvent été indiquées comme ayant une influence sur l'alternance du schwa. Citons notamment la loi des trois consonnes de Grammont (1914) qui stipule que le schwa sera moins élidé si la séquence consonantique résultante contient plus de deux consonnes. Malécot (1976) a montré que l'élision du schwa sera plus rare si la séquence de deux consonnes ainsi obtenue diffère en termes d'aperture, et plus particulièrement si la deuxième consonne est moins ouverte que la première.

Plusieurs variables ont été choisies pour décrire la composition de la séquence consonantique quand le schwa est élidé. Nous avons codé : (a) le nombre de consonnes (les semi-voyelles ont été comptées comme des consonnes) en utilisant deux mesures différentes. Pour la première, le nombre de consonnes autour du schwa à l'intérieur du mot. Pour la deuxième, on a ajouté au nombre de consonnes autour du schwa, la présence éventuelle de la consonne finale du mot précédent (par exemple « stm » pour la séquence « cette semaine »). (b) Nous avons également codé si la séquence consonantique résultant de l'élision respectait le principe de sonorité (Clements, 1990). (c) Nous avons évalué par une variable catégorielle à trois niveaux si la séquence résultante existe en français (n'existe pas vs. existe et forme une syllabe vs. existe mais hétérosyllabique). (d) Pour finir, nous avons inclus la fréquence lexicale cumulée des mots contenant la séquence consonantique résultante quelle que soit sa position dans le mot.

Nous avons évalué la complexité morphologique sur une échelle de 1 à 5 en indiquant si le schwa est clairement à l'intérieur du morphème (valeur = 1) ou en frontière du morphème (valeur = 5) (pour plus de détails voir Bürki et al., 2011). Pour finir, la classe grammaticale du mot (mot grammatical vs. mot lexical) a été prise en compte puisque ce facteur joue un rôle dans la durée des phonèmes en général.

Plusieurs variables prosodiques ont été présentées comme ayant un effet à la fois sur l'alternance du schwa et sur la durée de la voyelle. Les schwas sont plus fréquents en syllabe initiale, et les syllabes initiales sont plus longues toutes choses égales par ailleurs que les syllabes internes de mot. Puisque les schwas en syllabe finale n'ont pas été pris en compte, nous avons créé une variable catégorielle à deux niveaux séparant les schwas en syllabe initiale des schwas en syllabe non initiale. Deuxièmement, la position du mot dans la phrase/séquence semble être un paramètre pertinent. Selon Malécot (1976), les schwas sont plus souvent réalisés en position initiale et les voyelles tendent à être plus longues vers la fin. La position dans la séquence a ainsi été codée selon quatre degrés : initiale, finale, interne, ou isolée. Les limites gauche et droite de chaque séquence ont été définies par les

transcripteurs du corpus (voir détail sur http://trans.sourceforge.net/en/transguid.php). Troisièmement, la longueur du mot est également un facteur considéré comme significatif dans la littérature : plus le mot est long, moins la présence du schwa est attendue. Nous avons considéré ici deux mesures de la longueur du mot, le nombre de phonèmes et le nombre de syllabes.

Le débit a été ajouté, et calculé comme le nombre de syllabes dans une séquence commençant deux mots avant et terminant deux mots après le mot clé, et divisé par la durée de la séquence (les pauses ont été exclues).

La fréquence lexicale des variantes (avec et sans schwa) estimée par Racine (2007) a été également insérée comme variable. La fréquence lexicale a été insérée en tant que logarithme de la fréquence naturelle obtenue dans la base de données Lexique, et celle observée dans ESTER a également été retenue (les deux étant significativement corrélées à r=0.7). Une mesure de probabilité du mot en fonction du contexte précédent ainsi qu'une mesure de probabilité du mot en fonction du contexte suivant ont été proposées pour finir.

4.3.1.3.1 Modélisation de l'alternance du schwa

Des modèles mixtes linéaires généralisés ont été estimés, en insérant le locuteur et le mot comme termes aléatoires. Pour chaque modèle, nous avons eu recours à une procédure par étapes en incluant toutes les variables de manière séquentielle et en ne retenant que les variables significatives à $p<0.01$. Comme précédemment, les variables montrant de la colinéarité ont été orthogonalisées. Dans toutes les analyses, la variable dépendante indiquait la présence vs. l'absence du schwa. Le modèle final a permis de prédire 90 % des observations et le Tableau 24 fournit les valeurs statistiques associées à chaque prédicteur significatif ainsi que ses interactions significatives. Quatre effets principaux ont été observés : la position dans le mot, la position dans la séquence, le débit de parole et le respect du principe du sonorité.

Predictor	Df	F	p
Schwa position in word	1, 4137	119.25	<0.0001
Word position in utterance	3, 4137	23.41	<0.0001
Speech rate	1, 4137	231.80	<0.0001
Respect of sonority principle	1, 4137	8.23	<0.01
Number of consonants in sequence	1, 4137	4.95	<0.05
Number of consonants in sequence by Respect of sonority principle	1, 4137	9.60	<0.01

Tableau 24 : résumé des variables prédisant l'alternance du schwa, d'après Bürki et al. (2011)

Pour résumer ces résultats, le schwa est plus souvent présent en position initiale de mot, en initiale de séquence, pour un débit plus lent, et si le respect du principe de sonorité n'est pas assuré en cas d'élision. Tous ces résultats correspondent aux hypothèses que nous avions avancées (Bürki et al., 2011).

4.3.1.3.2 Modélisation de la durée du schwa

La durée du schwa a été modélisée dans quatre analyses différentes utilisant une nouvelle fois des modèles à effets mixtes. La première analyse a prédit la durée du schwa pour l'intégralité des 3096 mots réalisés avec un schwa, et nous avons pris en compte les variables disponibles pour la plupart des mots.

Dans cette analyse (Bürki et al., 2011), la durée du schwa est en moyenne de 51 ms (écart-type : 18) avec 8 ms pour minimum et 150 ms pour maximum, et une médiane de 50 ms. Nous avons appliqué à la durée un logarithme (naturel) dans le but de réduire l'asymétrie de la distribution. Comme précédemment, les locuteurs et les mots ont été ajoutés comme termes aléatoires. Les résultats montrent que le schwa est plus court : (1) quand le débit est plus rapide, particulièrement s'il est suivi d'une consonne voisée, (2) si le schwa n'est pas dans la syllabe initiale de mot, particulièrement s'il est précédé d'une obstruante voisée, (3) si le schwa est entouré d'obstruantes voisées. Le modèle a également montré que les locuteurs sont sensibles à la position du schwa dans le mot. Les effets de l'entourage lexical ont été analysés et ont montré que plus la prédictibilité du mot en fonction du mot suivant est élevée, et plus le schwa est court.

L'effet de la fréquence de la variante sans schwa a également été observé (seulement pour les items pour lesquels nous avions trouvé une fréquence dans les travaux de Racine), et ce nouveau prédicteur a été ajouté au modèle précédemment obtenu, nous avons ainsi obtenu que les mots dont la variante avec schwa était plus fréquente (que dans leur variante sans schwa) avaient des schwas plus longs.

Predictor	Df	F	P
Schwa position in word	1,2909	21.60	<0.0001
Speech rate	1,2909	359.45	<0.0001
Following consonant voicing	1,2909	7.82	<0.01 (N.S)[a]
Preceding consonant voicing	1,2909	4.72	>0.01
Following consonant voicing by preceding consonant voicing	1,2909	16.48	<0.0001
Speech rate by following consonant voicing	1,2909	13.15	<0.001
Position by preceding consonant voicing	1,2909	7.43	0.01

Tableau 25 : résumé des effets du modèle mixte utilisé pour modéliser la durée du schwa, d'après Bürki et al. (2011)

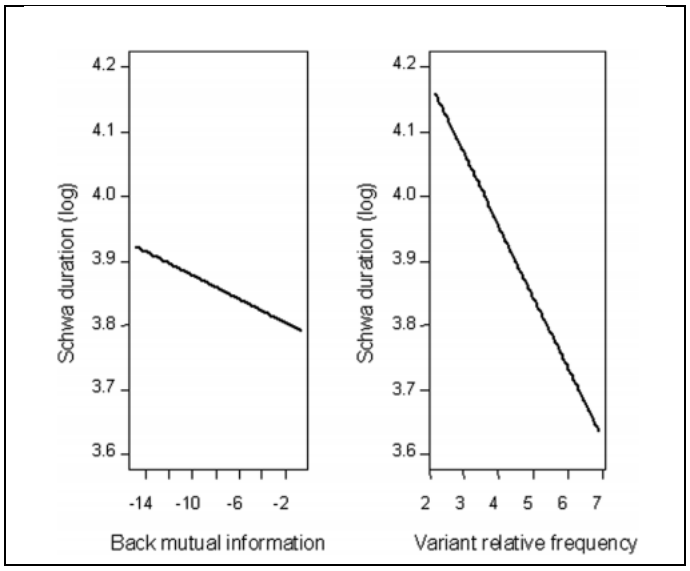


Figure 47 : durée du schwa en fonction du contexte lexical suivant (à gauche) et de la fréquence de la variante sans schwa (à droite) prédite par le modèle statistique, d'après Bürki et al. (2011)

Le premier but de ce travail était d'examiner parmi les nombreuses variables mentionnées dans la littérature celles qui influencent la présence ou l'absence des schwas internes de mots dans un corpus de parole continue. Le petit nombre de variables significatives peut être expliqué par l'intégration de toutes les variables simultanément tout en s'étant débarrassé de la colinéarité, et par l'utilisation de statistiques conservatrices (i.e. le seuil de significativité fixé à 0.01 et l'utilisation de termes aléatoires pour prendre la variabilité en considération). Ces résultats permettent d'affiner les modèles théoriques psycholinguistiques puisque nous avons pu montrer que la mesure du principe de sonorité qui distingue les fricatives des occlusives modélise mieux l'élision du schwa. De plus, les théories linguistiques prédisent en général que l'absence du schwa peut être prédite par la position du schwa dans le mot et par les consonnes entourant le schwa, mais nous avons pu montrer que le débit de parole ainsi que la position du mot dans la séquence doivent être pris en compte également.

Le deuxième objectif de cette recherche (Bürki et al., 2011) était de déterminer si des analyses d'un grand corpus pouvaient fournir des informations sur la nature (graduelle vs. catégorielle) et la phase du processus d'alternance. Le Tableau 26 résume les variables impliquées significativement, et les variables communes aux deux analyses sont indiquées en italique. Ce sont des variables qui peuvent intervenir à la fois au niveau phonétique et phonologique.

Predictors of Schwa Alternation	Predictors of Schwa Duration
Speech rate	*Speech rate*
Schwa position in word	*Schwa position in word*
Word position in utterance	Following consonant voicing
Number of consonants in sequence	Preceding consonant voicing
Respect of the sonority principle	Back mutual information

Tableau 26 : prédicteurs significatifs de l'alternance du schwa et de sa durée, d'après Bürki et al. (2011)

A l'inverse, les trois variables qui conditionnent seulement la durée du schwa sont connues pour être actives dans des processus phonétiques de réduction temporelle et spectrale. L'alternance du schwa par contre est conditionnée par trois paramètres prosodiques par nature : la position dans la phrase, le nombre de consonnes environnantes et le respect du principe de sonorité. Ces trois paramètres sont dépendants de la langue (par rapport au berbère par exemple) et du locuteur (certains locuteurs ne sont pas sensibles au principe de sonorité), ce qui exclut le fondement phonétique de ces variables. Puisque la durée n'est pas affectée par ces trois variables, nous pouvons défendre l'hypothèse que la sélection de la forme avec ou sans schwa se fait avant l'implémentation phonétique, que ce soit au niveau lexical (sélection de la forme) ou au niveau de l'encodage phonologique du mot. Nous défendons également l'idée que ces deux variantes sont stockées dans le lexique mental et que la réalisation acoustique d'une variante de prononciation est affectée par l'autre variante par analogie. Plus la variante sans schwa est fréquente et plus haut est son niveau d'activation, ce qui par conséquent affecte d'autant plus la réalisation acoustique de la variante avec schwa en induisant une variante plus courte (Bürki et al., 2011). Bien sûr, il serait utile de pouvoir confirmer ces résultats sur d'autres données, à la fois acoustiques et comportementales, mais le regard nouveau fourni par l'analyse automatique de grands corpus ne peut pas être négligé. Dans la section suivante, je présente un second cas d'étude impliquant à la fois phonétique et phonologie : l'analyse du /ʀ/ en français.

4.3.2 Le /ʀ/ : variation phonétique et statut phonologique en français

J'ai obtenu en 2013 une ANR JCJC (jeunes chercheurs) sur l'analyse de la variabilité du /ʀ/[3] uvulaire (i.e. la variété standard) en français. Un des objectifs était d'interpréter la variabilité de ce phonème à la lumière des analyses d'hypo- et d'hyper-articulation observées sur les voyelles. Le /ʀ/ est un phonème identifié comme très variable, plus variable que les autres phonèmes (Chafcouloff, 1980 ; Meunier, 1994), et dont l'interprétation a souvent été mise de côté dans la littérature en la considérant comme sujette à variation libre, mêlant à la fois les réalisations fricatives et approximantes, et un voisement sourd ou sonore. Mon hypothèse était que la réalisation non voisée est essentiellement fricative et correspond à une variante hyper-articulée, alors que la réalisation voisée est une approximante dans sa production hypo-articulée. La source de variabilité plus importante serait liée au fait que l'articulation étant très postérieure, avec un espace naturellement réduit à la courbure du conduit vocal, le voisement serait donc « éteint » rapidement par une équivalence entre la pression intra-orale et la pression sous-glottique due à la constriction linguale.

Le deuxième objectif était de s'intéresser de plus près au statut phonologique du /ʀ/ uvulaire standard en français. En effet, le /ʀ/ a des propriétés acoustiques qui le rapprochent d'une fricative, mais des propriétés phonotactiques similaires à celles des liquides. Une analyse acoustique permet-elle de reconsidérer le statut phonologique de cette consonne ? Nous avons également ajouté à l'analyse acoustique une analyse physiologique (articulatoire et aérodynamique) dont le but était de comprendre quels mécanismes sous-jacents pouvaient expliquer la variabilité observée pour le /ʀ/.

[3] J'ai préféré l'utilisation du symbole /ʀ/ à celle de /ʁ/ : le /ʀ/ est ici considéré comme l'archiphonème dont on étudie la variation en contexte.

Une première expérience (Gendrot, 2017 ; Gendrot et al., à paraitre) avec des données obtenues à partir d'un articulographe électro-magnétique (EMA) nous a permis de comprendre si les différentes réalisations de /ʀ/ variaient en termes de constriction entre la langue et le palais. Dans un deuxième temps, une étude aérodynamique nous a permis d'analyser les mêmes productions en termes de flux d'air, avec des données de pression sous-glottique en variable de contrôle. Dans ces deux expériences, de par le petit nombre de données, et considérant que celles-ci ont été correctement contrôlées, des ANOVAs ont été effectuées pour les tests statistiques. Le but de ces deux études était également de proposer des mesures acoustiques compatibles avec les réalisations articulatoires et aérodynamiques observées, et pertinentes pour capter la variation du /ʀ/. Ces mesures acoustiques seront ensuite utilisées dans plusieurs études sur des grands corpus de parole non préparée.

4.3.2.1 Acoustique et articulatoire

Les données EMA ont été collectées à l'Institut de Phonétique de Munich sur cinq locuteurs natifs du français avec l'EMA 3D AG500 de Carstens (Figure 48 gauche). Les locuteurs étaient des locuteurs natifs du français âgés de 29 à 50 ans. Il a été observé dans la littérature que la luette peut entrer en contact avec l'arrière de la langue afin de favoriser la friction. Dans d'autres données auxquelles j'ai pu avoir accès, il est possible d'observer un abaissement et une antériorisation de la luette, mais dans une moindre mesure que pour les nasales. Des données cinéradiographiques ont également montré que la luette pouvait s'abaisser pour entrer en contact avec la langue tout en maintenant la cavité nasale fermée. Les données aérodynamiques visaient à vérifier si un flux d'air nasal est détecté pendant la production du /ʀ/.

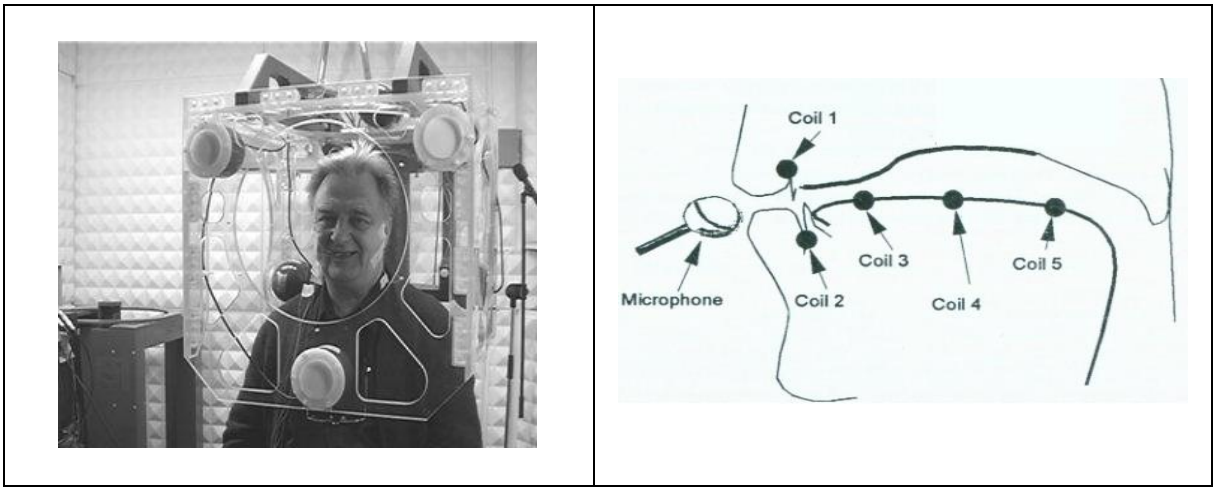


Figure 48 : illustration de l'EMA 3D AG500 de Carstens (à gauche) et de la position des capteurs EMA (à droite)

Nous avons analysé la position des capteurs 3, 4 et 5 (voir Figure 48 droite) dans les contextes suivants :

- /ʀ/ précédé d'une obstruante sourde
- /ʀ/ précédé d'une obstruante sonore
- /ʀ/ en initiale de mot précédé d'une pause

Dix répétitions ont été effectuées, avec les voyelles /a/ et /i/. La Figure 49 montre que la position du capteur 5 est significativement plus élevée ($p<0.01$) en contexte voisé par rapport au contexte non voisé, et ce pour le contexte labial, dental et vélaire, ce qui confirme l'hypothèse que la constriction

est plus importante en contexte non voisé. La Figure 50 montre également la trajectoire du capteur pour l'ensemble de la séquence CRV en fonction du voisement de la consonne initiale, et nous pouvons observer que la trajectoire est différente dès le départ et que la constriction moindre est anticipée dès la consonne initiale.

Quant à la position horizontale du capteur 5, elle est également significativement plus antérieure ($p<0.01$) dans les contextes labial et dental non voisés. La position des capteurs 3 et 4 suit les mêmes tendances que celle du capteur 5. Le /R/ initial montre quant à lui des valeurs intermédiaires entre le contexte labial voisé et non voisé. Quant à la différence entre /a/ et /i/, la réalisation du /R/ n'est pas antériorisée mais la langue est significativement plus élevée ($p<0.0001$) quand celui-ci précède un /i/, plus particulièrement pour les capteurs 4 et 5 (Gendrot, 2017).

Une analyse acoustique de ces données montre que le /R/ en contexte non voisé est significativement plus long ($p<0$ .0001), avec une intensité plus faible ($p<0.0001$) et un HNR plus bas ($p<0.0001$).

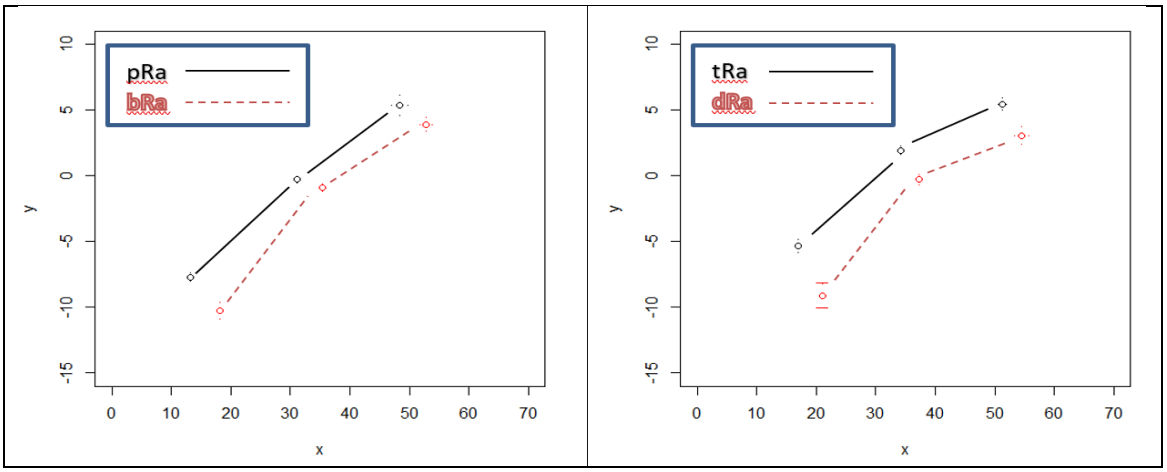


Figure 49 : positions verticales et horizontales des capteurs 3, 4 et 5 en fonction du contexte segmental, d'après Gendrot (2017)

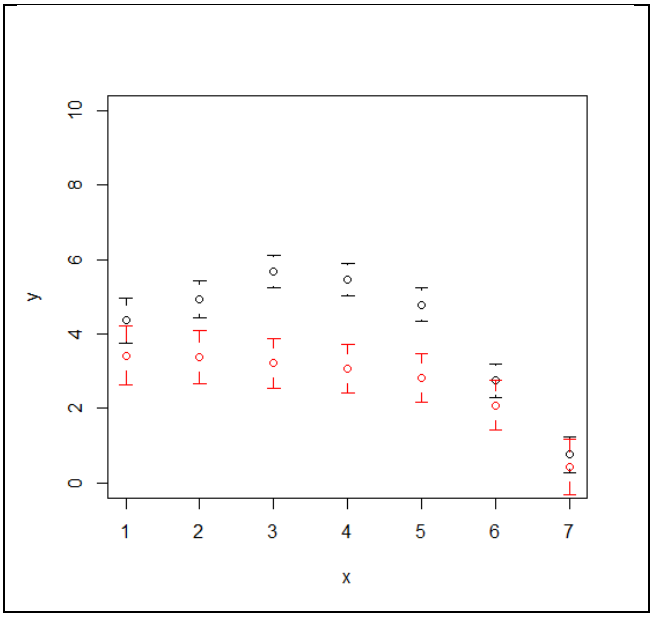


Figure 50 : mouvements du capteur 5 en fonction du temps pendant la réalisation de la séquence CʀV (en noir [tRa] ; en rouge [dRa], d'après Gendrot (2017)

Les données aérodynamiques ont été collectées par Didier Demolin (Professeur, Université Sorbonne Nouvelle) à l'hôpital de Bruxelles à l'aide de la station d'acquisition Physiologia. Des mesures de débit d'air oral ont été obtenues par l'intermédiaire du masque de Rothenberg, le débit d'air nasal au moyen de deux tubes fixés aux narines avec une olive en silicone. Un tube flexible en plastique inséré dans l'oropharynx via la cavité nasale a permis de recueillir la pression intra-orale et la pression sous-glottique a été mesurée directement via une aiguille insérée dans la trachée sous le cartilage cricoïde. Seuls deux locuteurs ont été analysés pour cette expérience. Puisqu'un masque de Rothenberg recouvre le visage du locuteur, les mesures acoustiques sont rendues difficilement fiables et seule la durée a été mesurée.

Une mesure de flux de résistance à l'air (mesuré comme le ratio entre la pression intra-orale et le flux d'air), qui est un paramètre relié au degré de constriction a également été calculé. Si le flux est élevé, cela indique une constriction importante dans la cavité buccale, et vice-versa. Dans nos analyses, le point d'inflexion maximal pendant la réalisation du /ʀ/ a été pris en compte. Le corpus consistait en une série de phrases répétées trois fois avec /ʀ/ dans des contextes C/ʀ/V, V/ʀ/C, V/ʀ/V et V/ʀ/# (en finale de phrase). Les voyelles étaient /a/, /i/ et /e/ et les consonnes /p,b,t,d,k,g,f,v,s,z/. Bien que le corpus ne soit pas identique à celui de l'expérience EMA, les contextes phonétiques semblables ont été analysés.

La pression intra-orale a été mesurée comme significativement plus élevée ($p<0.0001$) pour le /ʀ/ en contexte non voisé par rapport au contexte voisé (voir une illustration en Figure 51), le /ʀ/ initial de mot isolé étant systématiquement voisé. Le flux d'air oral n'est pas significativement différent entre les contextes sourd et sonore, et le flux de résistance à l'air est donc logiquement significativement plus élevé en contexte non voisé (Gendrot et al., 2017). Comme pour l'expérience EMA, le /ʀ/ en contexte non voisé est caractérisé par une durée plus longue ($p<0.001$). Un des locuteurs a volontairement réalisé des trilles uvulaires voisées dans quelques enregistrements, et celles-ci sont caractérisées par une pression sous-glottique plus élevée (avec des stries) et un flux de résistance à l'air intermédiaire amenant à une fricative voisée. A la fin des phrases lues, les deux locuteurs ont réalisé leur /ʀ/ avec une pression sous-glottique descendante, mais un flux de résistance à l'air bas, qui a généré des réalisations approximantes non voisées. Concernant la pression sous-glottique pendant la production du /ʀ/, elle ne varie pas à l'exception de productions emphatiques (où elle augmente) ou de fins de phrases (où elle baisse). Ces variations permettent d'expliquer les deux autres variations du /ʀ/ mentionnées dans la littérature, quoique plus rares : la fricative voisée et l'approximante sourde (voir Figure 52).

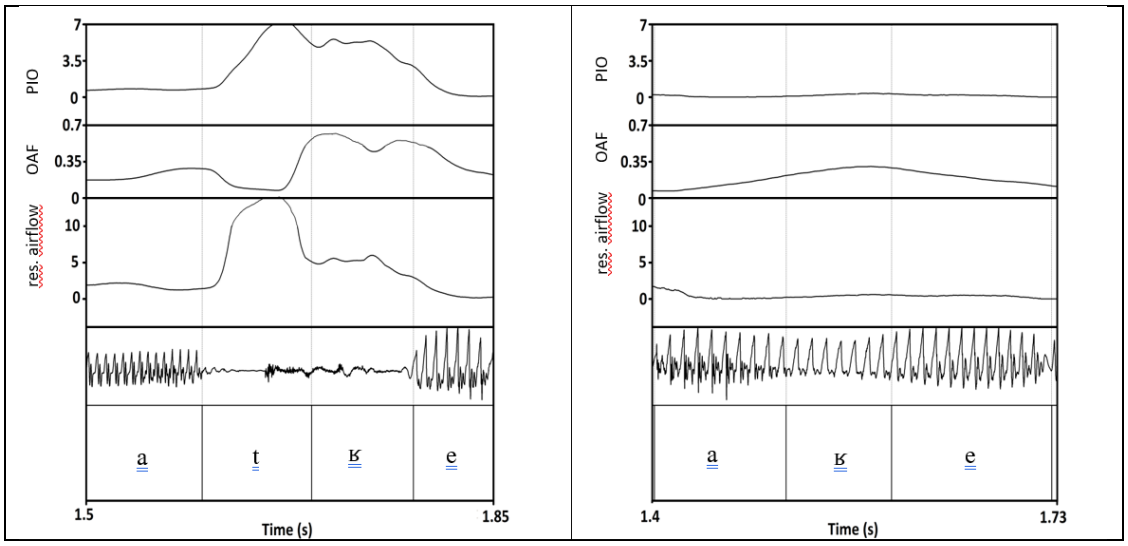


Figure 51 : résultats aérodynamiques pour le locuteur 1. Pression intra-orale, flux d'air oral et flux de résistance à l'air pour deux variantes du /ʁ/ : en contexte non voisé (gauche), et en contexte vocalique (droite), d'après Gendrot et al. (2017)

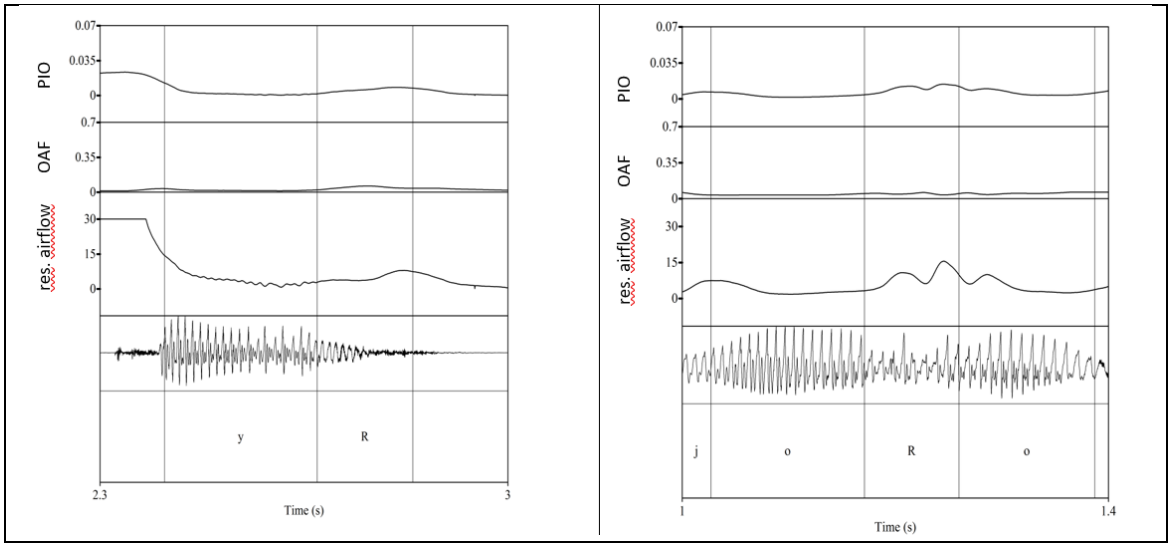


Figure 52 : résultats aérodynamiques pour le locuteur 1. Pression intra-orale, flux d'air oral et flux de résistance à l'air pour deux variantes du /ʁ/ : en position finale de phrase (gauche), et réalisés sous la forme d'une trille voisée (droite), d'après Gendrot et al. (2017)

La Figure 53 récapitule les différentes réalisations du /ʁ/ présentées ci-dessus en fonction des mesures aérodynamiques de pression sous-glottique et de résistance à l'air.

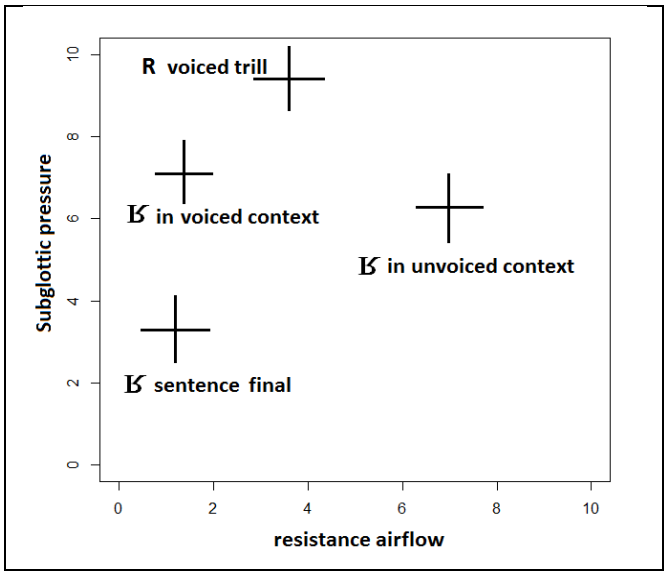


Figure 53 : résumé des résultats aérodynamiques (pression sous-glottique vs. résistance à l'air) pour le locuteur 1 en fonction des différentes réalisations du /ʀ/ observées dans le corpus, d'après Gendrot et al. (2017)

Les résultats présentés pour ces deux expériences articulatoires montrent que les réalisations peuvent se résumer de façon simplifiée selon un continuum allant de l'approximante voisée à la fricative sourde, et on ne peut observer, plus rarement, des approximantes sourdes et des fricatives voisées qu'en cas de mouvement important de la pression sous-glottique. Nous avons ensuite tenté d'appliquer ces connaissances sur des corpus acoustiques de parole non contrôlée, i.e. ceux que nous avons utilisés jusqu'à présent.

4.3.2.2 Analyses sur grands corpus

Comme observé dans la littérature (Meunier, 1994 ; Fougeron, 2007) et dans nos données, la variabilité de la réalisation du /ʀ/ est particulièrement avérée en position finale de mot, et ce notamment avant une pause. C'est cette position que nous avons choisie de présenter ici, filtrée selon les procédés précédemment évoqués (voir section 4.2.2). Une analyse du /ʀ/ post-vocalique sur le corpus ESTER nous a permis de montrer (voir Figure 54, et une illustration dans la Figure 55) - en dehors des observations que nous avions déjà faites sur les corpus articulatoires - qu'en position finale de mot (quand /ʀ/ est précédé d'une voyelle et suivi d'une pause), celui-ci est caractérisé par une valeur de HNR significativement plus basse lorsque le contour de f0 est montant. Ces résultats montrent qu'en position finale pour un contour montant (approchant les conditions d'une fin de groupe intonatif), la réalisation du /ʀ/ est plus proche d'une approximante, alors qu'en position finale avec un contour descendant (approchant les conditions d'une fin de phrase), la réalisation du /ʀ/ est plus proche d'une fricative. Le débit de parole ajoute encore à ces différences, avec un /ʀ/ de plus en plus approximant à mesure que le débit de parole augmente.

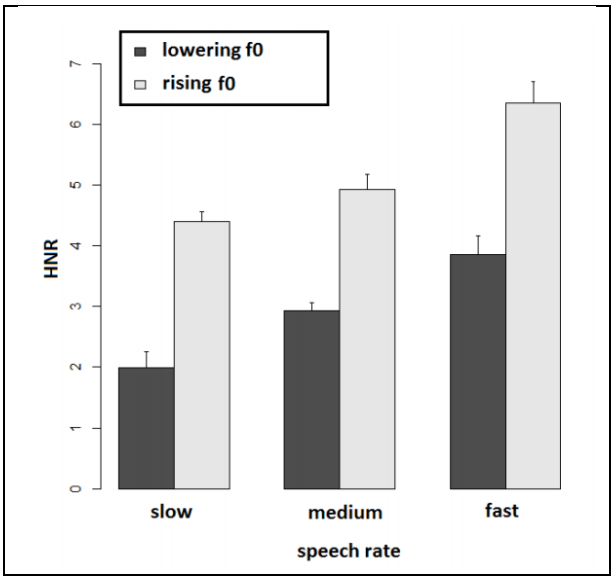


Figure 54 : HNR en fonction du début de parole et du contour de f0 précédant le /ʀ/ final de mot (précédé d'une voyelle et suivi d'une pause) , d'après Gendrot (2017)

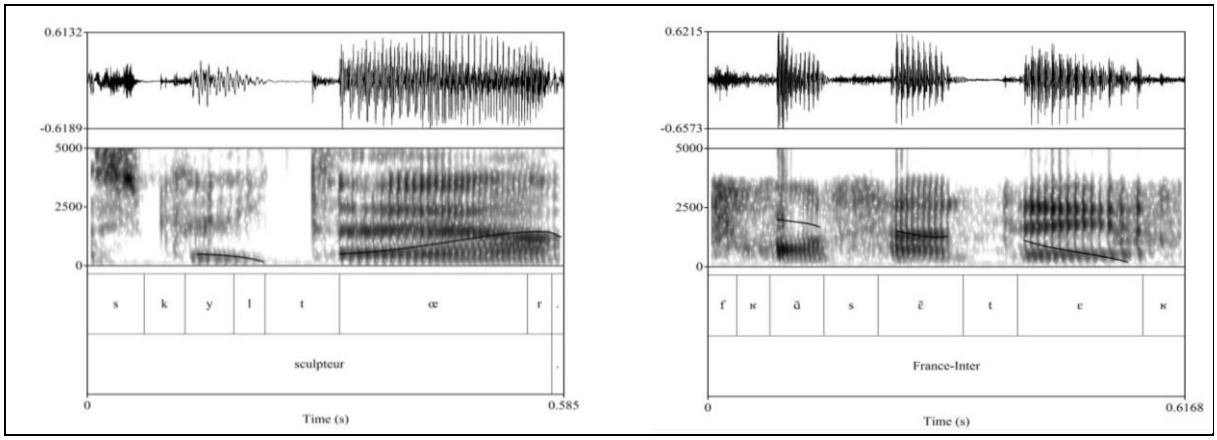


Figure 55 : deux exemples de variation du /R/, voisé à gauche et dévoisé à droite, en fonction du contour de f0 respectivement montant et descendant, d'après Gendrot (2017)

Suite à ces études montrant la variation du /ʀ/ en fonction de la position prosodique et du débit, nous avons mené à bien une étude analysant la présence ou l'élision du /ʀ/ en position finale de mot (Wu, Gendrot, et al., 2019). Pour détecter automatiquement la présence vs. élision du /ʀ/ et/ou du schwa, nous nous sommes appuyés sur le système d'alignement comme analyseur (pour plus de détails, voir Adda-Decker et al., 2008, 2013), l'efficacité de cette méthode ayant été vérifiée par une comparaison avec une annotation manuelle sur un extrait des données (voir Wu et al. 2019 pour plus de détails). Des /ʀ/ et des schwas optionnels ont été inclus comme variantes dans le dictionnaire de prononciations du système. La présence ou l'absence de /ʀ/ a donc été décidée automatiquement en utilisant l'alignement forcé puisque celui-ci sélectionne la variante qui correspond le mieux à la réalisation acoustique du locuteur. La Figure 56 illustre des spectrogrammes de deux productions alignées automatiquement du mot « quatre » avec et sans le /ʀ/.

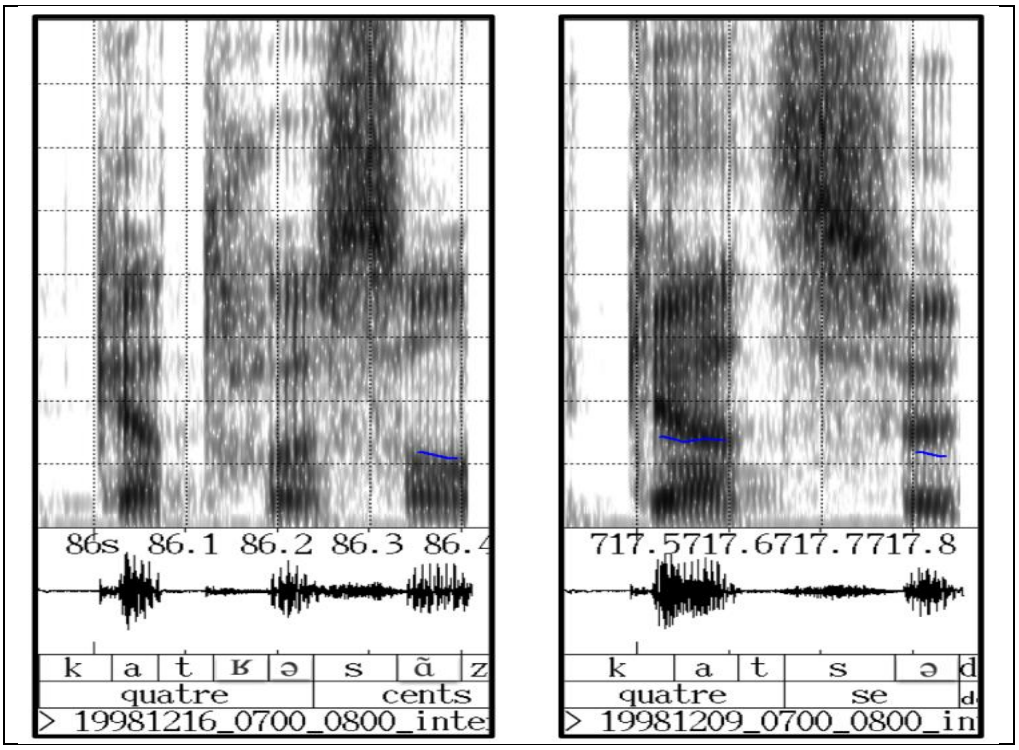


Figure 56 : le mot "quatre" avec (à gauche) et sans (à droite) le /R/ : exemple d'alignement automatique du LIMSI, d'après Wu et al. (2019)

La réalisation du /R/ en position post-consonantique finale est connue pour être instable en français, et celle-ci peut prendre régulièrement les formes suivantes : la forme canonique C/R/# (par exemple quatre [katR]), la forme canonique + l'insertion du schwa C/R/ə# (par exemple quatre [katRə]), l'élision du /R/ sans le schwa (quatre [kat]), ou avec un schwa (quatre [katə]).

La séquence C/R/#C (« quatre sacs ») par exemple est une séquence qui tend à violer le principe de sonorité. En analysant différents styles de parole, nous avons pu vérifier différentes stratégies qui permettent d'éviter cette contrainte et qui conditionnent l'apparition du /R/. Les résultats (voir Figure 57) ont montré que la présence du /R/ est dépendante de la présence du schwa final de mot : /R/ n'est (presque) jamais absent lorsque le schwa est inséré (par exemple [katRə#sak]). Plus la parole est formelle (dans la parole journalistique), plus on observe une insertion du schwa, et moins /R/ est élidé (ESTER: 65 % d'insertions du schwa, 15 % d'élisions du /R/ ; NCCFr : 13 % d'insertions du schwa, 69 % d'élisions du /R/).

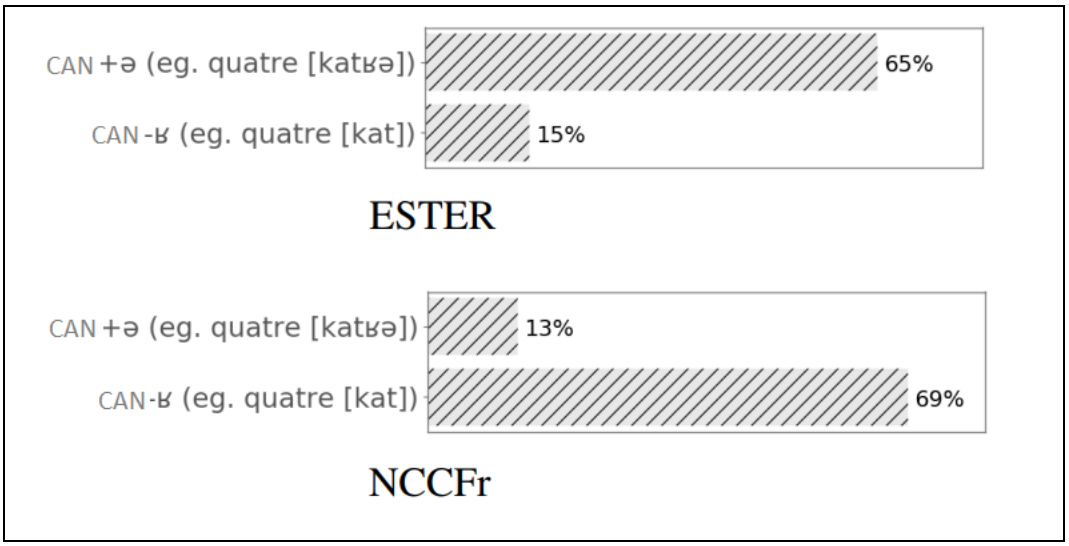


Figure 57 : réalisations de la séquence C/ʀ/#C pour deux styles de parole (journalistique en haut, et spontané en bas). CAN+ ə correspond à la forme C/ʀ/ sans le schwa, CAN - /ʀ/ correspond à la forme C/ʀ/ avec élision du /ʀ/, d'après Wu et al. (2019)

Dans une étude parallèle (Wu et al., 2019), nous avons analysé l'influence du contexte post-lexical (#C vs. #V) sur la production de ce cluster, en nous intéressant également au style de parole (journalistique vs. spontané). La Figure 58 montre les résultats d'alignement pour la forme canonique et ses variantes en fonction de la classe phonémique qui suit la frontière de mot (consonne ou voyelle), et en fonction du style de parole. Quand le contexte post-lexical est une consonne en parole journalistique, la forme la plus commune est « CAN + ə », mais lorsque le contexte post-lexical est une voyelle, la forme la plus commune est la forme canonique « CAN » (C/ʀ/#). Par contre, ces résultats changent complètement en parole spontanée où la forme « CAN - /ʀ/ » est privilégiée quand le contexte post-lexical est une consonne. Si le contexte post-lexical est une voyelle, la forme « CAN » reste la plus fréquente, mais à peine plus que la forme « CAN - /ʀ/ ».

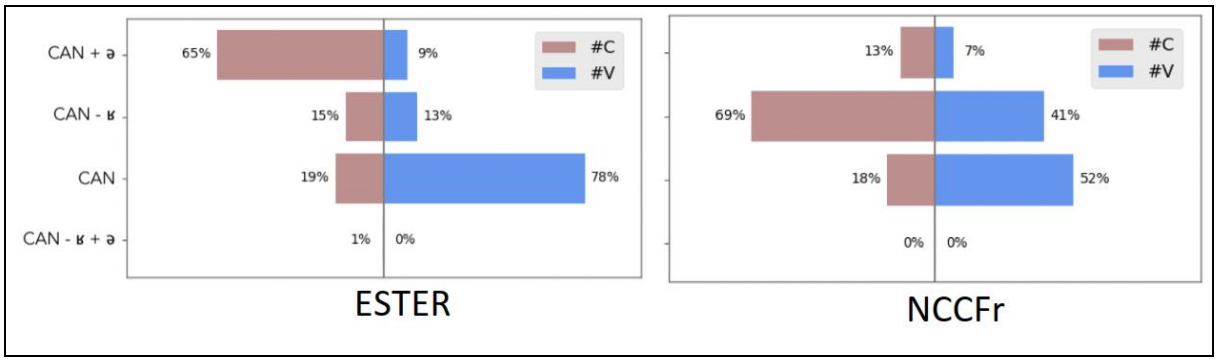


Figure 58 : taux d'alignement des quatre variantes de prononciation (CAN = forme canonique ; CAN, CAN+ ə, CAN - ʀ, CAN - ʀ + ə) en fonction du contexte post lexical (#C en marron, #V en bleu). A gauche le corpus de parole journalistique, à droite le corpus spontané

4.3.2.3 Apports didactiques

L'utilisation de ces travaux nous a permis d'aborder le thème de la remédiation pour la prononciation des sons dans le cadre de l'apprentissage d'une langue étrangère. Les locuteurs du chinois sont connus pour avoir des difficultés pour la prononciation du /ʀ/ voisé (dans le cas de la séquence /ʀ/V par exemple) quand ils apprennent le français : ils ont tendance à produire leur son natif /x/ qui est une fricative uvulaire sourde. Nous avons comparé (Wu et al., 2015) deux méthodes d'apprentissage afin d'améliorer leur prononciation : une méthode classique utilisant des explications vulgarisées de la réalisation articulatoire et une méthode utilisant l'ultrason lingual afin de montrer les mouvements de la langue en temps réel. Deux groupes distincts de dix locuteurs chinois ont participé à cette expérimentation. Des mesures acoustiques de HNR et de COG sur leurs productions pré- et post-entrainement nous ont permis de quantifier leurs progrès. La Figure 59 montre qu'en utilisant le feedback réalisé au moyen de l'ultrason en temps réel, les apprenants ont amélioré leur capacité à s'éloigner de leur prononciation du /x/ mandarin pour s'approcher du /ʀ/ (voisé) français à la fin de l'entrainement.

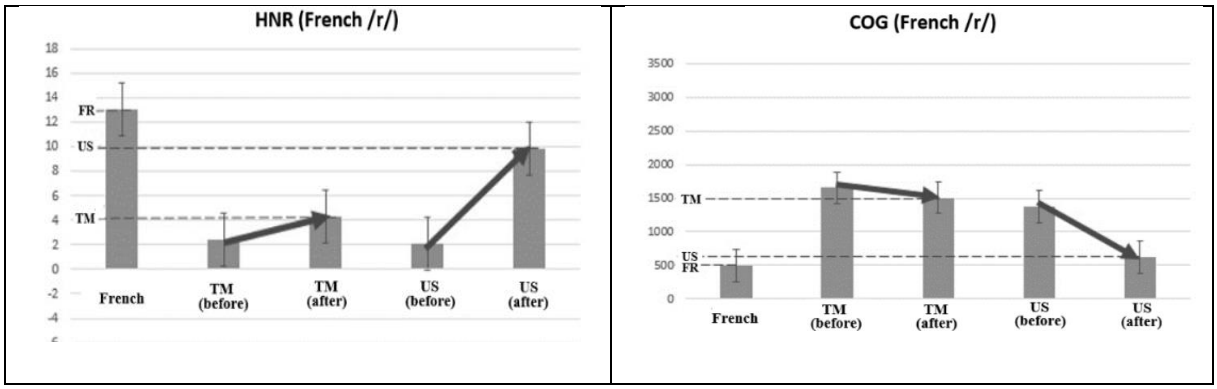


Figure 59 : valeurs de HNR (à gauche) et COG (à droite) du /ʀ/ (suivi d'une voyelle) prononcé par un locuteur natif du français et par des apprenants chinois avant et après l'entrainement des deux groups (TM : méthode traditionnelle ; US : méthode incluant la visualisation par l'ultrason), d'après Wu et al. (2015)

Dans la lignée de cette étude, nous avons mis en place un outil d'aide à la prononciation du /R/ français disponible en ligne sous le nom de « Reper Me ! » en collaboration avec Emmanuel Ferragne, mettant en œuvre les résultats acoustiques détaillés ci-dessus à partir des variations de COG et de HNR mais non encore validé dans une étude expérimentale : (http://www.univ-paris3.fr/anr-reper-231657.kjsp?RH=1378813965532&RF=1378805498389)

4.3.2.4 Discussion

Le lieu d'articulation du /ʀ/ français le rend particulièrement sujet à une forte variation, et un changement subtil de constriction lui permet de passer d'une approximante voisée à une fricative sourde. S'il est probable que dans une opposition entre deux fricatives sourdes et sonores (par exemple /s/ et /z/), la pression intra-orale plus importante des obstruantes sourdes induise une constriction plus importante, cette différence ne devrait pas être aussi importante que pour le changement de sonorité du /ʀ/. On peut conclure dans un premier temps que /ʀ/ peut être considéré comme une fricative lorsqu'il est réalisé dans sa forme hyper-articulée (plus longue avec une constriction plus importante), et comme une approximante lorsqu'il est plus court. Les variations observées lorsque le débit est plus rapide vont également dans ce sens. Dans une modélisation des

fricatives pharyngales et uvulaires, Yeou et Maeda (1995) ont suggéré que celles-ci devraient être considérées comme des approximantes et non comme des fricatives, notamment parce que les valeurs de débit d'air sont plus élevées que pour les autres fricatives. De par ses propriétés phonotactiques similaires à une consonne liquide et non à une fricative, nous concluons de ces travaux que le /ʀ/ est avant tout une approximante voisée dont la réalisation en contexte sourd se fait sous la forme d'une fricative. Le fait que la réalisation voisée et approximante du /ʀ/ soit la plus fréquente (la moins « marquée » au sens phonologique) et la plus intuitive (dans les mots monosyllabique #/ʀ/V# par exemple) appuie cette hypothèse. Considérant que le /ʀ/ français a beaucoup changé diachroniquement, une évolution de la réalisation du /ʀ/ en français dans les prochaines années pourrait être observée à la lumière de ces résultats. Dans la section suivante, je présente une étude sur corpus qui se consacre précisément à l'apport diachroniques des grands corpus.

4.3.3 La fusion e/ɛ : apports diachroniques des grands corpus multilocuteurs

Les grands corpus de parole non préparée ont souvent l'avantage de fournir des données provenant d'un nombre important de locuteurs, ce qui autorise la mise en évidence d'une tendance globale. Cette tendance portée par une majorité de locuteurs permet d'amorcer une évolution dans les langues, et le français n'en est pas exempt bien sûr malgré la normalisation apportée par les media nationaux et les efforts consentis par l'Académie Française. Dans cette dernière partie consacrée aux apports linguistiques des grands corpus, je présente un travail (Gendrot et Audibert, 2019) qui appuie l'hypothèse que les voyelles mi-ouvertes /e/ et /ɛ/ sont ancrées dans un processus de fusion phonétique pour ne devenir qu'une.

Dans les ouvrages de phonétique et phonologie, on distingue pour le français deux voyelles mi-ouvertes antérieures : /e/ et /ɛ/, dans un système vocalique dense composé de 12 voyelles orales (en incluant /ɑ/ et /ə/) et 4 voyelles nasales (si on inclut /œ̃/). Toutes ces voyelles, si elles font effectivement partie du système vocalique du locuteur (à l'exception possible de /ə/ que nous ne considérerons pas ici) sont considérées comme des phonèmes puisqu'elles peuvent former des paires minimales telles que [bat] (« batte ») ~ [bɛt] (« bête »), ou [ge] (« gué ») ~ [gɛ] (« guet »). En latin classique l'opposition /e/ ~ /ɛ/ n'existait pas et consistait en une opposition de durée /e/ ~ /e:/. Notons que la coexistence d'une distinction de timbre entre ces deux voyelles était très probable, ce qui a pu favoriser l'émergence de cette nouvelle opposition /e/ ~ /ɛ/, que l'on retrouve aussi pour la paire /o/ ~ /ɔ/ lorsque l'opposition phonologique de durée a disparu (Vaissière, 2001).

Les phonèmes /e/ et /ɛ/ sont régis par une distribution partiellement complémentaire. En effet, en syllabe fermée, seule la voyelle mi-ouverte /ɛ/ peut être réalisée ([bɛt] « bête »), alors qu'en syllabe ouverte on peut trouver les deux variantes (mi-ouverte et mi-fermée), ce qui amène parfois à considérer que /e/ et /ɛ/ sont deux variantes d'un archiphonème /E/. Ces règles sont également dépendantes des phénomènes de resyllabation en français, dus à l'élision du /ə/ ou d'un autre phonème. Dans le cas du mot « médecin » par exemple, les locuteurs pourront prononcer [me.də.sɛ̃], [mɛ.də.sɛ̃] ou bien [mɛd.sɛ̃] mais pas [med.sɛ̃]. Il est à noter que les locuteurs n'ont pas une réelle conscience du phonème qu'ils produisent, étant probablement influencés par l'orthographe et dans le cas présent par l'accent aigu du « é ». Nous émettons l'hypothèse qu'ils produisent de plus en plus une version intermédiaire entre /e/ et /ɛ/, comme cela a pu être le cas pour /a/ et /ɑ/. Un autre phénomène linguistique, celui de l'harmonie vocalique (Nguyen et Fagyal, 2008) peut intervenir dans le cas de ces voyelles au timbre intermédiaire, prenons le cas par exemple des mots « César » et « désir », de par la voyelle présente dans la deuxième syllabe [a] et [i] respectivement, celle-ci pourra ouvrir ou bien

fermer la première voyelle qui s'en trouvera alternativement plus proche d'un /ɛ/ ou d'un /e/. Il est probable bien sûr que, de par leur statut de voyelle mi-haute et mi-basse, l'écart entre /e/ et /ɛ/ soit naturellement plus réduit que celui entre /e/ et /i/ ou /ɛ/ et /a/, mais notre objectif était de montrer ici un processus en cours dans lequel ces deux voyelles tendent vers une seule entité.

L'avantage des grands corpus de parole non contrôlée se présente très naturellement ici puisqu'ils permettent d'analyser une grande quantité d'occurrences produites naturellement. Ils permettent de se dissocier des productions élicitées peu écologiques sur la base de mots orthographiques. Au contraire d'un petit corpus construit ad-hoc, un grand corpus de parole contiendra des contextes phonétiques et lexicaux non contrôlés, et donc qui ne respectent pas le principe du « toutes choses égales par ailleurs » : il faudra alors à la fois filtrer les contextes semblables et s'appuyer sur un nombre important et équilibré de données qui permettront statistiquement d'effacer les effets des cas particuliers.

Notre but étant de mettre en évidence le processus en cours qui tend à fusionner /e/ et /ɛ/ en français standard, nous émettons l'hypothèse que s'il existe un processus de fusion entre /e/ et /ɛ/, ce processus sera plus marqué en parole spontanée qu'en parole journalistique où le degré d'articulation reste soutenu avec un débit de parole moins important (Audibert et al., 2015). Dans cette analyse, nous avons sélectionné exclusivement les voyelles finales de mots, voyelles au sein de la syllabe portant l'allongement en français et donc mieux articulées. Si la distinction entre /e/ et /ɛ/ est maintenue, il y a fort à parier qu'elle soit avant tout maintenue dans cette position, et inversement si cette distinction s'estompe, elle le sera d'autant plus dans les autres positions lexicales moins renforcées.

Nous avons également retiré les mots grammaticaux les plus fréquents, qui contreviennent souvent aux règles phonologiques de la langue et peuvent avoir un fort impact statistique de par leur fréquence. Dans les corpus présentés ici, 110 mots représentent 21 % des occurrences de /e/ et /ɛ/ au total. Pour ces mots dont la phonétisation est variable, on ne peut pas considérer la catégorie de la voyelle comme indiquant la prononciation de référence : ils ont donc été retirés du reste de l'analyse. C'est également le cas du mot « ouais », fréquemment produit isolément avec un allongement important en parole spontanée.

Les mesures de formants ont été réalisées sur un total de 193 000 occurrences des voyelles /e/, /ɛ/ et /a/ produites par 157 locuteurs, selon la méthode employée précédemment, à savoir en mettant en place des filtres de valeurs acceptables pour chaque phonème en fonction du sexe du locuteur. Ces valeurs obtenues en Hertz ont ensuite été converties en Bark afin de mieux correspondre à la perception humaine (Traunmüller et Eriksson, 1995), puis une distance euclidienne a été calculée dans l'espace formé par les deux premiers formants F1 et F2. Rappelons que pour une valeur inférieure à 0.4 Bark, la distance n'est pas perçue par les auditeurs (Kewley-Port et Zheng, 1999). Nous comparons donc dans les figures suivantes les distances dans l'espace F1/F2 entre /e/ et /ɛ/, mais également entre /ɛ/ et /a/. Ces mesures ont été effectuées pour chaque locuteur puis moyennées entre locuteurs.

Sur les figures ci-dessous, quatre conditions sont analysées séparément :

- La distance entre /ɛ/ et /a/ en position finale de mot (120k voyelles, dont 72 % correspondent à des paires minimales)

- La distance entre /e/ et /ɛ/ en position finale, tous mots confondus à l'exception des 21 % exclus par filtrage (76k voyelles)

- La distance entre /e/ et /ɛ/ pour l'opposition entre « et » et « est » seulement (29k voyelles)

- La distance entre /e/ et /ɛ/ en position finale pour l'opposition de prononciation théorique entre l'infinitif en « er » et l'imparfait ou conditionnel « ais », « ait » ou « aient » (23k voyelles)

Pour chacune de ces conditions, nous comparons les mesures effectuées dans les deux corpus ESTER (parole journalistique) et NCCFr (parole spontanée). La Figure 60 ci-dessous montre que l'écart entre /ɛ/ et /a/ (première double barre à gauche) est environ deux fois supérieure aux écarts entre /e/ et /ɛ/ dans les trois conditions définies ci-dessus (les trois doubles barres à droite de la figure).

Nous pouvons également observer que la condition NCCFr présente des distances moins élevées dans toutes les catégories ce qui valide l'hypothèse selon laquelle le processus de rapprochement est plus important en parole spontanée. La Figure 61 présente des résultats identiques pour les femmes, à l'exception de la différence entre « et » et « est » pour lesquels la distance est plus élevée dans le corpus de parole spontanée.

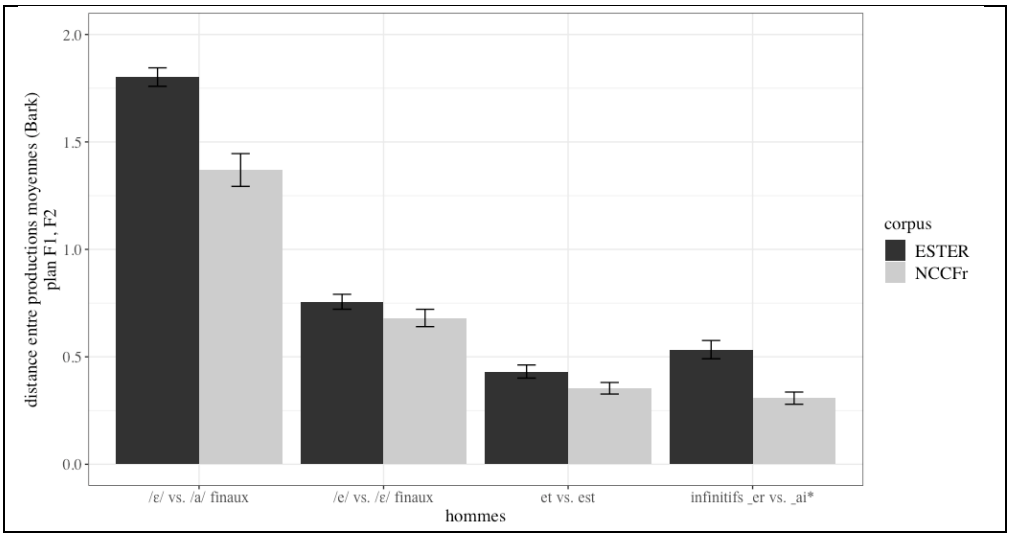


Figure 60 : distance acoustique mesurée en Bark pour les hommes entre /ɛ/ et /a/, et entre /e/ et /ɛ/ pour plusieurs conditions lexicales, et dans deux corpus de parole, journalistique (ESTER) et spontanée (NCCFr), d'après Gendrot et Audibert (2019)

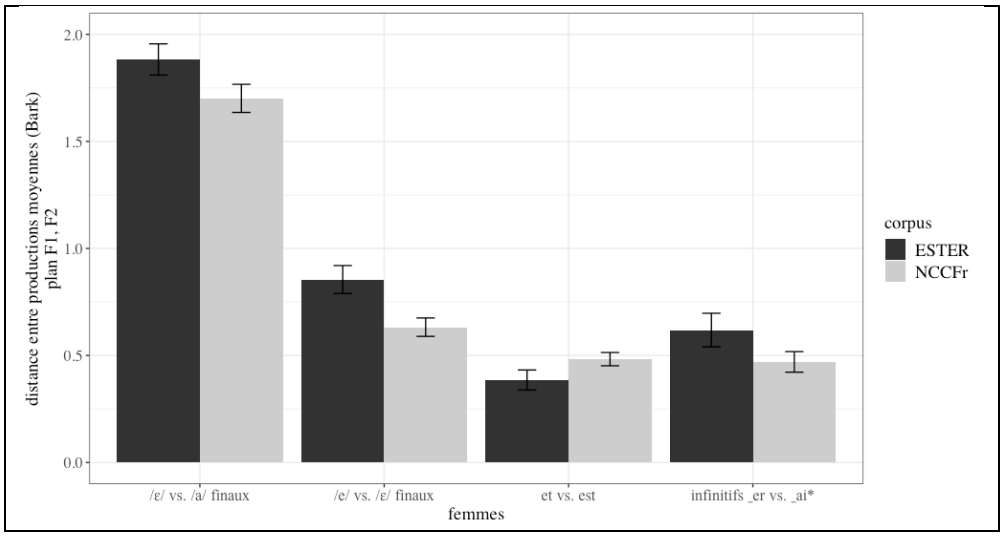


Figure 61 : distance acoustique mesurée en Bark pour les femmes entre /ɛ/ et /a/, et entre /e/ et /ɛ/ pour plusieurs conditions lexicales, et dans deux corpus de parole, journalistique (ESTER) et spontanée (NCCFr), d'après Gendrot et Audibert (2019)

La Figure 62 présente des ellipses de dispersion intégrant 95 % de toutes les occurrences des voyelles /a/, /e/ et /ɛ/ analysées dans la Figure 60 et la Figure 61 pour le corpus NCCFr, cette fois-ci présentées en Hertz telles qu'elles ont été mesurées. Les ellipses de dispersion permettent de visualiser la variation des voyelles en condition naturelle lorsque l'on analyse un grand nombre de données. Le recoupement entre /e/ et /ɛ/ est presque total, alors qu'il n'est que partiel pour les voyelles /a/ et /ɛ/, ce qui confirme une nouvelle fois l'hypothèse selon laquelle les voyelles /e/ et /ɛ/ se sont rapprochées acoustiquement.

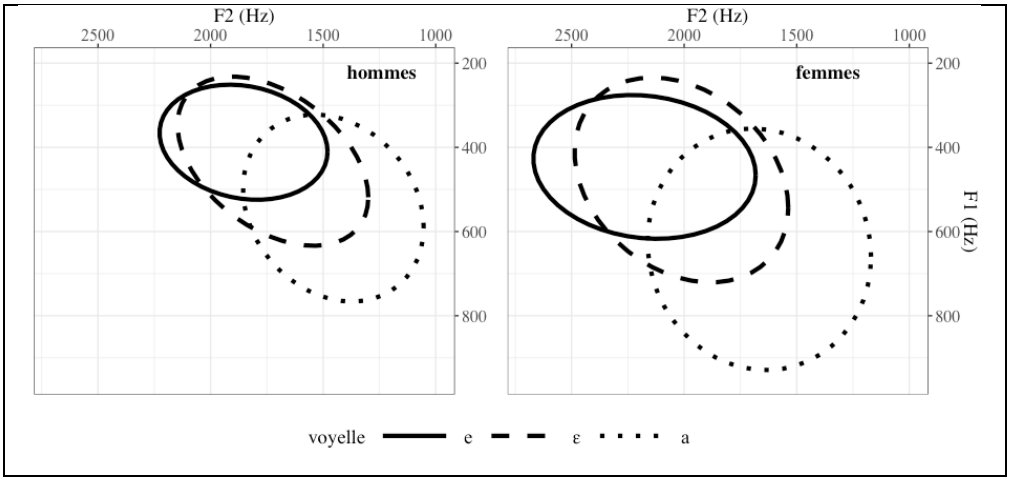


Figure 62 : ellipses de dispersion intégrant 95 % de toutes les occurrences des voyelles /a/, /e/ et /ɛ/ pour les hommes (gauche) et les femmes (droite). Mesures effectuées sur le corpus NCCFr, d'après Gendrot et Audibert (2019)

Certains contextes tels que la position initiale ou la syllabe initiale de mot n'ont pas été intégrés à notre analyse car ils ne représentaient pas une quantité de données suffisamment conséquente ou équilibrée. Il aurait pu être utile également de prendre en compte la durée phonétique des voyelles

qui a une importance non négligeable dans leur réalisation (Gendrot et Adda-Decker, 2005, cf. Figure 4), mais la prise en compte de nouvelles catégories d'analyse nécessite de multiplier la quantité d'occurrences à analyser afin de pouvoir maintenir la normalisation apportée par les grands corpus. A l'heure où nous écrivons ces lignes, il devient possible d'analyser non plus des dizaines d'heures de parole, mais des centaines ou des milliers d'heures, ce qui permettra sans nul doute d'affiner nos connaissances. Ainsi, la question du maintien ou non de la différence de prononciation entre le futur en « ai » et le conditionnel en « ais », qui n'a pu être prise en compte dans cette étude faute d'un nombre suffisant d'occurrences, pourrait trouver une réponse via le recours à des corpus de plus grande taille encore.

Une étude est actuellement en cours pour aller au-delà des simples moyennes par locuteur, et qui s'inscrit dans le cadre de mon projet décrit en dernière section. En effet, la diversité des locuteurs fait qu'inévitablement, certains locuteurs de par leur idiolecte vont maintenir l'opposition entre ces deux voyelles. On peut voir dans la Figure 63 le locuteur conservant le mieux l'opposition entre /e/ et /ɛ/, pour lequel il semble que le /e/ se rapproche acoustiquement d'un /i/. Un approfondissement de ce type permet de s'intéresser non plus aux grandes tendances, mais à la variation inter-locuteurs et surtout à la caractérisation phonétique du locuteur trop peu documentées dans mes travaux. C'est sur ces aspects de la variation que porteront mes prochains travaux.

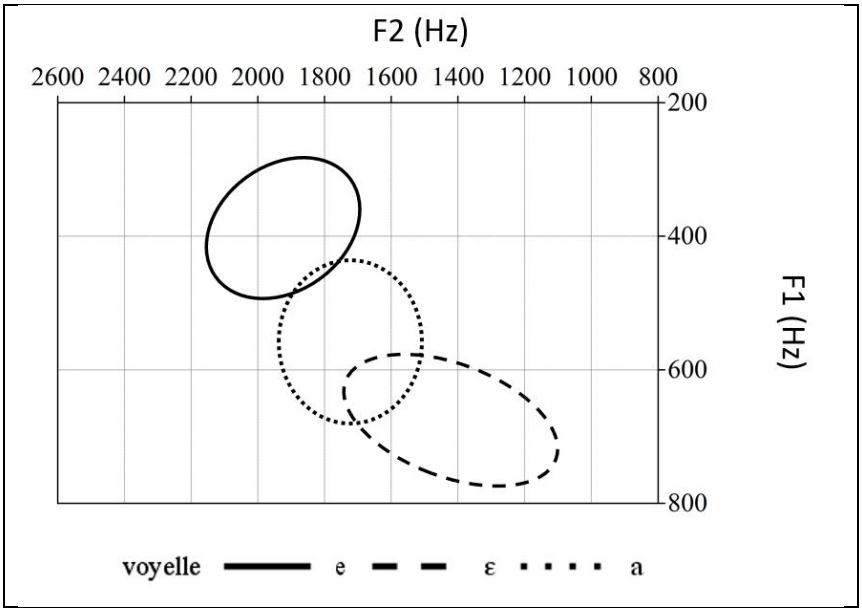


Figure 63 : ellipses de dispersion intégrant 95 % de toutes les occurrences des voyelles /a/, /e/ et /ɛ/ pour le locuteur ayant les productions les plus distinctes. Mesures effectuées sur le corpus NCCFr (non publié)

4.3.4 Résumé des trois études segmentales

Cette section a permis de présenter trois analyses de type *phonologie de laboratoire* sur des grands corpus de parole non contrôlée.

La partie concernant le schwa a montré que deux types de réduction peuvent coexister, une réduction phonologique qui permet au locuteur de choisir entre la forme lexicale avec ou sans schwa, et une réduction phonétique semblable à celle des autres voyelles.
La deuxième étude postule que la forme non voisée du /ʀ/ peut être considérée comme la réalisation hyper-articulée de la forme voisée non marquée, et montre que la variation est grandement influencée par la position prosodique et le style de parole, au-delà du contexte consonantique.
Pour finir, j'ai présenté une étude sur des corpus de parole journalistique et spontanée qui a pour but de montrer que malgré des disparités inter-locuteurs, et dans des positions lexicales qui restent encore ciblées, un processus de fusion entre /e/ et /ɛ/ est en cours en français.

Ces études ont également été l'occasion de tester perceptivement les variations spectrales mesurées et ainsi valider leur pertinence dans le cadre de la communication parlée. Plusieurs aspects méthodologiques fondamentaux sont également détaillés dans l'ensemble de cette section : équilibrage des données, utilisation complémentaire de petits corpus de données articulatoires, utilisation d'un système d'alignement pour analyser la variation, normalisation des mesures, analyses statistiques.
Tous ces aspects techniques ainsi que la complémentarité des grands corpus de parole non contrôlée avec des corpus de parole lue sont abordées dans la section suivante, dédiée à une discussion générale, avant de présenter mes travaux en cours en guise de projet de recherche.

4.4 Discussion et projection

4.4.1 La problématique des grands corpus : apports et limites

La quasi-intégralité des travaux présentés dans ce travail portent sur l'analyse de grands corpus de parole non contrôlée où plusieurs dizaines d'heures d'enregistrements ont été segmentés automatiquement au niveau du phonème et du mot. Les premiers corpus de ce type en France[4] - par exemple le corpus ESTER (Galliano et al., 2006)- étaient à l'origine utilisés pour la reconnaissance de la parole, et l'alignement était utilisé uniquement pour reconnaitre les triphones qui menaient ainsi à la reconnaissance des mots. Les linguistes ont su rapidement reconnaitre l'intérêt de travailler en collaboration avec des laboratoires en informatique afin d'utiliser les corpus d'entrainement pour la reconnaissance de la parole pour des problématiques de recherche en linguistique. En compensation, les laboratoires d'informatique pensaient pouvoir déduire de ces analyses linguistiques des possibilités d'amélioration des systèmes de reconnaissance de la parole. Dans cette discussion générale, je vais présenter les intérêts principaux à utiliser ces corpus, mais également leurs limites, et de façon plus générale les apports et limites des mesures phonétiques automatiques telles qu'elles sont effectuées le plus souvent dans les travaux actuels en phonétique.

[4] *Notons que le corpus Switchboard pour l'anglais américain était distribué par LDC dès 1992.*

4.4.2 Précision de l'alignement et validité des mesures

J'ai présenté une évaluation de la précision des systèmes d'alignement phonémique dans la section 4.1.3.3, et j'ai pu montrer que pour l'étude des voyelles, il est plus raisonnable de réaliser des mesures médianes. Ces travaux datant de 2008, la question se pose quant à l'évolution de l'alignement en près de 13 ans. La principale différence est venue d'un accès aux systèmes d'alignement de plus en plus faciles à utiliser. Les systèmes les plus connus sont webMAUS (Kisler et al., 2017) et ProsodyLab-Aligner (Gorman et al., 2011), tous deux tirés du toolkit de reconnaissance automatique de la parole HTK (Young et al., 2006) qui est basé sur des modèles acoustiques à partir de modèles de Markov cachés (HMM). Un autre système est apparu plus récemment : le système Montreal Forced Aligner MFA (McAuliffe et al., 2017) tiré du toolkit KALDI (Povey et al., 2011), basé quant à lui sur des réseaux de neurones profonds. A notre connaissance, il n'existe pas encore de comparaison quantitative de la qualité de la segmentation fournie par les deux outils, ou plus précisément de comparaison entre HTK et Kaldi, mais il semble que le plus grand nombre d'heures de corpus entrainés par le premier lui confère toujours un avantage. D'autres systèmes d'alignement dits « clique-bouton », et accessibles sans connaissance informatique particulière ont également vu le jour, comme Easyalign (Goldman, 2011) et SPPAS (Bigi et Meunier, 2018). L'avantage de ces outils multiples, outre leur facilité d'utilisation, consiste en l'ajout de réglages qui permettent de s'adapter aux besoins de l'utilisateur. SPPAS permet par exemple l'ajout de mots dans le dictionnaire de prononciation, ou bien d'entrainer les modèles acoustiques sur ses propres données (voir également Train&Align de Brognaux et al., 2014). Les corpus utilisés dans ce manuscrit ayant été transcrits orthographiquement manuellement dans un premier temps, et les systèmes d'alignement utilisés étant ceux de laboratoires spécialement impliqués dans la reconnaissance automatique de la parole[5], nous avons de bonnes raisons de croire que la segmentation automatique en phonèmes que nous avons utilisée est au moins d'aussi bonne qualité que celle dispensée par les outils sus-cités.

De manière plus générale, l'argument fréquemment présenté pour valider l'analyse de grands corpus automatiquement segmentés est que la quantité de données (souvent plusieurs centaines de milliers de phonèmes analysés) permet de neutraliser le bruit généré par l'imprécision de la segmentation. J'ai montré que cet argument peut être valable si l'on considère des moyennes générales, mais si l'on considère un phénomène dans le détail, comme l'étiquetage et la segmentation du schwa dans un contexte segmental précis, il reste possible de trouver des biais systématiques. Un autre argument mis en avant est qu'une segmentation automatique – par opposition à une segmentation humaine - a l'avantage d'être prévisible puisqu'elle ne se fatigue pas, et qu'elle est objective et apportera une réponse systématique à un input systématique. Le dernier point n'est plus tout à fait exact lorsqu'il s'agit d'intelligence artificielle, où le système d'alignement peut également s'améliorer au fur et à mesure que la quantité de données introduite permet d'affiner les réponses attendues. Si l'on redoute malgré tout que des imprécisions de segmentation génèrent des erreurs d'analyse, il existe des solutions. Par exemple, il est possible de filtrer les résultats obtenus à partir de fourchettes de valeurs attendues, mais cette méthode reste discutable et sera abordée dans le paragraphe suivant. Une deuxième solution serait d'effectuer des mesures toutes les 10 millisecondes, ce qui permettra d'analyser la modulation de valeurs acoustiques : la variabilité des données d'une fenêtre d'analyse à la suivante (comme pour une analyse de f0 basée sur un signal quasi périodique) permettra alors de

[5] *Plus récemment, les laboratoires se sont moins impliqués dans la reconnaissance automatique de la parole 'standard' puisque cette dernière a été prise à son compte par des grands groupes industriels tels que Apple, Google, Amazon, etc.*

localiser les mesures problématiques. Dans la continuité de ce raisonnement, la notion même d'analyse phonémique peut s'effacer au profit d'une analyse dynamique sur plusieurs segments : l'analyse d'une séquence de phonèmes dans son ensemble, et non plus de phonèmes isolés, permet de réduire les problèmes de précision de la segmentation temporelle. La compréhension des variations syntagmatiques devrait également permettre de mieux appréhender les phénomènes linguistiques, mais ces mesures sont encore trop peu utilisées dans la littérature phonétique de par la difficulté à traiter statistiquement ce type de résultats.

La validité des mesures acoustiques est également un point sur lequel nous devons porter notre réflexion. Lorsque l'on considère les mauvaises détections, le cas le plus évident correspond à des segmentations phonémiques trop décalées et/ou un renforcement des harmoniques pas suffisamment proéminent pour détecter correctement les formants. Le premier cas peut être évité en prenant des mesures dans la partie médiane de la voyelle comme précisé ci-dessus. Pour le deuxième cas, de multiples circonstances peuvent être en cause. Les mesures de formants peuvent être erronées si la voyelle est dévoisée par exemple, mais il est alors peu utile d'essayer de mesurer des formants de la même façon qu'il serait peu utile de mesurer une f0 pour de la voix craquée (Gendrot et al., 2004). Dans un autre cas de figure, nous avons montré (Gendrot et Adda-Decker, 2008) que des voyelles hyper-articulées favorisaient une mauvaise détection des formants, aussi contradictoire que cela puisse paraitre. Par exemple, un /y/ est caractérisé par un F2 et un F3 proche (voir section 4.1.4.3), et un /y/ plus long et/ou hyper-articulé sera caractérisé par un rapprochement supplémentaire de ces deux formants ce qui les rendra plus sujets à une mauvaise détection de la part d'un algorithme de détection automatique des formants : l'algorithme détectera alors un seul formant en lieu et place des deuxième et troisième formants. Une analyse des formants du /i/ posera des problèmes identiques, mais pour F3 et F4. Dans ces cas précis, il serait préjudiciable de se débarrasser des mauvaises détections de formants car les voyelles sont quant à elles de très bons prototypes. Quant au /u/, caractérisé par F1 et F2 proches, les deux résonances de Helmholtz seront d'autant plus difficiles à distinguer que le /u/ est hyper-articulé (Gendrot et Adda-Decker, 2008). Comment résoudre ce problème ? Il est possible de forcer une détection de formants dans une fourchette de fréquences plus restreinte, mais cette méthode pourra à l'inverse laisser échapper des cas où une voyelle hypo-articulée a des formants plus éloignés qu'à l'habitude. Dans tous les cas, ce raisonnement impose de connaitre les variations des formants avant de les mesurer ce qui peut poser un problème de circularité du raisonnement : on ne garde que les résultats que l'on s'attend à obtenir et le reste sera traité comme anormal et sera rejeté. Cela pose le problème des mesures de formants telles qu'elles sont réalisées par de nombreux phonéticiens. Les algorithmes de détection de formants se sont-ils suffisamment améliorés en plus de dix ans pour limiter ce type d'erreurs ? Un article récent (Chen et al., 2019) a montré qu'un des problèmes vient de l'estimation faite sur l'enveloppe spectrale de type LPC qui se base sur la source et donc sur la f0 ; il est souvent reconnu que les mesures de formants sont biaisées vers l'harmonique le plus proche, ce qui peut s'avérer problématique pour des f0 élevées. De même, l'estimation ne sera pas identique en fonction de la f0 intrinsèque de la voyelle, voire des variations de f0 intra-locuteur, ce qui implique une remise en cause de la fiabilité de ces mesures. /i/ a une valeur intrinsèque de f0 plus élevée que les autres voyelles et verra ses valeurs de formants sous-estimées. Des algorithmes basés sur des prédictions linéaires pondérées ont été proposés (Gowda et al., 2017) pour résoudre ce problème spécifique, mais ceux-ci nécessitent des informations précises sur l'impulsion glottale (par le biais d'un Electro-Glotto-Graphe par exemple), ce qui limite leur utilisation. Le retour à une analyse semi-automatique, par exemple avec utilisation de spectrogrammes 'réassignés' (reassigned spectrograms), serait finalement la solution la plus fiable (Shadle et al., 2016), mais repose le problème de l'analyse de très grandes quantités de données. Des travaux dans le sens d'un système expert (i.e. à base de règles) appliqué sur des analyses automatiques de type LPC

pourraient s'avérer utiles. Il est possible également que des tentatives de rétro-ingénierie permettent de réaliser un filtrage inverse plus fiable en se basant sur des grands corpus, et ainsi déterminer le flux glottique pour améliorer la détection de formants, mais à ma connaissance, les corpus ne servent pour le moment qu'à confirmer les modélisations effectuées sur de la parole synthétisée ou des voyelles tenues (Drugman et Dutoit, 2019).

Quelles sont les autres solutions ? Est-il possible de se limiter à une analyse du centre de gravité spectral tel qu'utilisé pour les fricatives (Jongman et al., 2000) ? Si cette mesure a l'avantage d'être plus robuste, elle sera évidemment moins précise, notamment pour distinguer les voyelles arrondies des voyelles non arrondies. Il semble au final que les mesures de formants proposées actuellement soient un pis-aller dont le phonéticien devrait se contenter. Les laboratoires d'informatique ont depuis longtemps réglé ce problème en ayant recours à des MFCC et non pas à des formants en Hertz, et certains ont tenté par intermittence de revenir à ces méthodes. Leur principale limite est la difficulté de les relier à des informations articulatoires. Ces interrogations doivent pousser les phonéticiens à s'interroger sur les raisons pour lesquelles ils effectuent des analyses de formants. Les formants permettent de caractériser articulatoirement des voyelles selon la disposition habituelle du triangle vocalique et de relier l'articulation à des phénomènes linguistiques. Ces caractéristiques articulatoires peuvent ainsi être interprétées en traits phonologiques selon le système de la langue. Mais dans d'autres cas, où l'on souhaite quantifier le degré de réduction d'un phonème, ou bien modéliser sa présence, il n'est pas systématiquement nécessaire de s'en tenir à des mesures de formants et d'autres mesures peuvent être envisagées : des MFCC, une probabilité de bonne classification, etc. Ce point est également valable pour l'ensemble des mesures acoustiques considérées par les phonéticiens.

Concernant les mesures de f0, comme précisé ci-dessus, elles sont souvent imprécises quand la voix sort des limites de la modalité (sans même aborder la parole pathologique, ou la parole dans des conditions bruitées) comme par exemple la voix soufflée ou craquée. La f0 est une mesure insuffisante à bien des égards car elle ne prend pas en compte le timbre de la voix dans son ensemble : un ténor et un baryton pourraient avoir la même valeur de f0, par exemple un Do_2 (130.8 Hz), alors que cela pourra apparaitre perceptivement comme deux hauteurs très différentes (on retrouve la même nuance entre un violon et un violoncelle). De même la f0 d'un homme dont la f0 moyenne est de 100 Hz peut atteindre aisément 250-300 Hz lors d'une montée de continuation et restera très différente de la voix d'une femme dont la f0 moyenne est de 250 Hz. Bien sûr les valeurs de f0 peuvent être présentées à l'aide d'une autre échelle, celle des demi-tons étant la plus connue, mais celle-ci ne tient toujours pas compte des différences de voix, et seule une mesure de f0 combinée à une mesure de timbre sera réellement pertinente dans cette optique.

Quelles sont les autres mesures acoustiques possibles ? Une grande variété de mesures a été évaluée dans la littérature, et plus particulièrement ces dix dernières années. Est-ce là un signe que les mesures classiques proposées ne permettent de répondre de façon satisfaisante aux questions des phonéticiens ? Il est vrai qu'une forte variabilité se retrouve dans les résultats ; cette variabilité a été prise en compte plus récemment par des analyses statistiques plus complexes et plus conservatrices comme les modèles mixtes à effets aléatoires. Mais une classification de type LDA (analyse linéaire discriminante) faite à partir de mesures classiques aboutit fréquemment à des résultats décevants, de l'ordre de 30-40 %, ce qui interroge sur la validité de ces mesures et de la possibilité de tester d'autres mesures acoustiques et de les combiner entre elles. Des outils en libre accès comme VoiceSauce (http://www.phonetics.ucla.edu/voicesauce/) ou plus récemment OpenSMILE (https://www.audeering.com/opensmile/) montrent l'étendue des mesures acoustiques utilisées de plus en plus fréquemment par des phonéticiens. VoiceSauce est un outil conçu pour des phonéticiens et implique des mesures qui peuvent essentiellement être reliées à des critères physiologiques tels *h1-*

h2 qui vise à estimer l'ouverture glottique, ou bien le *Cepstral Peak Prominence* qui est lié à la quantité de bruit dans la voix périodique. Par contre, OpenSMILE est un outil plus fréquemment utilisé dans le traitement du signal et de nombreux paramètres tels que *audSpec_Rfilt_sma* ou *pcm_fftMag_mfcc_sma_de* sont calculés. Ceux-ci relèvent de caractéristiques du spectre ou de bandes spectrales et sont difficiles sinon impossibles à interpréter d'un point de vue phonétique. Mais leur efficacité pour classifier des phénomènes pré-établis linguistiquement (comme la nasalité par exemple) et leur capacité à surclasser des paramètres phonétiques classiques sont forcément tentants. L'utilisation d'un pourcentage de probabilité d'appartenance sert alors de mesure graduelle d'une classe à une autre. Dans cette dernière optique, il serait encore plus tentant d'avoir recours à des Réseaux de Neurones Profonds qui sont connus pour leurs capacités d'extracteurs de paramètres. Je présenterai plusieurs pistes de cet ordre dans la dernière section et qui sous-tendent les dernières recherches vers lesquelles je me suis orienté depuis 2018.

4.4.3 Performance vs. compétence

Cet accès au « big data » (ou mégadonnées) a permis aux linguistes d'envisager les analyses sous un autre angle, en étudiant des données plus naturelles que de la parole lue. Selon Chomsky, ce type de corpus permet d'évaluer la performance plutôt que la compétence d'un locuteur, ce qui n'approcherait qu'une facette des capacités illimitées du locuteur. Nous allons discuter dans cette section les apports et les limites que représentent les grands corpus de parole non contrôlée (préparée ou spontanée) par opposition à des petits corpus de parole lue.

Dans le cadre de petits corpus construits ad-hoc, contrairement à ce qui est obtenu en parole non contrôlée, chaque détail est contrôlé : le nombre de syllabes, la position sérielle du phénomène analysé dans la phrase, l'entourage phonémique, le contexte sémantique, la fréquence lexicale des mots utilisés, l'amorçage, etc. Ces petits corpus permettent d'analyser tout type de réalisation phonétique, mais en respectant généralement le principe du « toutes choses égales par ailleurs ». Il est également fréquent que les origines linguistiques du locuteur et ses habitudes soient mieux connues puisque souvent, l'expérimentateur est en contact direct avec les informateurs. A l'inverse, dans les corpus de parole non contrôlée, les contextes ne sont par définition pas équilibrés. Malgré la taille des corpus analysés, quand bien même elle correspondrait à des centaines d'heures de parole, il est rare que l'ensemble des paramètres mentionnés ci-dessus puisse être observé simultanément en nombre suffisant pour une analyse statistique. Le phonéticien se retrouve ainsi face à l'illusion d'un corpus illimité. La notion de distribution est ici en cause : contrairement au corpus ad-hoc, les contextes ne sont pas en distribution équilibrée (Adda-Decker et al., 2008). Comme mentionné tout au long de ce travail, certains phonèmes et/ou combinaisons de phonèmes sont plus fréquents que d'autres, et il en va de même pour les mots et leurs combinaisons, dont la taille varie également de façon déséquilibrée. Ces distributions inégales vont avoir pour conséquence d'influencer la réalisation moyenne de certains sons.

Par exemple, la présence de la consonne /ʀ/ mentionnée à plusieurs reprises dans ce travail a pour conséquence de postérioriser les sons de son entourage et donc de modifier leur articulation : elle abaisse les fréquences de résonances de F1 et F3, alors qu'elle augmente celles de F2. Or on trouve fréquemment le /R/ suivi d'un /ə/ dans les mots commençant par « re » : « refaire », « regarder », « repas », etc.). En conséquence, le /ə/ est une voyelle que l'on rencontre plus fréquemment que la moyenne dans le contexte /R/ + /ə/. Si l'on mesure la valeur moyenne des formants pour le schwa – y compris au milieu de la voyelle - tous contextes et toutes occurrences confondues, le schwa aura des

valeurs de formants F1 et F3 plus basses que dans la représentation traditionnelle (i.e. au centre du triangle vocalique) et des valeurs de F2 plus élevées. Si l'on prend en compte de façon équilibrée (en nombre d'occurrences) tous les contextes, les valeurs de formants correspondront alors aux valeurs traditionnelles, à savoir centrales par rapport aux autres voyelles (cf. Figure 64).

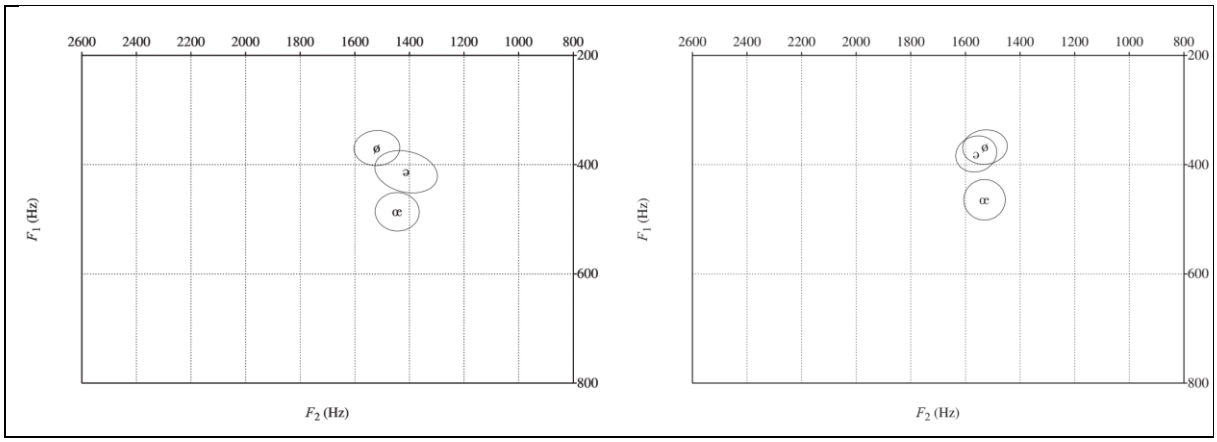


Figure 64 : mesures de formants pour les voyelles centrales du français. Les résultats diffèrent selon que les contextes sont ré-équilibrés (à droite) ou non (à gauche), d'après Adda-Decker et al. (2014)

Doit-on considérer que la représentation sous-jacente du /ə/ est celle d'une voyelle centrale et que sa représentation de surface serait celle d'une voyelle plus postérieure (et plus arrondie) ? Une analyse réalisée sur de grandes quantités de données prises dans des conditions naturelles (i.e. sans équilibrer les contextes) pourrait alors remettre en question un certain nombre de représentations sous-jacentes des phonèmes. Hormis le contexte uvulaire majoritairement présent dans ce cas, les langues favorisent naturellement les contextes alvéolaires qui ont tendance à antérioriser les articulations. Quelle doit donc être la représentation du schwa, comme de tout autre phonème, celle obtenue en équilibrant tous les contextes, ou bien la moyenne de tous les contextes sans pondérer la fréquence des contextes ? Il semble que cette dernière version s'accorde mieux aux théories de la perception dites à exemplaires (Pierrehumbert, 2001) et adaptées et dans les travaux de Bybee (2001). Dans tous les cas, il est vraisemblable - pour reprendre l'exemple du schwa - que la perception d'un schwa dans un contexte où celui est majoritairement plus postérieur et arrondi ne peut qu'influencer les auditeurs dans leur représentation sous-jacente de ce phonème (voir Todd et al., 2019 pour une discussion récente sur ce sujet). Ce phénomène est vérifiable dans les mots rares où la prononciation s'en retrouve hésitante comme par exemple « reblochon » que les locuteurs prononcent alternativement [Rəbloʃɔ̃], [Robloʃɔ̃] ou [Rɔbloʃɔ̃] (bien qu'une harmonisation vocalique ne soit pas à exclure dans ces exemples). Une autre indication de l'influence de la production sur la perception des phonèmes est fournie pas l'évolution des langues (Ohala, 1993) et les résultats des analyses sur les voyelles moyennes /e/ et /ɛ/ (Gendrot et Audibert, 2019) présentés en section 4.3.3 vont également dans ce sens. Quoi qu'il en soit, les questions soulevées par les grands corpus ne font qu'appuyer leur nécessité d'utilisation dans la recherche en phonétique.

Cependant, il sera souvent utile, dans un corpus de parole non contrôlée, de prendre en compte les contextes d'analyse malgré la disparité de leurs fréquences respectives, afin de mesurer l'effet d'une variable toutes choses égales par ailleurs. Il faudra tout de même se méfier des conséquences des tailles d'effectifs qui ont tendance à fournir des résultats significatifs trop 'facilement'. Il s'avère souvent plus utile de considérer les tailles d'effet au-delà des valeurs de p qui pourraient qui plus est

mettre en valeur des coïncidences statistiques. Des statistiques plus complexes que les ANOVAS classiques, tels que les modèles mixtes, les analyses en composantes principales, etc. ont permis de prendre en compte les différences de distributions de contextes dans des proportions inégales, alors que les ANOVAS nécessitent des catégories globalement équilibrées. Les modèles mixtes à effets aléatoires permettent également de prendre en compte la variabilité inhérente à la parole et représentent en cela des tests plus conservateurs.

Une solution intermédiaire consiste finalement à récupérer dans un corpus de parole non préparée des occurrences satisfaisant les conditions d'une analyse sur un petit corpus, c'est-à-dire un entourage segmental équilibré, des contextes sémantiques et pragmatiques cohérents, etc. Cela permet souvent d'obtenir plusieurs centaines d'occurrences qui respectent à la fois le critère naturel de la parole et le contrôle de la variabilité. Les travaux que j'ai effectués avec Audrey Bürki ou encore ceux avec Mathieu Avanzi en sont des illustrations.

Est-ce à dire que les petits corpus de parole lue n'ont pas d'utilité en phonétique ? A mon sens, ceux-ci peuvent être parfaitement complémentaires des grands corpus de parole non contrôlée dans la mesure où ils permettent de vérifier l'effet d'une variable dans un contexte précis, notamment parce que l'ensemble des paramètres susceptibles d'influencer la réalisation des unités de parole ne saurait être entièrement appréhendée dans la parole non contrôlée. Par exemple, les contextes sémantiques et pragmatiques restent encore difficiles à décrire au moyen de variables catégorielles ou même graduelles. Les grands corpus de parole non contrôlée pourront permettre de mesurer l'effet d'une variable dans des conditions naturelles, mais la variabilité devra soit être considérée statistiquement, soit considérée comme partie prenante des résultats. Pour des données pathologiques, des langues peu dotées, des mesures physiologiques, ou des stimuli pouvant être utilisés dans des expériences perceptives, il reste complexe et coûteux de collecter des enregistrements au-delà de plusieurs dizaines de minutes de parole et on pourra avantageusement avoir recours à des corpus de parole lue. Un dernier exemple qui relate l'utilité indéniable des petits corpus lus : on retrouve fréquemment dans les enquêtes sociolinguistiques la liste de lecture de mots qui permet de rendre compte de l'inventaire consonantique et vocalique. Or cette liste ne prend que quelques minutes à enregistrer et quelques heures à traiter, et permettra de générer l'espace vocalique du locuteur à peu de frais et aussi efficacement qu'un grand corpus de parole non préparée. L'espace vocalique ou les caractéristiques consonantiques du locuteur pourront ainsi être utilisés pour un profil vocal individuel, une normalisation, un exemplaire de la langue analysée, etc. Dans tous les cas, on pourra également avoir recours à des méthodes d'automatisation et de normalisation (Gendrot et al., 2010 pour le Tseltal ; (Ridouane et Gendrot, 2017 pour le Mehri). Dans l'article détaillé de Wagner et al. (2015) sur la complémentarité des corpus de parole lue et des corpus de parole, les auteurs comparent des travaux portant sur la réalisation de la proéminence en allemand dans différents types de corpus et pointent les résultats parfois contradictoires ainsi obtenus. Ils montrent que ceux-ci peuvent être dus aux différents styles de parole représentés par ces différents corpus, et insistent sur la nécessité de pouvoir comparer les résultats sur différents types de données afin de généraliser les résultats ou bien au contraire afin de pouvoir déterminer ce qui est spécifique au style de parole ou au contrôle expérimental en question. Ils pointent aussi le côté artificiel de certains protocoles expérimentaux tels que 'maptask' qui tentent de combiner les avantages des corpus contrôlés et non contrôlés. Les auteurs recommandent finalement de s'inspirer des travaux de 'terrain' où les chercheurs ont pour habitude d'obtenir une complémentarité entre parole contrôlée et non contrôlée.

Comme nous avons pu voir en section 4.1.6, mes travaux sur des grands corpus de parole ont permis de proposer une mesure de ralentissement pondéré par le contexte (Gendrot et Adda-Decker, 2016). Dans ce type de modélisation, les dizaines d'heures de parole transcrites et segmentées au niveau du

phonème sont bien utiles. Sur le même principe, les nombreuses mesures de rythme (Dellwo et al., 2015 ; Wagner et Dellwo, 2004) auront avantage à être effectuées sur des corpus conséquents si l'on souhaite prendre en compte la variabilité intra-locuteur et s'assurer de la stabilité des mesures. Les systèmes de synthèse de parole actuels, s'ils ont atteint un niveau de naturel acceptable au niveau du timbre pourraient en bénéficier pour une amélioration de leur prosodie. En effet, la synthèse de parole doit encore être améliorée dans sa variabilité des phrases produites et cette modélisation effectuée sur grands corpus pourrait justement prendre en compte l'éventail des différentes réalisations possibles.

4.4.4 Mesures de réduction phonétique : tout est dit ?

Les travaux plus récents sur les phénomènes de réduction (voir Brandt, 2019 pour une revue exhaustive) ont montré que les proéminences prosodiques des énoncés (frontières de groupes prosodiques et les accents nucléaires et/ou lexicaux) sous-tendaient les phénomènes d'hyper-articulation. Celles-ci sont ensuite combinées à la redondance sémantique observée dans la parole, où des mots peu attendus seraient mieux articulés et donc acoustiquement plus distinctifs (la 'Smooth Signal Redundancy hypothesis' de Aylett et Turk, 2006). Gahl et al. (2012) ont également montré que les locuteurs maximisent l'intelligibilité des mots (ainsi réalisés avec moins de réduction phonétique) qui seraient plus difficiles à identifier de par un voisinage phonologique plus dense. Ces modèles sont cohérents avec les résultats que nous avons observés tout au long de ce travail. Ils sont également plus précis que ceux mentionnant la fréquence comme un des facteurs prédisant la réduction phonétique. En effet, nous avons vu que la fréquence brute observée dans un corpus ne pouvait être prédicteur significatif que si elle était utilisée de manière plus fine. Dans l'exemple de la prédiction du schwa, la fréquence d'une variante lexicale avec schwa pouvait être un bon prédicteur mais pas la fréquence globale du mot. Il semble logique *in fine* qu'on ne puisse pas prendre la fréquence lexicale observée dans un corpus de films (e.g. Lexique3 : http://www.lexique.org/) et l'appliquer de façon globale sur un corpus spécifique. La fréquence doit être considérée comme un paramètre local, au niveau de la conversation in situ. Gahl et al. reconnaissent pour finir que leur modèle ne prend pas en compte les effets des mots à venir, mis à part la probabilité du mot suivant dans le bigramme analysé. Notre étude sur le schwa allait un cran plus loin en analysant la probabilité dans le trigramme. Il y a fort à parier que l'extension des corpus que nous aurons à disposition dans les années à venir permettra d'avancer encore de quelques crans supplémentaires. Cependant, ces modèles ne prennent pas en compte la structure prosodique dans son ensemble telle que nous l'avons envisagée en section 4.2, en considérant les frontières des groupes prosodiques de différents niveaux. Reste que l'analyse que nous avons effectuée porte sur de la parole journalistique (et donc plus proche de l'écrit) et des modèles d'analyse syntaxique de l'oral (Gerdes et al., 2019) devront être améliorés en collaboration avec des phonéticiens pour appliquer ces travaux sur de la parole spontanée. Notons également que dans les travaux présentés sur la modélisation du schwa en section 4.3.1.3 où une analyse approfondie de tous les paramètres disponibles a été réalisée, nous avons pu observer que pour deux questions portant sur la réduction, le raccourcissement et l'élision, les prédicteurs observés n'étaient pas les mêmes. Il convient donc d'être très précis sur le type de réduction observé et les objets analysés pour pouvoir mettre en parallèle différents travaux du domaine.

Un point supplémentaire qui pourra être amélioré dans les recherches futures serait la prise en compte du contexte sémantique par des mesures de plongements de mots ('word embeddings') capables d'estimer la similarité sémantique entre deux mots. Une extension de ces mesures au contexte pragmatique saurait certainement faire sauter le verrou de la modélisation de la variation acoustique

des mots, quoi que le locuteur pourra toujours nous réserver des surprises, comme le soulignait Bolinger dès 1972 (Bolinger, 1972) : « accent is predictable (if you're a Mind-reader) ». Cette citation doit nous ramener de façon systématique au locuteur et à sa variabilité inhérente. Le locuteur peut avoir recours à de multiples stratégies, volontaires ou non, dans sa manière de communiquer l'information, et chaque locuteur peut avoir recours à des stratégies différentes des autres locuteurs. C'est dans cette direction qu'une amélioration des travaux sur la réduction doit être envisagée, et dans cette optique, il sera indispensable de dissocier la variation propre au locuteur et son invariance : la variabilité intra-locuteur et la variabilité inter-locuteurs. Les lacunes et améliorations discutées ici, tant sur les outils d'analyse que sur l'absence de prise en compte de l'invariance du locuteur m'ont amené à faire évoluer la direction de mes recherches depuis 2018. Je présenterai dans la dernière partie de cette section l'avancement de ces travaux en guise de projet de recherche.

4.4.5 Évolutions du Traitement Automatique de la Parole

Dans les travaux présentés tout au long de ce manuscrit, je n'ai pas analysé les différentes tendances entre locuteurs, mais au contraire je me suis concentré sur les tendances globales tous locuteurs confondus. Dans les travaux que je vais présenter dans cette dernière section, effectués en collaboration avec Emmanuel Ferragne et Thomas Pellegrini, je me suis intéressé à ce qui permet de caractériser le locuteur.

Comme précisé dans la section 4.4.2, traditionnellement les phonéticiens ont utilisé l'analyse spectrale afin de décrire des entités linguistiques. Depuis l'avènement de l'apprentissage profond en 2010, des outils puissants sont disponibles pour effectuer ces tâches automatiquement sans nécessairement avoir recours à une analyse spectrale. Un avantage spécifique des réseaux de neurones profonds par rapport à l'intelligence artificielle plus traditionnelle est qu'ils sont capables d'extraire des paramètres d'analyse depuis les données brutes sans qu'un expert humain les fournisse de façon explicite au modèle (Goodfellow et al., 2016). Il semble légitime après 70 ans d'analyse phonétique de se demander si les réseaux de neurones profonds ne pourraient pas remplacer les humains dans le choix des paramètres pertinents pour l'analyse phonétique. Malheureusement, les réseaux de neurones profonds ont la réputation de « boites noires » qui manquent de transparence et n'autorisent pas une interprétation des données. Un des objectifs de nos travaux est de montrer qu'il est possible de surmonter cette difficulté et de tenter de comprendre les caractéristiques sur lesquelles les réseaux ont basé leur décision. En effet, de nombreuses techniques de visualisation[6] ont été mises au point dans le cadre de la reconnaissance d'images et nous tenterons d'en appliquer une sur des spectrogrammes. Un deuxième objectif était de mesurer à quel point une tâche peut être accomplie au moyen de réseaux de neurones profonds, et de la comparer à des mesures acoustiques plus traditionnelles. Ces tentatives se sont faites dans le champ de la comparaison de voix criminalistique, au sein de l'ANR Voxcrim (https://voxcrim.univ-avignon.fr/). Cette thématique était également l'opportunité pour moi de démêler la variation inter-locuteurs et la variation intra-locuteur comme précisé à la fin de la section précédente.

Les réseaux de neurones profonds (DNN désormais pour 'Deep Neural Networks') ont été utilisés avec une grande efficacité pour la modélisation acoustique au niveau du phonème dans le cadre de la reconnaissance de la parole (Hinton et al., 2012) ou la reconnaissance de la langue (Lozano-Diez et al.,

[6] *Voir par exemple le blog suivant qui retrace les avancées récentes : https://thegradient.pub/a-visual-history-of-interpretation-for-image-recognition/*

2018). En ce qui concerne la reconnaissance du locuteur, les DNN ont prouvé leur efficacité de façon indirecte, en étant utilisés non pas comme classifieurs mais comme extracteurs de paramètres (Pellegrini, 2017). Plus particulièrement, les 'Multi-Layer Perceptrons' (MLP) et les 'Convolutional Neural Networks' (CNN) sont capables, au niveau de certains neurones, de se spécialiser sur des traits phonétiques comme le lieu ou le mode d'articulation, ou bien sur des mesures acoustiques telles que les formants (Weber et al., 2016). L'objectif de ce travail n'était pas de s'inscrire totalement dans le domaine de la reconnaissance du locuteur, mais plutôt dans le cadre de la comparaison de voix : un petit extrait de données orales issues de quelques dizaines de locuteurs et contenant exclusivement la voyelle nasale /ɑ̃/. Bien sûr, il pourra toujours nous être rétorqué que les réseaux de neurones utilisés se seront sur-spécialisés sur un petit ensemble de données, mais notre objectif était d'identifier ce qui permettait de discriminer un nombre limité de locuteurs, et pas de créer un classifieur générique pour un nombre illimité de locuteurs. De façon plus générale, l'utilisation de l'apprentissage machine en se concentrant sur un phénomène phonétique précis permet de mieux comprendre la façon dont les données sont analysées. L'apprentissage machine permet également de connaitre ici le score maximal de classification que l'on peut atteindre sur la base d'un spectrogramme, en prenant non plus une analyse phonétique classique, mais une analyse du spectre effectué globalement sans connaissances à priori.

4.4.5.1 Interprétabilité par visualisation

L'objectif de ce premier travail était de déterminer les zones fréquentielles utiles aux CNN pour classifier des locuteurs. Des spectrogrammes ont donc été insérés en entrée de nos réseaux pour pouvoir les analyser en sortie. Dans une première expérimentation, 45 locuteurs ont été classifiés par un CNN sur leurs productions de voyelles /ɑ̃/. Dans un deuxième temps, le même CNN, incluant les mêmes paramétrages, a été ré-entrainé et testé sur les mêmes voyelles, mais auxquelles des filtres passe-bas de différentes valeurs ont été appliqués. Dans la troisième expérience, un algorithme de sensibilité à l'occlusion a été testé, et nous avons pu observer comment le masquage de certaines bandes de fréquences affectait les taux de classification.

La voyelle /ɑ̃/ a été utilisée car celle-ci a montré les taux de classification de locuteurs les plus élevés en comparaison avec d'autres voyelles (Gendrot et al., 2019), ce qui est cohérent avec de précédents résultats de la littérature (Ajili et al., 2016 ; Kahn et al., 2011). Les nasales, et plus particulièrement les voyelles nasales, semblent contenir plus d'informations propres aux locuteurs comparativement aux autres segments. L'origine de cette spécificité tiendrait au fait que la cavité nasale est moins malléable que la cavité buccale et la cavité pharyngale. Notons que les voyelles nasales sont généralement plus longues que leurs contreparties orales (de l'ordre de 40 ms en moyenne), ce qui leur permet de fournir plus d'informations. Les voyelles utilisées pour la classification ont été extraites du corpus ESTER, en utilisant les alignements phonétiques utilisés précédemment. Les voyelles ont été extraites à l'aide d'une fenêtre rectangulaire et sans leur contexte phonétique, celui-ci n'étant ni contrôlé ni fourni au réseau pendant l'entrainement ou le test.

Les spectrogrammes ont été obtenus en utilisant les paramètres qui correspondent au mieux aux spectrogrammes auxquels les phonéticiens sont habitués, toujours dans un but d'interprétabilité et plus de détails sur ces paramètres peuvent être consultés dans Ferragne et al. (2019). Les voyelles sélectionnées dans les corpus avaient une durée comprise entre 30 ms et 250 ms. Les voyelles dont la durée était inférieure à 250 ms ont été complétées par des valeurs nulles (zero padding) afin que toutes les images aient une largeur égale, et qu'il n'y ait pas de distorsion visuelle. Les images ont ensuite été converties en niveau de gris sur 8 bits et redimensionnées en 224 x 224 pixels, ce qui correspond pour chaque pixel à 1.15 ms sur la dimension temporelle, et 35.71 Hz sur la dimension

fréquentielle. Le modèle utilisé pour la classification est le VGG16 (Simonyan et Zisserman, 2014), un réseau de neurones convolutif reconnu pour la reconnaissance d'images, y compris dans le domaine de l'audio (Nagrani et al., 2017). Dans nos expériences, le modèle a été ré-entrainé intégralement avec des poids attribués aléatoirement.

Le plus petit nombre de /ɑ̃/ produit par un locuteur était 334 ; 334 occurrences ont été choisies aléatoirement pour chacun des autres 44 locuteurs, pour un total de 15030 spectrogrammes. Cet ensemble a été découpé selon un groupe d'entrainement de 70 %, un groupe de validation de 10 % et un groupe de test de 20 %. Une première classification a été effectuée à partir des spectrogrammes entiers avec une précision moyenne de 85.37 % pour des moyennes individuelles allant de 59.70 % à 98.51 %. Le score moyen pour les 10 femmes était de 83.88 % et de 85.80 % pour les 35 hommes, différence non significative selon le test U de Mann-Whitney (p=0.32). Au-delà de ces scores encourageants, notre objectif était de comprendre ce qui avait été utilisé par le modèle dans sa classification. Les spectrogrammes ont été progressivement coupés par des filtres passe-bas afin d'en obtenir différentes versions allant de 250 Hz à 6000 Hz ; chacune de ces différentes versions a ensuite été ré-entrainée et testée selon les mêmes paramètres que ceux de la version non altérée et les résultats de chaque version sont présentés dans la Figure 65 (le résultat pour la version non altérée y a également été laissé à titre de comparaison). Le score le plus bas (50.85 %) obtenu avec un filtre passe-bas de 250 Hz demeure significativement au-dessus du niveau de chance ($\chi 2 = 38479$, $p<0.001$). Ce résultat montre que l'on peut discriminer des locuteurs à un niveau largement supérieur à la chance avec très peu d'information spectrale. On peut observer ensuite que le taux de précision augmente rapidement (et significativement) entre 250 Hz et 1000 Hz, puis plus lentement entre 1000 Hz et 8000 Hz. Ces résultats montrent qu'une grande partie de l'information caractérisant le locuteur se trouve en dessous de 1000 Hz. Notons également une 2ème zone avec une augmentation plus rapide, significative, entre 3000 Hz et 4000 Hz, qui pourrait être mise en parallèle avec le 4ème formant fréquemment mentionné dans la littérature comme plus caractéristique du locuteur que les trois premiers qui seraient caractéristiques de l'articulation.

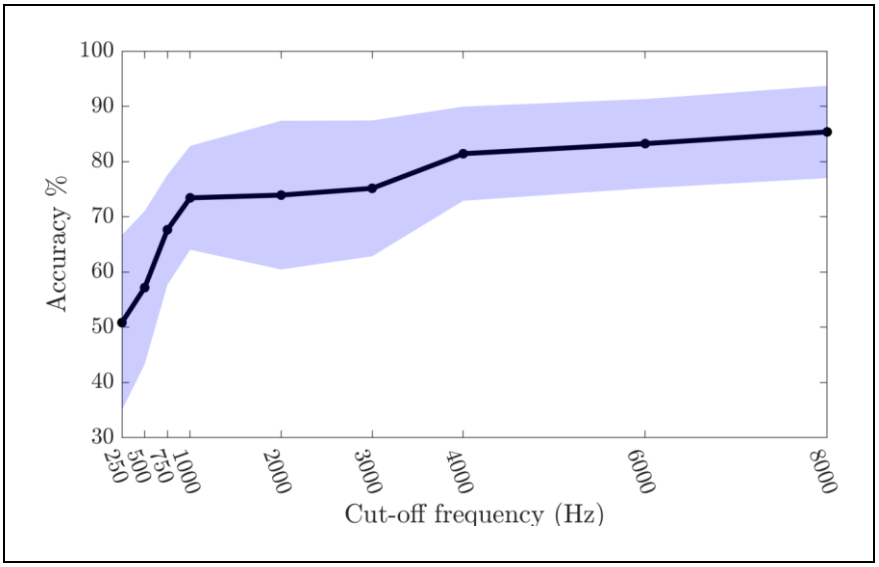


Figure 65 : précision du CNN en fonction de la fréquence du filtre passe-bas (la bande bleue représente l'écart-type), d'après Ferragne et al. (2019)

Un coefficient de concordance de Kendall nous a permis de conclure que ce résultat ne peut pas être observé pour chaque locuteur individuellement (W=0.11 ; $\chi 2$ = 43.39 ; p = 0.5) ce qui nous a poussé à approfondir nos analyses afin de localiser plus précisément les zones saillantes propres aux locuteurs. Nous avons donc réalisé une opération dite de sensibilité à l'occlusion : lors de la phase de test, un masque (réalisé au moyen d'une bande de zéros de 15 pixels, soit 536 Hz) sur toute la largeur de la voyelle a été positionné sur le spectrogramme. Cette opération a été effectuée en remontant l'ensemble des fréquences, pixel par pixel, générant systématiquement une nouvelle image de test. La position initiale ne masquait que la rangée de pixels la plus basse du spectrogramme, et la position finale la rangée la plus haute, avec un total de 252 nouveaux spectrogrammes pour chacune des 3015 voyelles (67 voyelles de test x 45 locuteurs). Pour finir, les 3015 voyelles de test ont été converties en cartes thermiques (désormais 'heatmaps') où la couleur d'une bande de fréquence reflète la probabilité que la voyelle corresponde à sa classe de locuteur quand sa bande de fréquence a été masquée[7]. Afin de déterminer quelles bandes de fréquences étaient typiques d'un locuteur, la moyenne et l'écart-type des 67 heatmaps de chaque locuteur ont été calculées et un Rapport de Signal sur Bruit a été obtenu en calculant le ratio par pixel entre la heatmap moyenne et l'écart-type de la heatmap de chaque locuteur. Une normalisation sur l'ensemble a finalement été effectuée pour faciliter la comparaison.

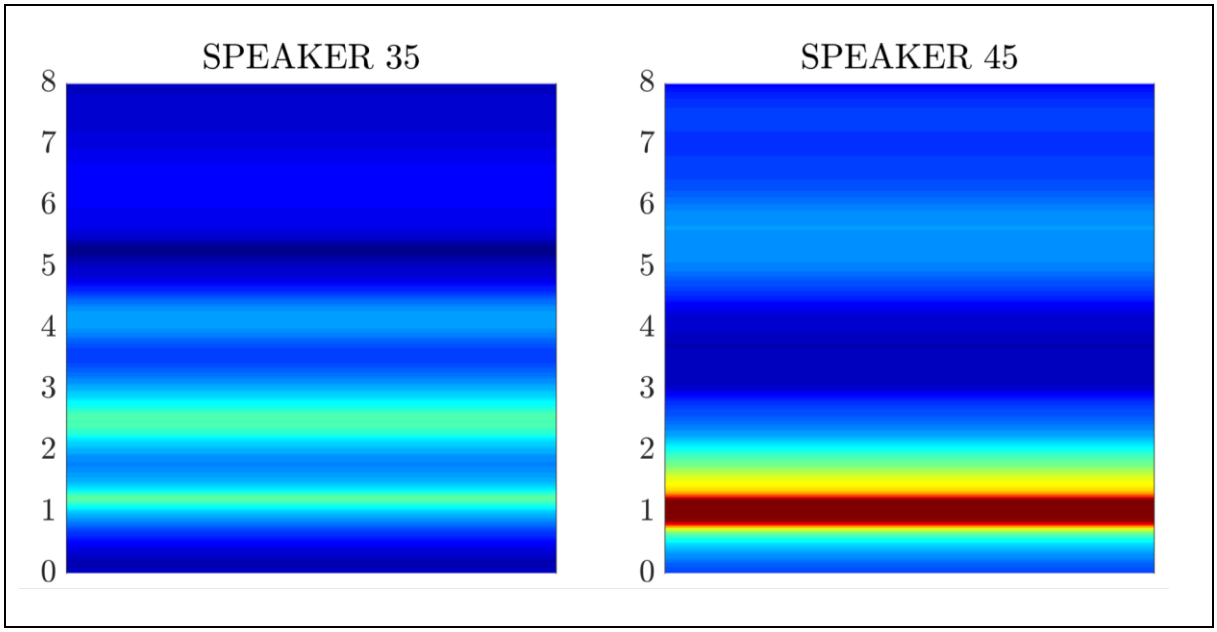


Figure 66 : deux heatmaps Signal sur Bruit (locuteur 35 et locuteur 45, sexe masculin), le bleu représente une valeur faible et le rouge une valeur élevée, d'après Ferragne et al. (2019)

La Figure 66 ci-dessus montre deux heatmaps avec les valeurs de Signal sur Bruit. Celle de gauche contient les valeurs les plus basses observées dans nos données (pour le locuteur 35), alors que la figure de droite correspond aux valeurs les plus élevées. La bande rouge autour de 1000 Hz montre que si cette région est masquée dans les spectrogrammes du locuteur 45, la probabilité que ce locuteur soit correctement classifié baisse de façon importante et cette détérioration est constante sur

[7] Pour des raisons de coût de traitement dans un premier temps, la même bande de masquage était appliquée sur toute la voyelle. Il serait intéressant par la suite de les moduler dans le temps, par tranches de 20 ms par exemple.

l'ensemble des 67 spectrogrammes tests de ce locuteur. A contrario pour le locuteur 35, bien que des régions critiques émergent (les bandes vert-jaunes légèrement au-dessus de 1000 Hz et entre 2000 Hz et 3000 Hz), celles-ci ne sont pas aussi pertinentes. Bien que ces deux locuteurs aient des heatmaps qui sortent du lot, en cela qu'elles contiennent les valeurs maximales et minimales de l'ensemble de données, leurs scores de classification ne se distinguent pas des autres et se trouvent entre le 1er et le 3ème quartile. Leurs courbes de précision de la 2ème expérience montrent des différences : le locuteur 35 révèle une augmentation abrupte de 42 % (250 Hz) à 94 % (3000 Hz) alors que le locuteur 45 voit sa précision augmenter plus lentement, de 64 % à 88 % sur le même intervalle. La relation entre la précision de la classification et les heatmaps n'est donc pas très claire (la corrélation est significative à $p<0.001$ mais peu élevée à 0.187) : deux autres heatmaps (locuteurs 17 et 18) visuellement très similaires n'ont pas des taux confusions très élevés contrairement à ce qui serait attendu.

Une inspection visuelle des 45 heatmaps a permis de montrer que pour au moins 35 locuteurs, on observe une bande de fréquence intense autour de 1000 Hz (comme le montre la Figure 66), ce qui montre l'importance de cette région spectrale pour caractériser la voix d'un locuteur. Cependant, cette bande de fréquence ne donne pas d'indication sur les modulations de fréquences pertinentes dans cette zone, et ce point reste à analyser. Ces résultats devront également être approfondis en vérifiant si d'un point de vue perceptif, la confusion entre deux locuteurs se rapproche des résultats de classification obtenus par le CNN ou bien se rapproche des heatmaps de saillance spectrale obtenues.

4.4.5.2 Comparaison des CNN avec des mesures acoustiques

Dans cette deuxième approche (Gendrot et al., 2019), nous nous sommes intéressés aux invariants phonétiques utiles pour la caractérisation du locuteur, qu'ils soient utiles aux phonéticiens ou aux systèmes de reconnaissance automatique du locuteur. En reprenant le principe de la comparaison de voix plus que de la véritable reconnaissance du locuteur, et en utilisant des réseaux de neurones convolutifs sur des voyelles extraites de leur contexte, nous avons souhaité fournir en entrée des mesures acoustiques interprétables, et les comparer aux résultats obtenus dans la section précédente. Nous avons choisi d'utiliser le même CNN que dans la section précédente, avec des paramétrages identiques.

Les 45 locuteurs analysés dans la section ont été repris, mais cette fois en plus de la voyelle /ɑ̃/, nous avons reproduit tour à tour nos analyses pour six autres voyelles : /a/, /ɛ/, /e/, /i/, /ə/ et /ɔ/. Cette étude s'est réalisée en deux étapes : (1) pour chacune des voyelles, le CNN a dû classifier les 45 locuteurs d'après le spectrogramme fourni en entrée (modèle SPECTR) ; (2) le même CNN a dû réaliser la même classification, mais à partir de mesures acoustiques effectuées au préalable (modèle ACOUS). Notre hypothèse était que le modèle SPECTR obtienait de meilleurs résultats en général, mais la comparaison des deux et leur éventuelle complémentarité pouvait fournir des éléments utiles à notre compréhension de la classification ainsi effectuée.

Comme dans la précédente section, toutes les voyelles des 45 locuteurs ont été extraites du corpus ESTER puis converties en spectrogrammes selon les mêmes critères. Sur ces voyelles, des mesures acoustiques ont été effectuées avec un pas d'une milliseconde à l'aide de PRAAT et de VoiceSauce. Pour ce dernier, l'ensemble des mesures possibles a été utilisée, soit :

f0	Cepstral Peak Prominence (CPP)
valeurs F1-F4	Harmonic to Noise Ratio : HNR15, HNR25, HNR35 (avec une fenêtre de 15, 25 et 35 ms)
largeurs de bande F1-F4	subharmonic to harmonic ratio (SHR)
Intensité (dB et RMS)	strength of excitation (SOE)

Notons que les mesures relatives à la f0 et aux formants sont mesurées à la fois avec PRAAT (p) et par SNACK (s). Ces paramètres ci-dessus ne sont pas corrigés en fonction des valeurs de formants. Pour les autres paramètres, on trouve une version corrigée ('c') et non corrigée ('u') (voir Kreiman et al., 2017 et Shue et al., 2011 pour plus de détails) :

H1 (amplitude du 1er harmonique)	2K (amplitude des harmoniques à 2000 Hz)	H1-A3
H2	5K	H4-2K
H4	H1-H2	2K-5K
A1 (amplitude de F1)	H2-H4	
A2	H1-A1	
A3	H1-A2	

Ces 62 paramètres sont considérés comme de bons descripteurs de la qualité de voix, ce qui est indéniablement une composante à prendre en compte dans la comparaison de voix (Nolan, 2007). Des mesures de moments spectraux ont été extraites à l'aide de PRAAT : centre de gravité spectral (COG), kurtosis (aplatissement), skewness (asymétrie), et écart-type de la distribution de l'énergie sur le spectre (SD). Les valeurs pour l'intégralité des paramètres ont été normalisées sur une échelle 0-255 pour correspondre à la conversion 8 bits des spectrogrammes, et ont ensuite été redimensionnées en une matrice 224 x 224 pour se conformer à la taille d'entrée des images de notre réseau. Pour l'analyse phonétique effectuée à la suite des résultats obtenus avec le CNN, des mesures ont été extraites à 25, 50 et 75 % de la durée de chaque voyelle puis moyennées sur une seule valeur pour plus de lisibilité et de facilité de traitement dans un premier temps. Nous sommes bien conscients que l'intégralité des points de mesures sur toute la durée du segment serait plus pertinente pour évaluer la variation spectrale à l'intérieur de la voyelle, et cette approche devra être traitée dans un travail futur. L'entrainement a été à nouveau effectué avec des poids attribués aléatoirement. Au total, 14 modèles ont été lancés, un pour SPECTR et un pour PHONET, et ce pour chacune des sept voyelles analysées.

Les résultats présentés dans le Tableau 27 montrent que les 'features' apprises par le spectrogramme (méthode SPECTR) aboutissent à des scores en moyenne 10 à 15 points au-dessus des 'features' acoustiques. Les scores supérieurs à 69 % sont surlignés en gras, et le tableau est divisé en cinq catégories pour la suite de notre analyse :

• catégorie 1 : PHONET a correctement identifié le locuteur (parmi les 45 locuteurs possibles)

• catégorie 2 : SPECTR a correctement identifié le locuteur

• catégorie 3 : PHONET correct. ; SPECTR incorrect.

• catégorie 4 : PHONET incorrect. ; SPECTR correct.

• catégorie 5 : PHONET incorrect. ; SPECTR incorrect.

cat.	ɑ̃	a	ɛ	e	i	ə	ɔ
cat. 1	**71.2**	**69.3**	63.6	64.8	53.1	57.9	63.2
cat. 2	**86.7**	**77.1**	**75.4**	**76.0**	**69.8**	**71.5**	**74.5**
cat. 3	5.8	10.5	10.3	9.7	10.5	10.1	11.6
cat. 4	21.3	18.3	22.1	20.9	27.3	21.4	25.2
cat. 5	7.4	12.4	14.3	14.3	19.7	25.2	16.9

Tableau 27 : taux de classification (en %) pour chaque voyelle en fonction des différentes catégories, d'après Gendrot et al. (2019)

Les deux réseaux montrent des résultats de classification identiques pour 68 % des voyelles testées (correctement pour 53.5 % des cas, ou incorrectement pour 14.5 %). 22 % des voyelles ont été correctement classifiées par SPECTR, alors que PHONET s'est trompé, et l'inverse a pu être observé pour 10 % des voyelles analysées. Pour les deux méthodes, /ɑ̃/ est la voyelle qui obtient les scores les plus élevés. Pour finir, les résultats sont meilleurs pour les hommes (67.6 %) que pour les femmes (62 %) quand PHONET est utilisé, alors qu'ils sont sensiblement équivalents avec SPECTR (74.5 % vs. 76.4 %). Sans que ce soit surprenant, les erreurs de discrimination sont plus fréquentes sont plus fréquentes au sein du même sexe, mais de façon plus prégnante pour les hommes (92 % vs. 70 % en moyenne pour les deux modèles).

La significativité des mesures acoustiques pour déterminer leur pertinence a été évaluée au moyen d'une MANOVA calculée pour chaque voyelle avec toutes les mesures acoustiques comme variables dépendantes et l'identité du locuteur comme variable indépendante. Une analyse linéaire discriminante a permis d'obtenir le poids de chaque variable ainsi que le degré de colinéarité aves les autres paramètres (certaines mesures obtenues par VoiceSauce sont identiques mais obtenues au moyen d'algorithmes différents). Les paramètres montrant une forte colinéarité ont été comparés entre eux, et celui avec le poids le plus fort a été conservé. Au final, les paramètres les plus pertinents pour la classification en locuteurs sont HNR35, l'énergie, H1u, H1H2u, H1H2c, H1A1u, skewness, COG, pf0, H2KH5Kc, A2u et A3u. Après avoir calculé la moyenne par locuteur de chaque paramètre pour chaque voyelle, nous avons calculé la différence entre chaque voyelle et la moyenne du locuteur et ainsi analysé l'écart à la moyenne. Si cette valeur est très éloignée de la moyenne, elle est alors considérée comme peu représentative de ce locuteur, ce qui peut apparaître comme un motif de mauvaise classification.

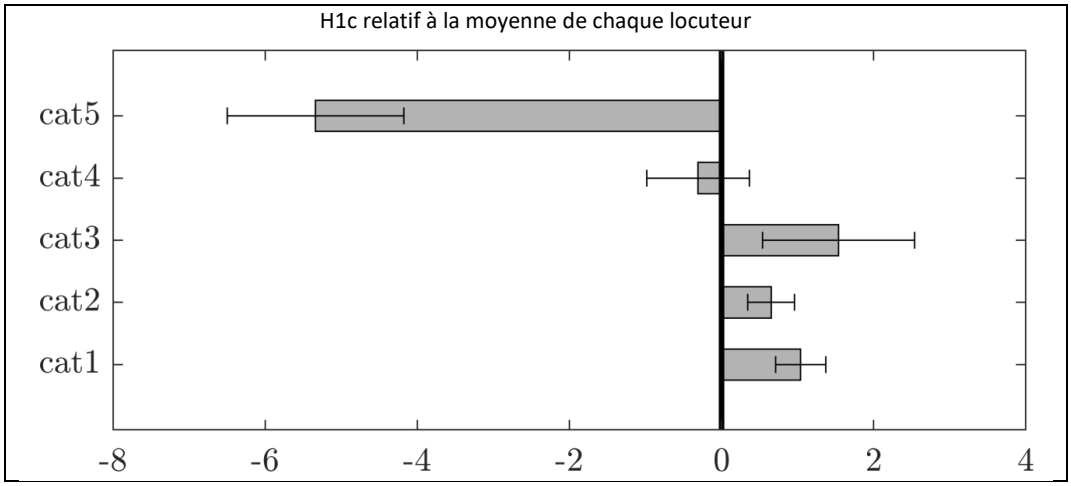


Figure 67 : valeurs moyennes de H1c pour toutes les voyelles en fonction des scores de classification, d'après Gendrot et al. (2019)

La Figure 67 montre les valeurs du paramètre H1c par rapport à la moyenne de chaque locuteur. On y voit de bas en haut les occurrences où la classification a été correcte pour PHONET (cat.1) et SPECTR (cat.2) comparées aux occurrences où la classification était incorrecte pour SPECTR (cat.3), PHONET (cat.4), puis SPECTR et PHONET (cat.5). Lorsque la classification était erronée pour SPECTR et PHONET, on remarque que les valeurs de H1c sont très en dessous de la valeur moyenne du locuteur. A l'inverse, lorsque la classification est incorrecte pour SPECTR (cat. 3), les valeurs de H1c sont au-dessus de la moyenne du locuteur. Les résultats fournis par la catégorie 3 montrent que la f0, l'énergie, H1, HNR et skewness sont très éloignés de la moyenne du locuteur alors que la classification est incorrecte, ce qui suggère qu'ils ne sont pas utilisés par le modèle SPECTR. Des travaux fusionnant l'utilisation de paramètres acoustiques et du spectrogramme dans un CNN devront être testés par la suite. Cette étude montre in fine que la caractérisation des locuteurs par des mesures phonétiques reste un exercice délicat, il est difficile d'établir une liste de paramètres qui puisse être mise à plat pour classifier de façon précise un locuteur. Nous avions en tête que la variation des paramètres au cours de la production de parole pouvait également être un facteur à prendre en compte, et nous avons donc abordé cet aspect dans la section suivante.

4.4.5.3 Du phonème à la phrase

Dans la suite de ces travaux, nous nous sommes intéressés à des séquences plus longues que de simples voyelles, en prenant en compte des unités de quatre secondes de parole. Nous cherchions ainsi à intégrer le caractère dynamique qui manquait dans les travaux précédents. Cette fois, le corpus utilisé était NCCFr car nous souhaitions des conditions d'enregistrements plus contrôlées (l'enregistrement du corpus a été effectué dans les locaux du Laboratoire de Phonétique et Phonologie (Torreira et al., 2015). Le corpus est constitué de 45 locuteurs puisque la segmentation d'un des locuteurs était incomplète, et nous nous sommes servis de la segmentation automatique pour récupérer 350 séquences de quatre secondes par locuteur contenant un minimum de 20 phonèmes. L'objectif de ce travail était de quantifier à quel point il était possible d'identifier un locuteur en utilisant seulement les paramètres prosodiques de f0 et d'intensité, et en les comparant aux résultats obtenus avec un spectrogramme en guise de référence. En ne prenant que quatre secondes de parole (là où les systèmes de reconnaissance du locuteur utilisent souvent 30 secondes au minimum), le contenu était volontairement restreint afin de pouvoir analyser ses spécificités ensuite. Ce travail avait pour objectif de se rapprocher des études sur la caractérisation du locuteur par ses aspects rythmiques (Dellwo et al., 2015), mais cette fois sur des contenus non lus (et donc non identiques). Nous avions également pour objectif de pouvoir prendre en compte les variations de f0 et d'intensité dans leur ensemble et non plus seulement des mesures moyennes, et le CNN était utilisé encore une fois pour ses capacités d'extracteur de paramètres.

À partir de ces extraits, nous avons obtenu quatre types de données : le spectrogramme complet de chaque séquence ; le contour de f0 ; les valeurs d'intensité ; la représentation conjointe de f0 et d'intensité. Le spectrogramme servira ici de référence puisqu'il comprend l'information la plus complète sur la séquence analysée. Les détails sur la procédure d'extraction des paramètres acoustiques et sur le modèle de CNN peuvent être consultés dans Chignoli et al. (2020). Les premiers résultats peuvent être présentés comme suit : les courbes de f0 permettent de classifier correctement les locuteurs à 28 %, les courbes d'intensité à 32 %, et lorsque l'on cumule les deux, les scores de bonnes classifications grimpent à 59 %. Les scores de classifications des spectrogrammes sont quant à eux de 93 %. Que faut-il retenir de ces résultats ? Bien sûr, une certaine variabilité inter-occurrences peut être observée : certaines séquences contiennent plus de phonèmes que d'autres (toujours pour

une durée de quatre secondes) et fournissent plus d'informations qui permettent d'identifier le locuteur. Certaines séquences ont des caractéristiques très inhabituelles et sont mal identifiées. Deux locuteurs ont des caractéristiques prosodiques très discriminantes qui ont permis d'atteindre des scores de classification légèrement plus élevées que les spectrogrammes. Pour finir sur l'utilité des paramètres prosodiques de f0 et d'intensité, pour le sous-groupe de données que le réseau n'arrive pas à classifier correctement en se basant sur le spectrogramme (7 % des données totales), l'utilisation conjointe du contour de f0 et d'intensité parvient à une bonne classification dans 33 % de ces occurrences. Des mesures acoustiques sur les séquences nous ont permis de montrer – au moyen d'une analyse en composantes principales – que les valeurs minimales et le dernier décile sont les mesures plus pertinentes pour l'intensité, et la valeur moyenne et le décile médian sont les plus représentatives du locuteur pour la f0. Ces travaux sont actuellement poursuivis pour appliquer ce principe à de multiples paramètres et ainsi obtenir des regroupements de locuteurs ; en utilisant l'intégralité des mesures fournies par VoiceSauce, on parvient à près de 90 % de bonnes classifications (Chignoli et Gendrot, soumis). Ces travaux montrent à quel point il est crucial de prendre en compte l'aspect dynamique quand des mesures phonétiques sont utilisées (à terme, il sera intéressant de tester perceptivement certaines séquences en demandant aux auditeurs de former des groupes de locuteurs « proches » en termes de voix. Cela nous permettra d'appréhender les paramètres sur lesquels se basent les humains pour identifier une voix.

Dans cette quête de compréhension des paramètres utiles à la caractérisation du locuteur, nous avons cherché à vérifier pourquoi certaines séquences étaient mieux reconnues que d'autres, en reprenant les spectrogrammes fournis en entrée dans le réseau (Gendrot et al., 2020). Dans un premier temps, nous avons retenu une mesure de flux spectral ('spectral flux') qui permet de quantifier la variabilité du spectre d'une fenêtre à la suivante, et nous avons observé que les séquences dont les valeurs moyennes de flux spectral étaient plus importantes avaient un taux de bonnes classifications significativement plus élevé. Malheureusement, en plus des phénomènes de coarticulation qui avaient une valeur de flux spectral élevée, on trouvait également les phonèmes suivis et précédés d'une pause (de par un changement spectral important), ce qui impliquait qu'un plus grand nombre de pauses augmentait la valeur de flux spectral de la séquence sans qu'elle soit corrélée à une meilleure classification. Pour finir, le flux spectral est fréquemment corrélé au débit de parole, ce qui rend ce paramètre délicat à utiliser dans un contexte non contrôlé. Nous avons poursuivi nos travaux sur ces séquences de quatre secondes en cherchant à savoir si certains phonèmes ou classes de phonèmes avaient été décisifs dans la classification du locuteur. Cette question est légitime puisqu'il est fréquent de constater que des locuteurs de notre entourage présentent une production particulière de certains phonèmes : on pense par exemple à la réalisation de /ʃ-s/ et /ʒ-z/ en fricatives latérales [ɬ] et [ɮ]. Sur le même principe que Ajili et al. (2018), nous avons réalisé une expérience de classification avec masquage ('occlusion'), où une partie de l'information contenue dans le signal acoustique est cachée afin de comparer la classification avant et après masquage. Cette étude se rapproche également des travaux effectués par Besacier et Bonastre (1998) dans lesquels des blocs temporels de signal sont sélectionnés pour améliorer les taux d'identification du locuteur. Cependant, contrairement aux études citées ci-dessus, nous avons utilisé des spectrogrammes à bandes larges en entrée avec des réseaux neuronaux, toujours dans notre objectif de phonéticien de retracer la correspondance acoustique-articulatoire. De plus amples détails sur les méthodes d'occlusion et sur les CNN utilisés peuvent être trouvés dans Gendrot et al. (2020). Pour rendre la tâche plus difficile aux CNN et ainsi obtenir plus de cas où la classification bascule, nous sommes passés à des séquences de deux secondes, au lieu de quatre. Dans un premier temps, nous n'avons masqué qu'un seul phonème à la fois en parcourant l'ensemble de la séquence. L'objectif était d'identifier un ou plusieurs phonèmes

susceptibles de faire basculer l'identification du locuteur (i.e. engendrer une classification erronée). Ce type de changement dans la classification n'a été obtenu que pour les séquences dont la probabilité de classification dans la classe correcte avant masquage était faible, inférieure à 50 %. Les taux d'identification étant globalement supérieurs à 90 % avec des probabilités d'identification très élevées, le masquage d'un seul phonème ne permettait que très rarement d'engendrer une erreur de classification. Au total, à l'issue du masquage par phonème, seuls 2.5 % des séquences présentaient un changement de classe de locuteur, ce qui est insuffisant pour effectuer une analyse quantitative. Notons tout de même que les phonèmes /s/ et /ʃ/ ont été identifiés pour deux locuteurs comme particulièrement pertinents, notamment parce que -après écoute des séquences concernées- ceux-ci étaient réalisés avec un chuintement. Mais ces cas étaient par trop rares et nous avons donc procédé dans un second temps à une occlusion par classes phonémiques, où tous les phones correspondant à une classe phonémique ont été masqués simultanément. En analysant les scores d'identification après masquage, nous avons pu déduire que les voyelles orales ont un effet important sur la classification, loin devant les occlusives puis les autres catégories (respectivement nasales, sonantes, fricatives), mais ces résultats sont en partie corrélés à la fréquence d'apparition de ces classes, et donc à la durée du masque sur la séquence. Il est à noter que lorsqu'une classe phonémique permet de faire basculer la classification du locuteur de correcte à erronée, il est très fréquent que les autres classes phonémiques testées fassent également basculer la classification (25 % de cas où une seule classe phonémique est impliquée dans un changement de catégorie pour une séquence, 24 % de cas où il y a 2 classes, 51 % de cas où il y a entre 3 et 5 classes), ce qui ne plaide pas en faveur de l'idée d'une classe phonémique cruciale pour la classification du locuteur.

Un approfondissement de ces analyses sur la variation inter-locuteurs nous a permis de montrer que 20 % des locuteurs ne sont pas sensibles au masquage, ces locuteurs attirant à eux les prédictions dont le score de probabilité est plus faible. Ces locuteurs que nous avons qualifiés d'attracteurs pourraient être considérés comme les agneaux ('lambs') selon la terminologie de Doddington et al. (1998) car ces locuteurs pourraient apparaitre comme faciles à imiter. Afin de comprendre pourquoi ces locuteurs recueillent un nombre important de faux positifs, nous avons effectué des mesures acoustiques sur les différentes séquences testées de ces locuteurs et avons pu constater qu'ils étaient caractérisés par une variation acoustique plus importante que les autres locuteurs, notamment pour leurs valeurs de f0 et d'intensité. Nous avons également pu faire ressortir des locuteurs qui se distinguent par leur caractère moyen sur l'ensemble des mesures acoustiques.

L'ensemble des travaux en cours mentionnés ici relèvent surtout la complémentarité qui peut exister entre les mesures acoustiques et les spectrogrammes (présentés comme des images), et la nécessité de comprendre les paramètres que les réseaux neuronaux parviennent à extraire. Ils montrent aussi toute la complexité à caractériser un locuteur par des mesures phonétiques interprétables. Dans les travaux que nous menons actuellement (Chanclu et al., soumis), plutôt que de procéder à une classification directe des locuteurs, nous essayons de caractériser des types de voix, en prenant en compte la qualité de voix. Bien que celle-ci soit complexe à décrire, elle reste une des caractéristiques propres à un locuteur. Par exemple, la nasalité de la voix d'un locuteur qui peut être due à une mauvaise fermeture du voile du palais, ou une voix soufflée générée par une fermeture incomplète des plis vocaux pendant la phonation ('glottal chink') sont des caractéristiques du locuteur qu'un humain peut identifier et sur lesquelles il peut s'appuyer pour reconnaitre perceptivement une voix. Ces caractéristiques restent complexes à analyser par des paramètres phonétiques, puisque basées sur l'amplitude des harmoniques (Klatt et Klatt, 1990 ; Pruthi et Espy-Wilson, 2004). Nous avons dans un premier temps effectué des mesures acoustiques intégrant des mesures fournies par openSMILE qui vont au-delà des mesures phonétiques classiques, bien qu'elles soient difficilement interprétables d'un point de vue articulatoire. Ces mesures ont ensuite été couplées à un classifieur basé sur des CNN.

Ces résultats -préliminaires- permettent de classer une voix comme orale/nasale ou craquée/modale/soufflée (voir Figure 68) et devraient pouvoir être intégrés dans une caractérisation du locuteur interprétable par l'humain (voir Kreiman et al., 2021 pour une discussion récente de l'importance de la validation par l'humain de la qualité de voix).

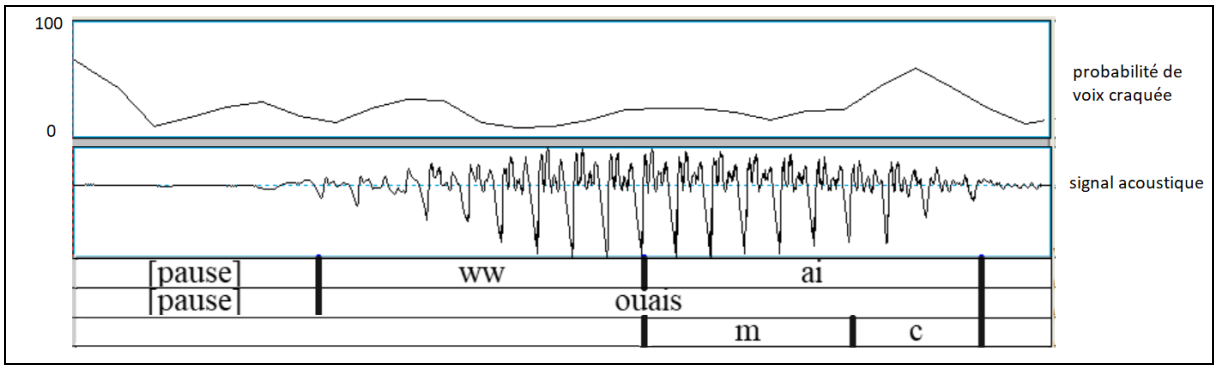


Figure 68 : calcul de la probabilité de voix craquée dans le signal. La fin du mot 'ouais' est réalisée craquée par le locuteur

A quelques lignes de la fin de ce document, il me semble utile de rappeler que l'ensemble de ces travaux n'a réellement de sens que rapporté à l'humain. C'est particulièrement vrai pour les derniers travaux que j'ai présentés ici, ceux-ci ayant pour but l'explicabilité de l'identification des individus par leur voix dans un cadre criminalistique. C'est également vrai lorsque l'on mesure des variations de durée, de formants, ou tout autre paramètre acoustique : il est nécessaire de se demander si l'humain est capable de les utiliser consciemment ou inconsciemment dans le cadre d'une interaction. A l'heure où des systèmes de classification basés sur des réseaux neuronaux peuvent surclasser (très largement dans certains cas) les capacités humaines, il n'est pas inutile de rappeler que l'injection de « l'humain » est le plus souvent sous-jacente à ces résultats, qu'elle soit sous la forme d'une annotation permettant un entrainement ciblé, d'une aide à la décision dans un système expert, d'une validation des résultats, ou de modifications à apporter pour s'adapter à un ensemble de données constamment en évolution.

5 Conclusion générale

Les travaux présentés dans ce document ont montré qu'en 15 ans de recherches en phonétique, les données, méthodes, et modélisations ont évolué rapidement pour une meilleure compréhension de la variation en parole. J'ai montré qu'il était possible de réaliser automatiquement des mesures acoustiques sur des grands corpus annotés automatiquement. Ces mesures ont permis de mettre en évidence des phénomènes de variations segmentales dans la parole et ce pour plus plusieurs styles, avec des spécificités propres à chacun.

Des analyses mettant en lien syntaxe et prosodie et comparant des langues avec des systèmes accentuels différents ont permis de montrer que les phénomènes d'hypo- et d'hyper-articulation peuvent être expliqués selon deux sources : (1) les aspects lexicaux (accentuation lexicale, position du

phonème au sein du mot) et (2) les aspects prosodiques (catégories prosodiques tels que le groupe accentuel et le groupe intonatif). Ces deux sources peuvent générer une hyper-articulation des voyelles mais la première se fera de façon prédominante sous la forme d'un renforcement spectral alors que la deuxième sera avant tout véhiculée par un allongement, qui aura ensuite pour effet un renforcement spectral.

J'ai également pu montrer dans le cadre de l'analyse du schwa que la prise en compte de multiples facteurs était possible et souhaitable dans des grands corpus de parole non préparée, et que celle-ci permettait de répondre à des questionnements phonologiques. En effet, la mise en évidence de variables différentes pour la réduction du schwa vs. son élision complète a permis de conclure à des mécanismes différents, l'un phonétique et l'autre phonologique.

L'analyse du /ʀ/ français standard d'après une combinaison de corpus de données articulatoires et de grands corpus de parole a permis de considérer la forme non voisée du /R/ comme la réalisation hyper-articulée de la forme voisée, et a montré que la variation du /R/ est grandement influencée par la position prosodique et par le style de parole, en plus du contexte consonantique.

Pour finir, dans une étude postulant que /e/ et /ɛ/ sont entrés dans un processus de fusion, j'ai montré que les grands corpus avec de multiples locuteurs sont des outils utiles pour repérer des tendances globales dans une langue malgré le maintien de variations inter-locuteurs.

Mes travaux récents m'ont guidé vers la recherche de stratégies propres au locuteur et de sa caractérisation phonétique. Depuis moins de dix ans, les réseaux de neurones profonds ont bouleversé le domaine de la classification en dépassant la plupart des systèmes, et il paraissait indispensable d'essayer de les utiliser pour l'analyse phonétique. En ayant recours à des réseaux de neurones convolutifs (CNN) par le biais des spectrogrammes, le but est double : (1) savoir jusqu'à quel point le spectrogramme permet de caractériser le locuteur au-delà d'une analyse phonétique classique et (2) au moyen de techniques de visualisation, parvenir à localiser les zones du spectrogrammes utilisées par les CNN. Des résultats encourageants présentés dans la discussion finale donnent un aperçu de mes projets de recherche.

6 Bibliographie

Adank, P. M. (2003). *Vowel Normalization. A Perceptual acoustic study of Dutch Vowels*. sn: sl.

Adda-Decker, M. (2007). Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole. *Actes Des 5èmes Journées Linguistiques de Nantes*, 211–216.

Adda-Decker, M., de Mareüil, P. B., Adda, G., & Lamel, L. (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, *46*(2), 119–139.

Adda-Decker, M., Gendrot, C., & Nguyen, N. (2008). Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique Des Langues*, *49*, 13–46.

Adda-Decker, M., Gendrot, C., Snoeren, N., & Nguyen, N. (2013). *Apport du traitement automatique à l'étude des voyelles*. Hermes Science Publications.

Ajili, M., Bonastre, J.-F., Rossetto, S., & Kahn, J. (2016). Inter-speaker variability in forensic voice comparison: a preliminary evaluation. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2114–2118.

Al-Tamimi, J.-E., & Ferragne, E. (2005). Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. *Ninth European Conference on Speech Communication and Technology*.

Audibert, N., Fougeron, C., Gendrot, C., & Adda-Decker, M. (2015). Duration-vs. style-dependent vowel variation: A multiparametric investigation. *18th International Congress of Phonetic Sciences (ICPhS'15)*, 5.

Avanzi, M., Brunetti, L., & Gendrot, C. (2012). Extra-Sentential Elements, Prosodic Restructuring, and Information Structure. A Study of Clitic-Left Dislocation in Spontaneous French. *Speech Prosody 2012, Sixth International Conference*.

Avanzi, M., Gendrot, C., & Lacheret-Dujour, A. (2010). Is there a prosodic difference between left-dislocated and heavy subjects? Evidence from spontaneous French. *Speech Prosody 2010-Fifth International Conference*.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Bagou, O., Michel, V., & Laganaro, M. (2009). On the production of sandhi phenomena in French: Psycholinguistic and acoustic data. *Tenth Annual Conference of the International Speech Communication Association*.

Barnes, J., & Kavitskaya, D. (2002). Phonetic analogy and schwa deletion in French. *Annual Meeting of the Berkeley Linguistics Society*, *28*(1), 39–50.

Béchet, F. (2001). LIA PHON: un systeme complet de phonétisation de textes. *Traitement Automatique Des Langues*, *42*(1), 47–67.

Beckman, M. E. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. *Papers in Laboratory Phonology II: Segment, Gesture, Prosody*.

Besacier, L., & Bonastre, J.-F. (1998). Frame pruning for speaker recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, *2*, 765–768.

Bigi, B., & Meunier, C. (2018). Automatic segmentation of spontaneous speech. *Revista de Estudos Da Linguagem*, *26*(4).

Bladon, A., & Fant, G. (1978). A two-formant model and the cardinal vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *19*(1), 1–8.

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

Bolinger, D. (1972). Accent is predictable (if you're a mind-reader). *Language*, 633–644.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3–4), 255–272.

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2014). Train & Align: Un Outil d'Alignement Phonétique Automatique Disponible en Ligne. *Paper Presented Atthe Journées d'étude de La Parole (JEP), Le Mans*.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180.

Bürki, A., Ernestus, M., & Frauenfelder, U. H. (2010). Is there only one “fenêtre” in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *Journal of Memory and Language*, *62*(4), 421–437.

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. *The Journal of the Acoustical Society of America*, *130*(6), 3980–3991.

Bürki, A., Fougeron, C., Gendrot, C., & Frauenfelder, U. H. (2011). Phonetic reduction versus phonological deletion of French schwa: Some methodological issues. *Journal of Phonetics*, *39*(3), 279–288.

Bürki, A., Gendrot, C., Gravier, G., Linarès, G., & Fougeron, C. (2008). Alignement automatique et analyse phonétique: comparaison de différents systèmes pour l’analyse du schwa. *Traitement Automatique Des Langues*, *49*(3), 165–197.

Bybee, J. (2001). Frequency effects on French liaison. *Typological Studies in Language*, *45*, 337–360.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Calliope, E. P. (1989). JP TUBACH. *Collection:“La Parole et Son Traitement Automatique,” Collection Technique et Scientifique Des Télécommunications*.

Chafcouloff, M. (1980). Les caractéristiques acoustiques de (j, ɣ, w, l, r) en français. *Travaux de l’Institut de Phonétique d’Aix Aix-En-Provence*, *7*, 7–56.

Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, *102*(4), 2360–2370.

Chen, W.-R., Whalen, D. H., & Shadle, C. H. (2019). F 0-induced formant measurement errors result in biased variabilities. *The Journal of the Acoustical Society of America*, *145*(5), EL360–EL366.

Chignoli, G., Gendrot, C., & Ferragne, E. (2020). Caractérisation du locuteur par CNN à l’aide des contours d’intensité et d’intonation: comparaison avec le spectrogramme. *6e Conférence Conjointe Journées d’Études Sur La Parole (JEP, 31e Édition), Traitement Automatique Des Langues Naturelles (TALN, 27e Édition),* 91–99.

Chistovich, L. A., Sheikin, R. L., & Lublinskaja, V. V. (1979). Centres of Gravity’and Spectral Peaks as the Determinants of Vowel Quality," in/bf Frontiers of Speech Comminication Research, B. Lindblom and S. Ohman, Eds. *Academic Press, London*, *143*, 157.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of/ɑ, i/in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878.

Clément, L., Lang, B., & Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. *LREC 04*, 1841–1844.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. *Papers in Laboratory Phonology*, *1*, 283–333.

Côté, M. H., & Morrison, G. S. (2007). The nature of the schwa/zero alternation in French clitics: Experimental and non-experimental evidence. *Journal of French Language Studies*, *17*(2), 159.

Cucchiarini, C., & Strik, H. (2003). Automatic phonetic transcription: An overview. *Proceedings of ICPHS*, 347–350.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

de Mareüil, P. B., d’Alessandro, C., Yvon, F., Aubergé, V., Vaissière, J., & Amelot, A. (2000). A French Phonetic Lexicon with Variants for Speech and Language Processing. *LREC*.

Delattre, P. (1965). *The general phonetic characteristics of languages.* Julius Gross Verlag.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773.

Dell, F. (1985). Les règles et les sons. *Paris, France: Hermann*.

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

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). *Sheep, goats, lambs and wolves: A statistical*

analysis of speaker performance in the NIST 1998 speaker recognition evaluation. National Inst of Standards and Technology Gaithersburg Md.

Drugman, T., & Dutoit, T. (2019). Glottal closure and opening instant detection from speech signals. *ArXiv Preprint ArXiv:2001.00841*.

Ernestus, M. T. C. (2000). *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: LOT.

Fant, G. (1970). *Acoustic theory of speech production* (Issue 2). Walter de Gruyter.

Ferragne, E., Gendrot, C., & Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. *ICPhS*, ISBN-978.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135.

Fougeron, C. (2007). Word boundaries and contrast neutralization in the case of enchaînement in French. *Papers in Laboratory Phonology IX: Change in Phonology*, 609–642.

Fougeron, C., Gendrot, C., & Bürki, A. (2007). Le schwa: une voyelle comme les autres? *5èmes Journées d'Etudes Linguistiques*, 191–198.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical Society of America*, *90*(4), 1816–1827.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*(1), 29–47.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. *Vocal Physiology: Voice Production, Mechanisms and Functions*, 347–355.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J. F., Mostefa, D., & Choukri, K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 139–142.

Gendrot, Cedric. (2017). Perception and Production of Word-Final/ʁ/in French. *INTERSPEECH*, 3926–3930.

Gendrot, Cédric. (2005). *Aspects perceptifs, physiologiques et acoustiques de différentes catégories prosodiques en français*. Paris 3.

Gendrot, Cédric. (2013). De la normalisation formantique des voyelles . In N Nguyen (Ed.), *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*. Hermes/Lavoisier. https://halshs.archives-ouvertes.fr/halshs-01422367

Gendrot, Cédric, & Adda-Decker, M. (2004). Analyses formantiques automatiques de voyelles orales: évidence de la réduction vocalique en langues française et allemande. *Workshop MIDL04*.

Gendrot, Cédric, & 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. *Ninth European Conference on Speech Communication and Technology*.

Gendrot, Cédric, & Adda-Decker, M. (2010). *Influence du contexte consonantique et de la durée des voyelles sur la centralisation des voyelles orales en français.* l'Harmattan.

Gendrot, Cédric, & Adda-Decker, M. (2007). Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1417–1420.

Gendrot, Cédric, Adda-Decker, M., & Schmid, C. (2012). Comparaison de parole journalistique et de parole spontanée: analyses de séquences entre pauses. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 1: JEP*, 649–656.

Gendrot, Cédric, Adda-Decker, M., & Vaissière, J. (2008). Les voyelles/i/et/y/du français: focalisation et variations formantiques. *XXVIIèmes Journées d'Étude Sur La Parole*, 205–208.

Gendrot, Cédric, & Audibert, N. (2019). La distinction/e/vs/ɛ/en français standard est-elle maintenue en finale de mot? Étude

sur des corpus de parole journalistique et de parole spontanée. *Langue Française*, *3*, 53–66.

Gendrot, Cédric, Demolin, D., & Kühnert, B. (n.d.). Aerodynamic, articulatory and acoustic realization of French /ʁ/. In *The socio-phonetics of rhotics, Studies on Language Variation series* (John Benja, pp. 3926–3930).

Gendrot, Cédric, Ferragne, E., & Pellegrini, T. (2019). Deep learning and voice comparison: phonetically-motivated vs. automatically-learned features. *ICPhS*.

Gendrot, Cédric, Ferragne, E., & Pellegrini, T. (2020). Informations segmentales pour la caractérisation phonétique du locuteur: variabilité inter-et intra-locuteurs. *Actes de La 6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition),* 262–270.

Gendrot, Cédric, Gerdes, K., & Adda-Decker, M. (2016). Détection automatique d'une hiérarchie prosodique dans un corpus de parole journalistique. *Langue Française*, *3*, 123–149.

Gendrot, Cédric, Henrich, N., Sshade, G., Muller, F., & Expert, R. (2004). Vocal folds vibratory patterns of laryngeal mechanism M0 as investigated with high speed cinematography and electroglottography. *International Conference on Voice Physiology and Biomechanics, Marseille, France*.

Gendrot, Cédric, Léonard, J. L., & Polian, G. (2010). Correlación laringovelar y variación dialectal del tseltal (Maya occidental, Chiapas, México): enfoques del proyecto ALTO (CIESAS Sureste/París 3). *Estudis Romànics*, 311–329.

Gendrot, Cédric, & Schmid, C. (2011). F0 Declination in French: Broadcast News versus spontaneous speech. *Nijmegen Workshop in Production and Comprehension of Conversational Speech.*, 15–17.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2019). Pourquoi se tourner vers le SUD: L'importance de choisir un schéma d'annotation en dépendance surface-syntaxique. *LIFT 2019-Journées Scientifiques" Linguistique Informatique, Formelle & de Terrain"*.

Goldman, J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. *Interspeech'11, 12th Annual Conference of the International Speech Communication Association*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, Issue 2). MIT press Cambridge.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Gowda, D., Airaksinen, M., & Alku, P. (2017). Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation. *The Journal of the Acoustical Society of America*, *142*(3), 1542–1553.

Grammont, M. (1914). Traité de phonétique. *Paris: Librairie Delagrave*.

Hansen, A. B. (1994). Etude du E caduc—stabilisation en cours et variations lexicales1. *Journal of French Language Studies*, *4*(1), 25–54.

Harmegnies, B., & Poch-Olivé, D. (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech Communication*, *11*(4–5), 429–437.

Hart, J. (1990). *A perceptual study of intonation: An experimental phonetic approach to speech melody/J.'t Hart, R. Collier, A. Cohen*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., & Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems. *Intonation Systems: A Survey of Twenty Languages*, 1–44.

Hualde, J. I. (2012). Stress and rhythm. *The Handbook of Hispanic Linguistics*, *69*, 153–172.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In *Intonation* (pp. 209–242). Springer.

Kahn, J., Audibert, N., Bonastre, J.-F., & Rossato, S. (2011). Inter and Intra-speaker Variability in French: An Analysis of Oral Vowels and Its Implication for Automatic Speaker Verification. *ICPhS*, 1002–1005.

Keating, P. (1985). *Universal phonetics and the organization of grammars. V. Fromkin (ed.) Phonetic Linguistics*. Academic Press.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, 143–161.

Kessens, J. M., & Strik, H. (2004). On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. *Computer Speech & Language*, *18*(2), 123–141.

Kewley-Port, D., & Zheng, Y. (1999). Vowel formant discrimination: Towards more ordinary listening conditions. *The Journal of the Acoustical Society of America*, *106*(5), 2945–2958.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

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.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261–2271.

Labov, W. (1972). *Sociolinguistic patterns* (Issue 4). University of Pennsylvania Press.

Lamel, L. F., Gauvain, J.-L., & Eskenazi, M. (1991). BREF, a large vocabulary spoken corpus for French. *Second European Conference on Speech Communication and Technology*.

Lanchantin, P., Morris, A. C., Rodet, X., & Veaux, C. (2008). Automatic Phoneme Segmentation with Relaxed Textual Constraints. *LREC*.

Lehiste, I. (1970). *Suprasegmentals.* (C. M. Press. (ed.)). Massachusetts Inst. of Technology P.

Levelt, W. J. M., & JM, W. (1989). *Speaking: From intention to articulation.-" A bradford book."*. MIT Press.

Lieberman, P. (1967). Intonation, perception, and language. *MIT Research Monograph*.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

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

Lozano-Diez, A., Plchot, O., Matejka, P., & Gonzalez-Rodriguez, J. (2018). DNN based embeddings for language recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5184–5188.

Maeda, S. (1976). *A characterization of American English intonation; PhD diss*. MIT.

Maeda, S., Huffman, M., & Krakow, R. (1993). *Phonetics and Phonology: Nasals, Nasalization and the Velum*. Academic Press, Ch. Acoustics of vowel nasalization and articulatory shifts ….

Malécot, A. (1976). The effect of linguistic and paralinguistic variables on the elision of the French mute-e. *Phonetica*, *33*(2), 93–112.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech*, *2017*, 498–502.

Meunier, C. (1994). *Les groupes de consonnes: problématique de la segmentation et variabilité acoustique*. Aix-Marseille 1.

Meunier, C., & Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, *39*(3), 271–278.

Meunier, C., Frenck-Mestre, C., Lelekov-Boissard, T., & Le Besnerais, M. (2003). Production and perception of vowels: does the density of the system play a role? *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, 723–726.

Moon, S., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, *96*(1), 40–55.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *ArXiv Preprint*

ArXiv:1706.08612.

Navarro Tomás, T. (194 C.E.). *Manual de entonación española* (New York:). Hispanic Institute in the United States.

Nespor, M., & Vogel, I. (2007). *Prosodic phonology: with a new foreword* (Vol. 28). Walter de Gruyter.

Nguyen, Noël, & Espesser, R. (2004). Méthodes et outils pour l'analyse acoustique des systèmes vocaliques. *Bulletin PFC (Phonologie Du Français Contemporain)*, *3*, 77–85.

Nguyen, Noël, & Fagyal, Z. (2008). Acoustic aspects of vowel harmony in French. *Journal of Phonetics*, *36*(1), 1–27.

Nolan, F. (2007). Voice quality and forensic speaker identification. *Govor*, *24*(2), 111–128.

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. *The Handbook of Phonetic Sciences*, *5*, 640–673.

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1–2), 155–161.

Ohala, J. J., & Ewan, W. G. (1973). Speed of pitch change. *The Journal of the Acoustical Society of America*, *53*(1), 345.

Padgett, J., & Tabain, M. (2005). Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica*, *62*(1), 14–54.

Pallier, C. (1994). *Rôle de la syllabe dans la perception de la parole: études attentionnelles*. Ecole des hautes Etudes en Sciences sociales (EHESS).

Pellegrini, T. (2017). Densely connected CNNs for bird audio detection. *2017 25th European Signal Processing Conference (EUSIPCO)*, 1734–1738.

Perrier, P., Ostry, D. J., & Laboissière, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech, Language, and Hearing Research*, *39*(2), 365–378.

Pierrehumbert, J. (2001). Lenition and contrast. *Frequency and the Emergence of Linguistic Structure*, *45*, 137.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, *CONF*.

Pruthi, T., & Espy-Wilson, C. Y. (2004). Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, *43*(3), 225–239.

Quilis, A. (1981). *Fonética acústica de la lengua española*. Gredos Madrid.

Racine, I. (2007). Effacement du schwa dans des mots lexicaux: constitution d'une base de données et analyse comparative. *Proceedings of JEL*, 125–130.

Racine, I., & Grosjean, F. (2002). La production du E caduc facultatif est-elle prévisible? Un début de réponse. *Journal of French Language Studies*, *12*(3), 307.

Ridouane, R., & Gendrot, C. (2017). On ejective fricatives in Omani Mehri. *Brill's Journal of Afroasiatic Languages and Linguistics*, *9*(1–2), 139–159.

Schmid, C., Gendrot, C., & Adda-Decker, M. (2012). Une comparaison de la déclinaison de F0 entre le français et l'allemand journalistiques. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 1: JEP*, 329–336.

Schwartz, J.-L., Beautemps, D., Abry, C., & Escudier, P. (1993). Inter-individual and cross-linguistic strategies for the production of the [i] vs.[y] contrast. *Journal of Phonetics*, *21*(4), 411–425.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, *25*(3), 255–286.

Shadle, C. H., Nam, H., & Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *The Journal of the Acoustical Society of America*, *139*(2), 713–727.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1–2), 3–45.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309–1325.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100.

Tabain, M. (2003). Effects of prosodic boundary on/aC/sequences: articulatory results. *The Journal of the Acoustical Society of America*, *113*(5), 2834–2849.

Tabain, M., & Perrier, P. (2005). Articulation and acoustics of/i/in preboundary position in French. *Journal of Phonetics*, *33*(1), 77–100.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1–20.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, *52*(3), 201–212.

Torreira, F., & Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology*, *2*(2), 331–353.

Traunmüller, H. (1984). Articulatory and perceptual factors controlling the age-and sex-conditioned variability in formant frequencies of vowels. *Speech Communication*, *3*(1), 49–61.

Traunmüller, H., & Eriksson, A. (1995). The perceptual evaluation of F 0 excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America*, *97*(3), 1905–1915.

Vaissière, J. (1985). The use of allophonic variations of/a/in automatic continuous speech recognition of French. *The Journal of the Acoustical Society of America*, *77*(S1), S12–S12.

Vaissière, J. (2001). Changements de sons et changements prosodiques: du latin au français: du latin au français. *Revue Parole*, *17*, 53–88.

Vaissière, J. (2007). Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. *Experimental Approaches to Phonology*, 54–71.

Vaissière, J. (2009). Articulatory modeling and the definition of acoustic-perceptual targets for reference vowels. *The Chinese Phonetics Journal*, *2*, 22–33.

Vaissière, J. (2010). Le français, langue à frontières par excellence. In M.-A. Delomiier, D., Morel (Ed.), *Du linguistique au sémiotique* (pp. 10–20). Lucas, Lambert.

Vaissière, J., & Michaud, A. (2006). Prosodic constituents in French: a data-driven approach. In Y. K. & T. M. I. Fónagy (Ed.), *Prosody and syntax* (pp. 47–64). John Benjamins.

Van Bael, C., Boves, L., Van Den Heuvel, H., & Strik, H. (2007). Automatic phonetic transcription of large speech corpora. *Computer Speech & Language*, *21*(4), 652–668.

Van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, *12*(1), 1–23.

Wagner, P., & Dellwo, V. (2004). Introducing YARD (Yet Another Rhythm Determination) and re-introducing isochrony to rhythm research. *Proceedings of Speech Prosody*.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12.

Weber, P., Bai, L., Russell, M. J., Jancovic, P., & Houghton, S. M. (2016). Interpretation of Low Dimensional Neural Network Bottleneck Features in Terms of Human Perception and Production. *INTERSPEECH*, 3384–3388.

Wesenick, M.-B., & Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, *1*, 129–132.

Wioland, F. (1985). Les structures Syllabiques du français : fréquence et distribution des phonèmes consonantiques, contraintes idiomatiques dans les séquences consonantiques,. In *Slatkine-Champion, Paris*.

Woehrling, C., & Mareüil, P. B. de. (2007). Comparing Praat and Snack formant measurements on two large corpora of northern and southern French. *Eighth Annual Conference of the International Speech Communication Association*.

Wright, R. (2004). Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology VI*, 75–87.

Wu, Y., Adda-Decker, M., Gendrot, C., & Lamel, L. (2019). Impact of post-lexical context and speech style on word-final/ʁ/realization in French using large corpora and automatic speech processing. *R-Atics 6*.

Wu, Y., Gendrot, C., Adda-Decker, M., & Fougeron, C. (2019). Post-consonantal word-final/k/realization in french: contributions of large corpora. *19th International Congress of Phonetic Sciences*.

Wu, Y., Gendrot, C., Hallé, P., & Adda-Decker, M. (2015). On Improving the Pronunciation of French/r/in Chinese Learners by Using Real-Time Ultrasound Visualization. *ICPhS 2015 (18th International Congress of Phonetic Sciences)*.

Yeou, M., & Maeda, S. (1995). Pharyngeal and uvular consonants are approximants: An acoustic modeling study. *Proceedings of the 13th International Congress of Phonetic Sciences*, 586–589.

Young, S. J., Evermann, G., & Gales, M. (2006). *The HTK Book Version 3.4, Cambridge University*.

Yuan, J., & Liberman, M. (2010). F0 declination in English and Mandarin broadcast news speech. *Eleventh Annual Conference of the International Speech Communication Association*.