Contributions du traitement automatique de la parole à l'étude des voyelles orales du français

Martine Adda-Decker[1] — Cédric Gendrot[2] — Noël Nguyen[3]

[1] *Groupe Traitement du Langage Parlé, LIMSI-CNRS Orsay*
madda@limsi.fr

[2] *LPP-CNRS et Université Paris 3*
cgendrot@univ-paris3.fr

[3] *LPL-CNRS et Université Aix-Marseille*
noel.nguyen@lpl-aix.fr

RÉSUMÉ. *Le traitement automatique de la parole peut concrètement contribuer à éclairer de nombreuses questions concernant la variabilité phonémique à l'oral. L'exploitation de grandes masses de données permet ainsi de dégager de grandes tendances, dont une interprétation plus fine repose ensuite à la fois sur un éclairage linguistique et sur un certain nombre de précautions méthodologiques. Nous allons focaliser l'étude sur la variabilité des voyelles orales en français. Des mesures de durée et de formants, à partir des grands corpus PFC et ESTER utilisés à la fois par les chercheurs en linguistique et en traitement automatique de la parole permettront d'illustrer l'impact de divers paramètres, notamment le style de parole, la position syllabique et l'origine régionale des locuteurs. Enfin la réalisation des voyelles mi-fermées antérieures sera examinée à l'aide de classification automatique de variantes dans un cadre bayésien.*

ABSTRACT. *Automatic speech processing methods and tools can contribute to shedding light on many issues relating to phonemic variability in speech. The processing of huge amounts of speech thus allows to extract main tendencies, for which detailed interpretations then require both linguistic and methodological insights. The experimental study focuses on the variability of French oral vowels in the PFC and ESTER corpora, which are widely used both by linguists and researchers in automatic speech processing. Duration and formant measures allow to illustrate global variations depending on different parameters, which include speech style, syllable position and the speakers' regional origins. The last part addresses the phonetic realization of close-mid front vowels, using automatic classification in a Bayesian framework.*

MOTS-CLÉS : *traitement automatique de la parole, RAP, alignement de la parole, transcription phonémique, classification, grands corpus, variation phonologique et phonétique, f_0, durée, suivi de formants, voyelles moyennes, harmonie vocalique.*

KEYWORDS: *automatic speech processing, ASR, speech alignment, phonemic transcription, classification, large corpora, phonological and phonetic variation, syllable, f_0, duration, formant tracking, mid vowels, vowel harmony.*

1. Introduction

À l'heure actuelle, les systèmes de transcription automatique de la parole permettent de produire de grandes quantités d'oral transcrit et annoté à des niveaux de granularité différents. Ils permettent ainsi d'associer au signal de parole les mots prononcés, les prononciations réalisées, et de mesurer les durées et autres descripteurs acoustiques. Afin de traiter la performance orale de locuteurs réels, les systèmes de transcription automatique reposent généralement sur une modélisation statistique de la parole à deux niveaux : la modélisation acoustique des mots (spécifique à la modalité audio), et la modélisation appelée *linguistique* (concernant la modalité écrite) pour la génération d'une séquence de mots. Dans ce cadre, la réalisation acoustique des sons de la langue est modélisée par des densités de probabilité de paramètres acoustiques, et l'ensemble des phrases possibles est approché par des distributions probabilistes de suites de mots (sous forme de chaînes markoviennes). L'estimation des modèles statistiques est faite alors à partir de corpus audio et textuels, ces derniers provenant surtout de sources écrites, mais aussi de parole transcrite. Ainsi, des suites de mots très fréquentes dans la langue, comme par exemple `de la`, auront une probabilité d'apparition élevée, alors que des séquences *a priori* très peu probables comme `la de` resteront néanmoins possibles avec une faible probabilité dans le modèle de langue n-gramme.

Au-delà de leur vocation première de produire une transcription de l'oral, les systèmes de reconnaissance automatique de la parole peuvent servir d'instruments linguistiques pour explorer de façon cohérente des corpus virtuellement illimités. Ainsi, l'utilisation pour les besoins de la phonétique et de la phonologie de cette nouvelle technologie permet d'envisager un large spectre d'études, en partie déjà imaginé par les grands linguistes d'il y a un siècle, études qui étaient alors matériellement irréalisables. En effet, la mise au point de systèmes de transcription automatique s'accompagne de la création de centaines, voire de milliers d'heures de ressources orales enrichies, qu'il s'agit maintenant d'explorer. On peut sélectionner ou filtrer, dans ces corpus, les occurrences d'un phénomène à l'étude, par exemple la réalisation du schwa, ce qui permet d'abord de quantifier son importance dans certaines configurations d'usage (parole familière ou formelle, monologue ou dialogue, français standard ou variante régionale…). Ensuite, cette démarche permet de décrire le périmètre de variation de ce phénomène et de dégager les principales variables exerçant un effet sur lui. Des travaux sur ces sous-corpus permettront de mieux qualifier et quantifier la variabilité, et de formaliser de nouvelles connaissances concernant la variabilité en fonction d'un grand nombre de paramètres.

Il apparaît que dans la parole, tous les phonèmes ne sont pas articulés avec la même précision. Ceci n'est pas toujours lié à la nature du phonème (comme pour le schwa), mais souvent à sa place dans le message parlé. Il apparaît dans les corpus que les mots fréquents et/ou à forte redondance sont souvent peu articulés, voire inexistants dans certains cas extrêmes. Il semble que si l'information est donnée, soit par le niveau syntaxique, soit par le niveau pragmatique, alors le niveau acoustique n'a pas besoin d'être très complet (Meunier, 2005). Nous manquons à l'heure actuelle d'une description

détaillée de ces phénomènes. Comment les modéliser pour le traitement automatique de la parole ? Il s'agit surtout de pouvoir proposer des modèles acoustiques de mots plus courts que ceux générés par une modélisation phonologique standard et l'hypothèse est que ces raccourcissements ne se font pas au hasard. Pour augmenter nos connaissances autour de ces questions, nous pouvons utiliser ces mêmes systèmes de transcription comme outil d'analyse afin de qualifier et de quantifier sous-articulations et réductions, appelées *métaplasmes*. L'exemple de la figure 1 montre la séquence de mots `je crois que c'est quelque chose`, dont la prononciation canonique est /ʒəkʁwakəsɛkɛlkəʃoz/. On peut observer non seulement la chute de toutes les voyelles schwa, mais plus spectaculairement, la séquence `c'est quelque` (zone entre 0,15 et 0,3 secondes sur le spectrogramme) est temporellement très réduite, admettant [sek] comme transcription phonétique approximative. Le mot `quelque` se trouve ainsi réduit à la seule consonne /k/. Alors que ces phénomènes ne représentent pas le cœur de l'étude présentée ici, ils ont contribué à orienter les études développées ci-après, notamment le volet sur les durées vocaliques.

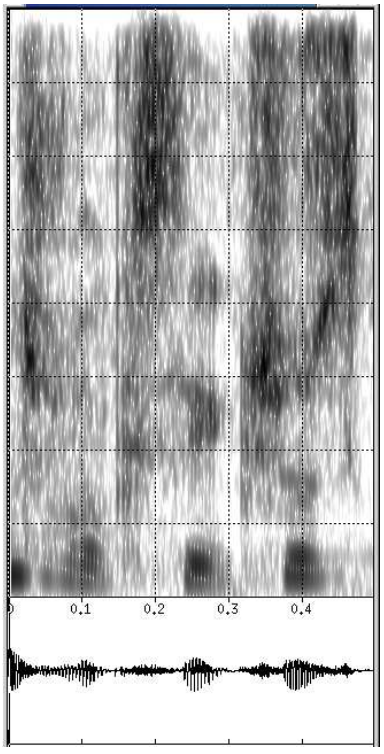

Figure 1. *Exemple de métaplasme extrait d'une interview politique :* `je crois que c'est quelque chose` *articulé comme* [ʃʁwaksekʃoz]

Dans cette contribution nous nous intéressons aux apports du traitement automatique de la parole à l'étude de la variation en français parlé, avec un intérêt tout particulier pour les voyelles. Nous décrirons rapidement quelques sources de variabilité telles qu'elles sont mises en évidence par les systèmes de transcription automatique. Nous montrerons ensuite plus en détail différentes utilisations des systèmes de transcription, et plus particulièrement des systèmes d'alignement, en tant qu'instrument de mesure à des fins de recherches linguistiques (Habert, 2005). Nous aborderons la question de la validité des mesures en lien avec la fiabilité des frontières des segments induites par l'alignement. La durée vocalique sera ainsi examinée en fonction de différents alignements, de différents styles de parole, en fonction de la position de la voyelle dans le mot lexical ou prosodique, et en fonction de la variante régionale. Finalement la question de la réalisation des voyelles mi-ouvertes sera examinée à l'aide de l'instrument de traitement automatique.

2. Corpus utilisés

Le corpus PFC (phonologie du français contemporain) vise à rassembler des centaines d'heures de parole (lecture et entretiens), dont quelques dizaines d'heures ont servi aux études présentées (Durand *et al.*, 2003). Le corpus PFC collecté dans des dizaines de points d'enquête de l'espace francophone, permet d'étudier l'influence du style de parole et de l'accent régional sur la production langagière (inventaire phonémique, variantes de prononciation, réalisations de schwa et de liaisons, lexique...). La comparaison, pour les mêmes locuteurs, entre la lecture d'un article journalistique (qui n'est qu'une oralisation de l'écrit) et des entretiens libres et guidés, qui correspondent à du « vrai » oral, est particulièrement intéressante.

Lors de la campagne ESTER (Galliano *et al.*, 2005) (évaluation des systèmes de transcription enrichie d'émissions radiophoniques), financée par le programme interministériel français TECHNOLANGUE et organisée conjointement par l'AFCP [1], la DGA/CTA [2] et ELDA [3], 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. 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 de centaines d'heures de journaux radiodiffusés à la fin des années 1990. Dans cette dyna-

1. Association Francophone de la Communication Parlée `www.afcp-parole.org`.
2. Délégation Générale à l'Armement, Centre Technique d'Arcueil.
3. Evaluations & Language resources Distribution Agency `www.elda.org`.

mique a été développé le logiciel TRANSCRIBER (Barras *et al.*, 2001), qui a trouvé un large succès pour la transcription de corpus oraux, bien au-delà de la communauté du traitement automatique de la parole. L'exercice de transcription manuelle pointe sur des problèmes qui se retrouveront également lors de la transcription automatique. Ainsi, on peut se rendre compte à l'écoute attentive que bon nombre de mots sont souvent réalisés de manière incomplète. On pourrait être tenté d'écrire `ad'taleur` pour `à tout à l'heure`, tout comme on peut trouver fréquemment des « trucages orthographiques » comme `'ya` pour `il y a`. Les transcriptions manuelles sont faites en orthographe normative, comme le préconisent également les conventions du GARS (Blanche-Benveniste, 1999), avec un minimum d'indications de prononciations. Ce principe permet de converger au mieux vers des transcriptions stables indépendantes du transcripteur et avec un temps de transcription plus faible que si des annotations spécifiques étaient effectuées. Le problème de « trucages » des prononciations (*cf.* métaplasmes) est cependant bien réel et nécessite des adaptations au niveau de la modélisation acoustique des mots.

Le passage de la lecture à la parole radiophonique a un impact au niveau des prononciations, avec des réalisations qui peuvent s'écarter de manière importante de prononciations canoniques (Duez, 2003). Les mots-outils fréquents, dont l'information est largement portée par le contexte, sont souvent mal et peu articulés ; sous l'effet de répétition, des mots pleins thématiques bien prononcés en début d'émission, finissent par être raccourcis, de telle sorte que ne subsistent que certaines parties (comme, par exemple, [aRʃɛktyR] pour le mot `architecture` /aRʃitɛktyR/)). 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.

Le tableau 1 donne des ordres de grandeur de quelques paramètres caractéristiques de systèmes de transcription en français, notamment les corpus d'apprentissage pour les modèles acoustiques et les modèles de langue n-grammes ainsi que les taux d'erreur de mots obtenus en 2005 (Gauvain *et al.*, 2005a). Afin de garantir une bonne couverture, le vocabulaire du système contient 200 000 mots. Les entrées lexicales sont directement les formes fléchies telles qu'elles sont observées dans le langage courant. Ainsi un verbe comme *aller* peut avoir des dizaines d'entrées différentes. Pour certains locuteurs professionnels, les taux d'erreur de mots peuvent descendre autour de 5 %. Mais il est actuellement difficile d'approcher des taux aussi faibles pour une population large de locuteurs, sur des thèmes variés et des styles de parole plus spontanés (Gauvain *et al.*, 2005b).

Style	*Mod. acoustique*	*Mod. n-gramme*	*Taille voc.*	*% err.*
Journalistique	100 h de radio $\sim$ 1 M mots	400 M mots	200 k mots	11 %

Tableau 1. *Quelques données concernant la transcription automatique du français : style de parole, quantité de données acoustiques et textuelles pour l'apprentissage des modèles, taille du vocabulaire et taux d'erreur de mots*

3. Traitement automatique de grands corpus

3.1. *Modélisation acoustique*

La reconnaissance automatique de la parole vise à convertir le signal acoustique (signifiant acoustique) en symboles graphémiques (signifiant écrit). La variabilité phonétique observée dans le signal acoustique devrait pouvoir être ignorée afin de recouvrer la suite de mots prononcée *via* une prononciation standard. Cette problématique s'apparente au moins partiellement à celle des premiers phonologues de la fin XIX^e et début du XX^e siècle, comme Baudouin de Courtenay, Saussure ou Troubetzkoy. Ainsi on peut lire : « *distinguer deux* phonétiques *descriptives distinctes, suivant qu'on veut étudier les sons phoniques comme des signaux physiques (phonétique) ou comme des éléments abstraits, sons distinctifs d'un système linguistique (phonologie)*. Les travaux de ces phonologues structuralistes visaient à établir pour une langue donnée le système de phonèmes minimal en recherchant *quelles différences phoniques sont liées dans la langue étudiée à des différences de signification, comment les éléments de différenciation se comportent entre eux et selon quelles règles ils peuvent se combiner les uns avec les autres... Le phonologue ne doit envisager en fait de son que ce qui remplit une fonction déterminée dans la langue* » (Troubetzkoy, 1939-1976).

La notation phonologique, très économe, permet d'associer à chaque mot (écrit) une prononciation idéale. La linéarité du signifiant acoustique (déroulement dans le temps) entraîne que les éléments se présentent *a priori* les uns après les autres : ils forment une chaîne. Cette représentation phonologique sous forme de chaîne de phonèmes est utilisée pour la modélisation acoustique des mots en reconnaissance automatique de la parole. À cette représentation sont alors associés des modèles de Markov cachés (MMC) à mélange de gaussiennes permettant de rendre compte de la variation phonétique effectivement observée dans le signal physique. Pour chaque phonème, la modélisation acoustique permet un nombre élevé de MMC élémentaires (qu'on peut considérer comme modèles d'allophones), afin de couvrir les variations induites par le contexte phonémique. Suivant que ce contexte se limite à un (respectivement deux) phonème(s) à gauche et à droite, on parle de modèles acoustiques triphones (respectivement quinphones). La modélisation par phonèmes en contexte reflète le fait que les phones (*i.e.* réalisations acoustiques des phonèmes dans les segments acoustiques) dépendent du contexte phonémique. Ce choix de modélisation permet, à partir d'un

niveau phonémique très générique, de faire appel à des modèles acoustiques précis, car tenant compte des contraintes de coarticulation induites par les contextes. Ainsi, dans les systèmes de transcription actuels, des milliers d'allophones en contexte sont modélisés. Ces modèles d'allophones en contexte sont estimés à partir de grands corpus, transcrits à un niveau phonémique *via* les dictionnaires de prononciation, et ils permettent de rendre compte de manière implicite d'un grand nombre de variations acoustiques. Parmi les facteurs de variation phonémique les plus importants, on peut notamment citer : le style de parole, le contexte phonémique environnant, la durée, la fréquence fondamentale f_0, la position du phonème dans la syllabe, dans le mot et dans l'énoncé, la fréquence lexicale, le sexe du locuteur, l'accent lexical, l'accent régional, etc. Pour la plupart, ces facteurs ne sont pris en compte par les modèles acoustiques que de manière implicite.

La modélisation acoustique des mots *via* une représentation phonologique permet de garantir une certaine indépendance entre le lexique du corpus d'apprentissage et le lexique du système de transcription. En effet, les modèles de taille phonémique permettent de « synthétiser » n'importe quel mot (d'où l'indépendance entre vocabulaire d'apprentissage et vocabulaire d'application). Cependant comme les modèles acoustiques reflètent la variation observée dans le corpus d'apprentissage, ils seront d'autant plus appropriés que les variations observées dans le corpus d'application sont similaires à celles de l'apprentissage. La similarité fait référence alors, non seulement à un lexique partagé, mais aussi à un style de parole similaire. Des modèles acoustiques estimés à partir d'un corpus de lecture ne seront pas nécessairement adaptés à une parole spontanée, même si (par construction) on garantissait une identité lexicale. Si la modélisation acoustique *via* des prononciations phonémiques est adaptée pour une parole bien articulée, le traitement d'une parole plus spontanée, plus relâchée, peut poser problème : les prononciations observées peuvent avoir un contenu segmental très différent de la forme canonique, car, comme nous l'avons évoqué auparavant, des phonèmes, voire des syllabes entières, peuvent disparaître (Duez, 2003; Adda-Decker *et al.*, 2005). Il est intéressant de lier ces phénomènes à la nature du phonème et à sa position dans le mot. La question est alors de savoir quels phonèmes sont les plus sujets à réduction temporelle (dont la disparition peut être vue comme le cas limite) et si certaines positions dans le mot (initiale, antépénultième, pénultième, finale) les rendent plus robustes que d'autres.

3.2. *Dictionnaires de prononciation*

Le dictionnaire de prononciation d'un système de reconnaissance automatique sert à déterminer, pour les mots qui y sont inclus, leur modélisation acoustique : un mot admettant une prononciation à N phonèmes sera représenté au niveau acoustique par une chaîne à $3 \times$ N états, ce qui implique alors une durée minimale de $3 \times \mathrm{N} \times \mathrm{t}$ pour chaque observation du mot. Trois états consécutifs sont associés à chaque phonème, et chaque état est représenté par un ensemble de densités gaussiennes de paramètres acoustiques décrivant les réalisations possibles pour cette partie de phonème. La durée

minimale provient du fait que tout passage d'un état à l'autre, y compris une boucle sur le même état, consomme une unité de temps t (ici t = 10 ms). Les gaussiennes sont estimées à partir d'un corpus d'apprentissage, segmenté au préalable utilisant le même type de dictionnaire de prononciation.

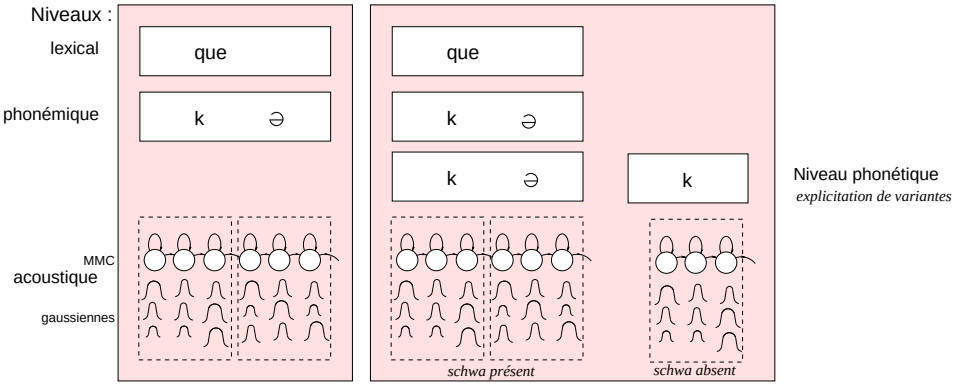

Figure 2. *Représentation schématique de la modélisation acoustique des mots via un dictionnaire de prononciation canonique (à gauche), incluant des variantes (à droite)*

Différentes options peuvent être prises pour le dictionnaire de prononciation : canonique ou incluant des variantes (voir figure 2). Prenons l'exemple du mot `que` avec une prononciation canonique /kə/. Le modèle acoustique correspondant est composé de six états et dure au moins 60 ms. La variante de prononciation [k] (chute du schwa) limitée à un seul phonème, n'admet que trois états avec une durée minimale de 30 ms. De manière générale, ces mots-outils monosyllabiques à noyau schwa (`de, le, ne, que, ce, se`...) sont modélisés avec leur forme canonique pour limiter des insertions erronées (d'autant plus faciles que les modèles sont courts) et la complexité de décodage (particulièrement élevée pour des phonèmes à la fois frontière gauche et droite de mot). Pour chaque entrée lexicale (forme fléchie), le dictionnaire de prononciation prévoit ainsi une prononciation canonique, qui correspond en général à une prononciation maximale (tous les phonèmes possibles sont prévus). Le nombre de variantes reste faible, afin de limiter le risque d'ambiguïté et la complexité de calculs, en particulier si les variantes concernent les frontières de mot. Ainsi, concernant les mots monosyllabiques à noyau schwa, le schwa est présent par défaut ; à l'inverse, pour les mots polysyllabiques, il n'y a en général pas de schwa final prévu dans le dictionnaire de prononciation. Pour la reconnaissance automatique de la parole, les dictionnaires de prononciation sont donc plutôt de type phonémique. La variabilité acoustique observée est modélisée de manière implicite *via* les mélanges de gaussiennes dans les MMC d'allophones. Cela implique alors qu'un modèle acoustique de consonne en fin de mot peut modéliser non seulement la consonne en question, mais également une voyelle épenthétique (*cf.* fig. 3 pour le /d/ final de `Bagdad` réalisé comme [bagdadə]).

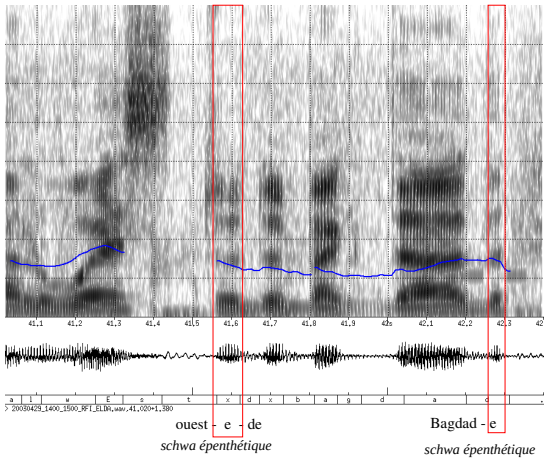

Figure 3. *Spectrogramme d'un extrait de radio, illustrant deux voyelles épenthétiques, la première (finale de* `ouest`*) était prévue par une variante de prononciation, la seconde (finale de* `Bagdad`*) est absorbée par le modèle du* /d/

De manière générale, plus la réalisation temporelle des mots s'écarte de la chaîne linéaire proposée *via* le dictionnaire de prononciation, moins les modèles acoustiques reflètent simplement le phonème visé, mais (également) son voisinage. Ce problème se pose particulièrement aux frontières des mots, où schwa et liaisons, assimilations et autres phénomènes de coarticulation, pauses et respirations, hésitations et disfluences sont autant de perturbations d'une modélisation acoustico-phonémique recherchée. Afin d'obtenir une idée quantitativement plus précise des variantes majeures dans la parole, on peut faire appel à l'alignement automatique (Adda-Decker et Lamel, 2000 ; Boula de Mareüil et Adda-Decker, 2002). En particulier pour les variantes concernant les voyelles à aperture moyenne, on peut commencer par expliciter les variantes ouvertes et fermées dans le dictionnaire de prononciation. L'alignement permettra de faire émerger des tendances, même avec des modèles acoustico-phonémiques « bruités ». Surtout, les comparaisons contrastives entre différents types de corpus (style, région ou autres variables) sont éclairantes, dans la mesure où l'on peut interpréter des variations, plutôt que des mesures absolues (Adda-Decker et Lamel, 1999).

4. Réflexions méthodologiques sur l'approche automatique

Nous abordons ici quelques questions importantes concernant l'utilisation de l'alignement automatique dans les études linguistiques, afin que le lecteur puisse se faire une idée de la validité et des possibilités d'exploitation de telles ressources, désormais produites en grande quantité.

4.1. *Alignement phonémique ou phonétique ?*

Il est important de prendre en considération le fait que la précision des transcriptions phonétique et phonémique, ainsi que de la segmentation associée dépend de la configuration du système d'alignement utilisé (Adda-Decker, 2007). En effet, nous savons qu'elle dépend des réglages de notre instrument de mesure, en particulier de la résolution temporelle induite par le calcul des paramètres acoustiques (ici 10 ms), des prononciations prévues, des modèles acoustiques utilisés, de certains réglages concernant les durées minimale et maximale, de la gestion des silences, pauses ou autres phénomènes extra-lexicaux. Et comme pour tout instrument complexe, ces réglages ainsi que l'interprétation des résultats obtenus, peuvent nécessiter une expertise importante. Un dictionnaire de prononciation incluant très peu de variantes (par exemple moins de deux variantes par mot) fournira un étiquetage plutôt phonémique, et les mesures acoustiques associées pourront alors donner des indices sur la variabilité phonétique du phonème en question. À l'inverse, un dictionnaire de prononciation avec un fort nombre de variantes appropriées, fournira un étiquetage plutôt phonétique : c'est l'étiquetage qui reflétera la variation observée dans l'audio.

Le dictionnaire de prononciation pourra, par exemple, permettre la présence/absence d'un phonème (variante séquentielle) (Adda-Decker et Lamel, 1999 ; Bürki *et al.*, 2007), mais également l'alternance (variante parallèle) entre les voyelles semi-ouvertes (e/ɛ) et (o/ɔ) souvent génératrices de variantes libres ([vɔlœʁ] / [volœʁ]). Dans (Adda-Decker et Lamel, 1999), de nombreuses expériences décrivant des alignements avec des variantes parallèles ou séquentielles et utilisant des modèles acoustiques indépendants et dépendants du contexte sont décrites. Ces expériences ont permis de mettre clairement en évidence qu'avec des modèles dépendants du contexte, l'alignement automatique a moins recours à des variantes qu'avec des modèles acoustiques indépendants du contexte. Ces modèles dépendants du contexte modélisent donc implicitement un grand nombre de variantes. Ils sont également reconnus pour être moins précis dans la segmentation phonémique (voir Bürki *et al.*, ce volume). Dans quelle mesure la précision de la segmentation peut influer sur les résultats d'analyses phonétiques est une question à laquelle nous devrons répondre.

4.2. *Précision de la segmentation : un problème spécifique au TAP ?*

L'intérêt que le linguiste phonéticien doit porter à la précision de la segmentation automatique dépend du type d'analyses qu'il souhaite effectuer. Si l'on ne s'intéresse qu'à des phénomènes globaux, tels que la prosodie de l'énoncé, par exemple, la précision de la segmentation s'avérera de moindre importance. Dans cette partie nous nous consacrerons à l'analyse phonémique : l'analyse au niveau du phonème est bien sûr plus sensible à la précision de la segmentation, même si les mesures dans les parties centrales y seront moins sensibles que des mesures concernant les frontières et les zones de transition.

La précision des alignements automatiques, entendue en termes d'écarts avec une transcription manuelle de référence n'a pas été faite ici, mais elle a été évaluée à plusieurs reprises dans la littérature. Des mesures du décalage entre frontières manuelles et automatiques sont, par exemple, fournies par (Auran et Bouzon, 2003), qui obtiennent une fiabilité de 70 % pour un seuil d'acceptabilité de 20 ms. (Nguyen et Espesser, 2004) précisent que la durée attribuée aux voyelles par l'outil d'alignement utilisé dans leur étude, est généralement plus courte, avec une précision moins importante en fin de voyelle. Malgré tout, le milieu de la voyelle est correctement localisé dans 75 % des cas avec une tolérance de 20 ms. Leurs résultats précisent également que les écarts entre les alignements manuel et automatique ne concernent pas de manière univoque tous les phonèmes ni tous les contextes phonémiques. À défaut d'avoir fait des comparaisons entre segmentations manuelle et automatique ici, nous proposons de comparer les mesures issues de différents systèmes d'alignement. Si les mêmes tendances émergent, la précision de la segmentation ne semble pas jouer un rôle crucial. Nous varions les systèmes d'alignement à la fois suivant leur laboratoire de provenance (IRISA, LIA, LIMSI), et suivant leur configuration (modèles acoustiques indépendant ou dépendant du contexte).

Comme précisé précédemment, les modèles acoustiques tiennent compte de façon implicite des facteurs de variation de la parole. La durée vocalique en fonction de la position de la syllabe dans le mot, par exemple, ou bien les valeurs des formants des voyelles en fonction de leur durée sont des facteurs importants de variation, et seront présentées dans cette section. D'après les résultats de la littérature présentés ci-dessus, la précision de la segmentation est donc toute relative dès lors que l'on s'applique à considérer les frontières de phonèmes, et elle serait susceptible de se dégrader pour certains phénomènes plus fins, tels que les transitions de voyelle à consonne, ou bien des cas extrêmes tels que les voyelles fortement réduites. Cependant, ce problème ne s'applique-t-il pas également dans le cadre d'une segmentation manuelle ? Les frontières difficiles à placer par les machines (pour les sonantes, les semi-voyelles ou le schwa, par exemple) sont également sources de difficultés pour les humains. L'alignement manuel peut être taxé de subjectivité, puisque fréquemment réalisé par l'auteur de l'étude qui découle de ce corpus. Si un corpus est aligné manuellement pour l'investigation des variantes de prononciation, par exemple, il est important de comprendre dans quelle mesure et comment les choix du transcripteur contribuent à la variation mise en évidence. Il est plus que probable qu'une machine ne sera pas sujette à cette subjectivité inconsciente du segmenteur humain. De même, si les alignements automatiques peuvent être suspectés d'être erronés ou imprécis dans certains cas, et peut-être de façon plus systématique pour certains phonèmes ou certains contextes que pour d'autres, le phonéticien qui segmentera manuellement n'est pas exempt non plus de ces suspicions. Il est par ailleurs reconnu que la segmentation humaine peut manquer de constance au cours des travaux de segmentation s'étalant sur une longue durée.

Il est à noter que la précision de la transcription et de la segmentation n'est pas uniforme sur l'ensemble d'un corpus. Pour des passages de parole dont le débit s'accélère, par exemple, la segmentation qui impose une durée phonémique à 30 ms au minimum (par la présence de trois états HMM) ne pourra pas saisir les voyelles les

plus courtes. Par opposition, les passages de parole plus soutenue et préparée, dont le rythme reste posé, verront la qualité de leur segmentation accrue. Si l'on savait détecter des segmentations imprécises, il ne serait pas pour autant forcément opportun de s'en débarrasser sans précautions particulières. En effet, ces cas problématiques peuvent être porteurs d'informations linguistiques pertinentes, à plus forte raison s'ils concernent certains phénomènes (comme les réductions vocaliques « extrêmes ») ou certains contextes segmentaux uniquement. L'éradication de ces cas limites est ainsi susceptible d'orienter sensiblement les résultats (Bürki et Gendrot, 2007).

4.3. *Des analyses automatiques combinées à un alignement automatique*

Des mesures et analyses automatiques sont tout aussi utiles que la segmentation et l'alignement automatiques, dès lors que de très grandes quantités de données sont prises en compte. Sur le même principe, la validité des mesures automatiques peut être mise en cause. Pour l'analyse phonémique, et plus particulièrement l'analyse des voyelles orales, les mesures qui nous intéressent sont la durée, la fréquence fondamentale et les formants. 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 autre solution peut consister à rejeter des mesures aberrantes par l'établissement de filtres pour les valeurs de formants ou de f_0 établis sur la base de connaissances acoustiques ou bien de vérifications visuelles, telles qu'effectuées par (Gendrot et Adda-Decker, 2004, 2005, 2006) et (Boula de Mareüil *et al.*, ce volume). Les mesures automatiques fournies par le biais de logiciels libres tels que Praat (http ://www.fon.hum.uva.nl/praat) ou la librairie Snack Sound Toolkit (http ://www.speech.kth.se/snack/) 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/ est une voyelle mieux perçue par le rapprochement des troisième et quatrième formants, augmentant respectivement leur amplitude d'un point de vue acoustique ; ceci peut contribuer à favoriser la non-détection du deuxième formant. Il en va de même pour les deux premiers formants de /u/, ou bien pour les mesures de f_0 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. Ce réglage a permis sur des données de parole radiophonique de réduire les taux d'erreur de détection de 45 % à 19 % (Gendrot et Adda-Decker, 2005).

Pour les quatre figures ci-après (voir figure 4), 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 un tiers et deux tiers 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 critique 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 (Bürki et al. ce volume) ont 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 (à l'instar de (Gendrot et Adda-Decker, 2005)) et indiquent des valeurs relativement proches : LIA (5,5 %), IRISA monophones (4,8 %), IRISA triphones (5,5 %) et LIMSI (4,1 %).

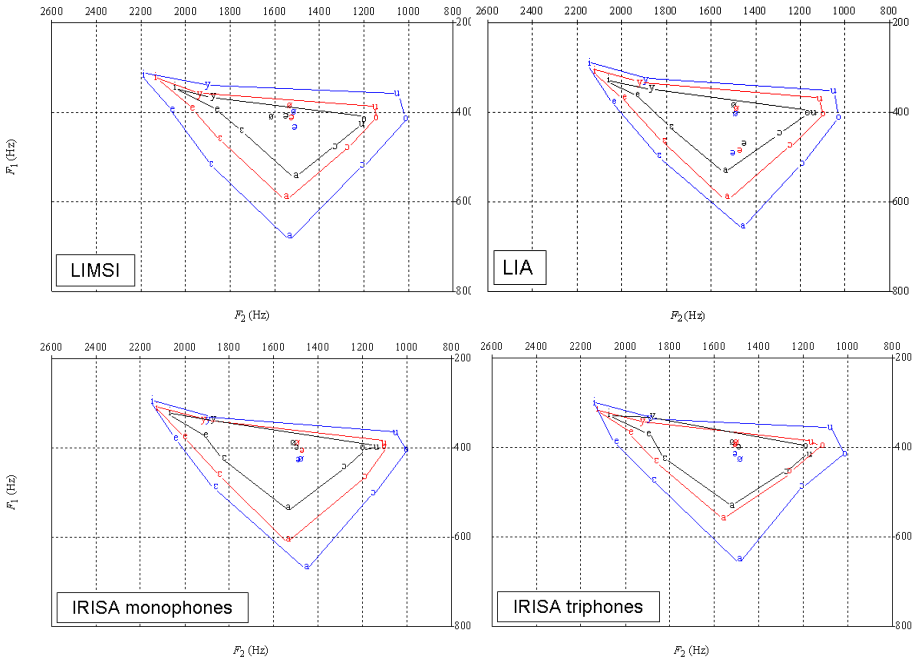

Figure 4. *Valeurs moyennes de F_1 et F_2 des voyelles orales du français en fonction de la segmentation utilisée et de la durée vocalique. De l'intérieur à l'extérieur segments de durée (en ms) : [30 - 50], [60 - 80], [90 - 110]*

D'autres analyses avec une résolution temporelle plus fine, telles que celle présentée ci-dessous (voir figure 5) avec des mesures prises au début de chaque voyelle (à un tiers du début, l'étiquetage choisi est IRISA monophones) en conservant les mêmes données et sans modifier le taux de rejet, révèlent des résultats beaucoup moins nets. En effet, on peut observer un nombre important de chevauchements entre voyelles (fermées) pour les différentes catégories de durée observées. Il est délicat de dire dans quelle mesure ces résultats traduisent la variation inhérente aux phénomènes de coarticulation, ou bien s'ils sont dus à un manque de précision soit de la segmentation, soit

de la détection automatique de formants. Ainsi, il semble difficile à l'heure actuelle d'utiliser de telles mesures sans précautions supplémentaires. En effet, si les mesures de formants sont prélevées à la périphérie plutôt qu'au centre, la coarticulation joue un rôle plus important, et ceci peut expliquer les « perturbations » des triangles vocaliques de la figure 5. Pour ce type de mesures excentrées, il faut séparer les segments suivant leur contexte phonémique gauche (ou droit pour des mesures sur le dernier tiers du segment).

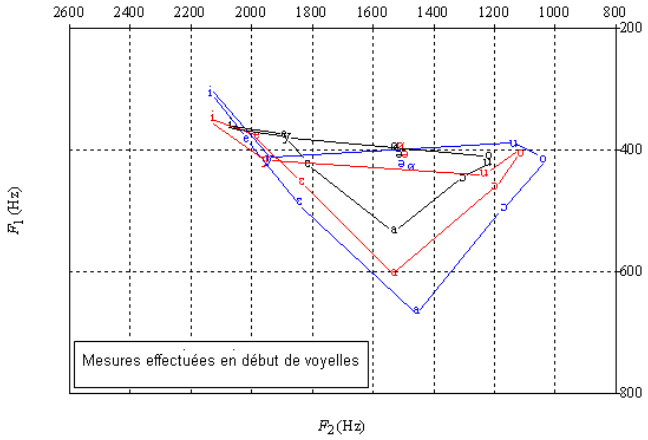

Figure 5. *Illustration de l'impact de la coarticulation. Valeurs moyennes de F_1 et F_2 des voyelles orales du français en fonction de la durée (catégories de durée identiques à la figure précédente), avec mesures effectuées au premier tiers de la voyelle*

4.4. *La taille des corpus : une façon différente d'aborder les données.*

Comme nous avons tenté de le montrer ci-dessus, les analyses de très grands corpus permettent de révéler des tendances nettes, les éventuelles imprécisions ou erreurs d'alignement phonémique étant compensées par la quantité importante des données. En effet, dans le cadre du traitement automatique de la parole, les données audio alignées sont disponibles en très grande quantité. Elles proviennent généralement de situations de parole contrôlées (par exemple, les journaux télévisés), en revanche le contenu lexical ou syntaxique n'est pas particulièrement contrôlé, comme ceci est souvent le cas pour des travaux sur corpus en linguistique. Nous n'excluons pas les cas où un alignement ainsi qu'une analyse automatique seraient appliqués à un corpus contrôlé dans cette dernière acception du terme, mais nous ne considérerons pas cet aspect dans cette section.

Afin d'analyser un flux de parole continue, il faut prendre en compte au mieux l'ensemble des phénomènes de variabilité, la méthodologie étant fondamentalement

inversée en comparaison des corpus linguistiques construits *ad hoc*. Non seulement la taille des corpus doit être considérée, mais la représentativité de certains phénomènes et/ou de certains contextes ne doit pas être négligée. Notons, par exemple, la très grande quantité de mots-outils dans un corpus de parole continue qui diffèrent considérablement des mots lexicaux dans leur réalisation, par leurs voyelles hypoarticulées, ou plus spécifiquement pour le cas de la voyelle /y/ dont les valeurs de formants, ou d'autres paramètres d'analyse, seront fortement altérées par la fréquence d'utilisation du pronom « tu », en parole conversationnelle, alors que la fréquence de ce même pronom est extrêmement rare en parole journalistique. De nombreux exemples peuvent être cités, mentionnons-en un qui touche à la phonotactique : la présence d'un /œ/ en français est très fréquemment suivie d'un /R/ (`bonheur`, `leur`, `heure`), ce qui modifiera considérablement sa réalisation moyenne. Ces différences de distribution doivent être prises en compte dans les analyses. Certes, elles sont intéressantes, puisqu'elles font partie des caractéristiques de la langue, mais elles influencent également la réalisation des sons concernés de manière significative (notons également le contexte alvéolaire majoritaire en parole continue, la fréquence lexicale, le voisinage phonologique et l'ordre d'apparition).

Les figures présentées ci-dessous (voir figure 6) avec des mesures de formants moyennes (l'étiquetage choisi est IRISA monophones) effectuées pour une étude sur le schwa (Fougeron *et al.*, 2007), révèlent pour des schwas internes de mots l'importance de prendre en compte la distribution du contexte phonémique. En effet, sans prendre en compte un nombre équilibré de contextes, les mots commençant par `re-` (`recommencer`, `refaire`, etc.) sont extrêmement fréquents, ce contexte précédant induisant ainsi une réalisation acoustique très différente. Les résultats présentés ici indiquent des valeurs de formants très différentes de celles des contextes équilibrés à cause de la phonotactique.

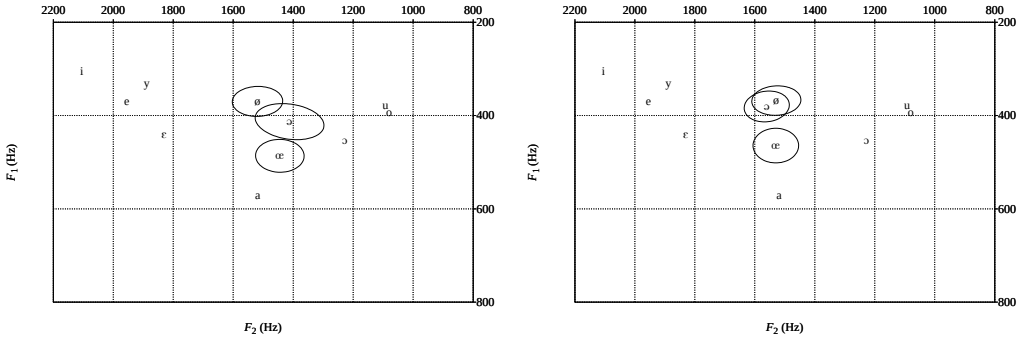

Figure 6. *Mesures de formants pour les voyelles centrales du français. Les résultats diffèrent nettement selon que les contextes sont rééquilibrés (à droite) ou non (à gauche). Les voyelles périphériques (sans ellipses) sont positionnées à titre indicatif*

Dans une étude linguistique d'après un corpus construit *ad hoc*, il est fréquent de considérer, pour le cas des voyelles, les voyelles produites en isolation, ou bien dans un contexte dit « neutre » pour produire des valeurs de référence. Les valeurs moyennes ici représentent les productions les plus communes dans un corpus de parole continue et pourraient modifier fondamentalement la notion de référence.

Dans des contextes très précis, par exemple pour une voyelle longue dans un contexte gauche/droit labial, le nombre de données peut n'apparaître plus suffisamment important, pour effectuer des comparaisons fiables. Si l'on analyse des tendances générales dans un premier temps, puis que l'on souhaite affiner les résultats, ou bien s'il apparaît nécessaire de contrôler tous les contextes (phonémique, syllabique, etc.) comme cela est réalisé dans un corpus « contrôlé », les occurrences s'y trouveront en nombre beaucoup plus restreint, et surtout dans des contextes déséquilibrés (Fougeron *et al.*, 2007). À défaut de corpus toujours plus grands, seul des corpus contrôlés permettront de vérifier, dans certains cas, des hypothèses linguistiques très précises.

5. Mesures objectives

5.1. *Durées segmentales en fonction du système*

Afin de mesurer l'influence de l'instrument de mesure sur la segmentation obtenue, nous avons utilisé les segmentations produites par différents systèmes d'alignement, développés dans différents laboratoires de recherches (LIA, IRISA et LIMSI) pour le corpus ESTER ; ceux-ci sont détaillés de manière exhaustive par (Bürki et al., ce volume). La figure 7, établie à partir de ces différents alignements, signale 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. Ces tendances sont identiques pour tous les systèmes, bien que les mesure 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. Notons que ces résultats fournis pour les mots trisyllabiques ne correspondent pas à ce qui est prédit par la littérature. Notamment, nous pourrions nous attendre à ce que la syllabe initiale, fréquemment porteuse d'un accent initial (que cet accent soit purement rythmique ou qualifié de « journalistique »), soit légèrement plus longue que les syllabes intermédiaires, ce que nous pouvons vérifier pour les mots quadrisyllabiques. En effet, la syllabe initiale est légèrement plus longue que la première syllabe intermédiaire (deuxième syllabe sur les quatre ; la tendance est moins nette pour le système IRISA monophones malgré tout). Cependant, la syllabe pénultième est systématiquement plus longue que la syllabe initiale. Ces observations invitent à repréciser les résultats mentionnés dans la littérature sur le français, puisque le rapport entre syllabes initiales et syllabes internes de mots semble fortement dépendre du nombre de syllabes dans le mot. La forte cohérence

entre les différents systèmes considérés permet également d'accroître la confiance apportée vis-à-vis des observations présentées ici.

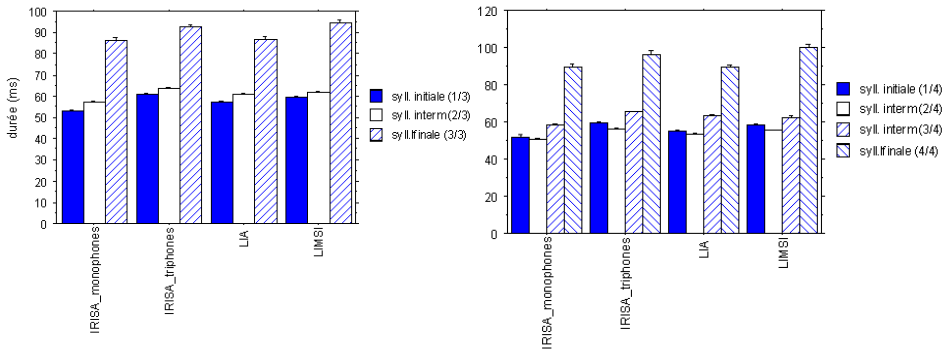

Figure 7. *Variations de la durée moyenne des voyelles en fonction de leur position syllabique dans le mot pour différents systèmes d'alignement. À gauche : mots trisyllabiques ; à droite : mots quadrisyllabiques*

5.2. *Durées en fonction du style*

Dans cette partie nous utilisons l'alignement automatique pour quantifier les différences entre styles de parole en termes de durées segmentales. Nous avons utilisé un corpus de données journalistiques (ESTER), un corpus de conversations téléphoniques de 60 heures et une trentaine d'heures du corpus PFC, correspondant approximativement pour un tiers à de la lecture de texte, un tiers d'entretiens guidés et un tiers de conversations libres.

La figure 8 montre l'effet du style de corpus. Pour le corpus journalistique et pour la partie de PFC correspondant à la lecture de texte, les distributions de durées réalisent leur pic à 60 ms globalisant respectivement 14 % et 12 % des corpus, la courbe de la lecture étant moins ramassée autour de ce pic que celle des données journalistiques. Pour la parole conversationnelle, le « pic » en question, plus étalé, se trouve plutôt autour de 50 ms avec seulement 10 % des données. En fait, le maximum des segments est concentré sur les durées courtes avec environ 20 % des segments sur la durée minimale de 30 ms. Ce maximum peut être vu comme le repli de tous les « segments » dont la durée aurait dû être inférieure ou égale à 30 ms sur ce point. Pour la parole journalistique et la lecture, ce taux reste inférieur à 10%.

Cette simple analyse des durées segmentales permet de mettre en évidence des différences de débit et de rythme de parole, entraînant pour la parole spontanée un débit plus variable, qui se traduit par un certain étalement de la distribution des du-

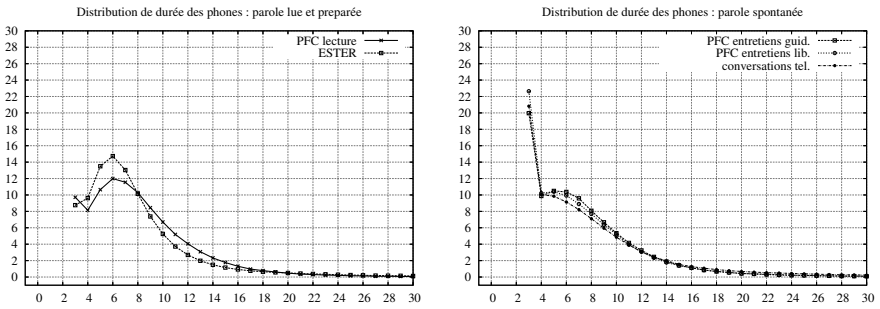

Figure 8. *Impact des styles de parole sur la distribution des durées segmentales (axe X : durée segmentale en cs ; axe Y : pourcentage des segments) pour différents corpus. À gauche : lecture (PFC - texte) et parole préparée (ESTER - émissions journalistiques). À droite : parole spontanée (PFC - entretiens libres et guidés ; conversations téléphoniques)*

rées, ainsi que des réductions temporelles des prononciations, matérialisées, dans nos mesures, par le pic autour de 20 % de segments de durée minimale. Il est intéressant de noter que ce pic est observé pour différents corpus de parole spontanée : 10 heures de PFC entretiens libres, 14 heures de PFC entretiens guidés, 60 heures de conversations téléphoniques. Comment faut-il interpréter ce pic étant donné notre instrument de mesure ? Il faut garder à l'esprit que l'alignement reposait ici sur des prononciations phonémiques (canoniques maximales) et tous les phonèmes prévus ont donc été alignés. Si un mot est articulé de manière non canonique avec des réductions temporelles (e.g. `v'là` pour `voilà`), l'alignement produira des segments de durée minimale dans la zone où a eu lieu la réduction (dans notre exemple dans la zone des phonèmes /w/ et /a/). Les segments de durée minimale pointent ainsi sur des zones avec des variantes de prononciation, entraînant un raccourcissement temporel. Les mesures ne permettent pas de conclure si les segments ont été effectivement réalisés avec une durée faible ou simplement omis, ou encore s'il y a d'autres phénomènes incluant, par exemple, des restructurations syllabiques. Une interprétation plus précise nécessite dans le futur des investigations approfondies en lien avec des études sur les variantes phonologiques du français, la prosodie, le débit et la modélisation des prononciations pour le traitement automatique. En particulier, des études en fonction des contextes phonémiques et des mots supports permettront de montrer si la réduction temporelle observée est plutôt conditionnée par les contextes phonémiques, par la fréquence phonémique ou lexicale, ou bien par la structure syllabique et le rang syllabique dans le mot. Nous relevons simplement ici quelques exemples typiques observés, de manière relativement peu fréquente dans les corpus journalistiques – `ça va être surtout` (prononcé approximativement comme [saɛtsyʁtu]]) –, et de manière récurrente en parole spontanée – `je crois que`, `moi je suis`, `je lui ai dit` –, exemples pour

lesquels nous laissons au lecteur le soin de se rappeler ou d'imaginer des prononciations couramment entendues. On peut se poser la question de savoir si tous les phonèmes sont impliqués de la même manière. On peut, par exemple, s'attendre à ce que le schwa ait une proportion élevée de durées minimales, dès lors que le schwa était obligatoire lors des alignements exploités, en particulier pour les mots-outils monosyllabiques très fréquents comme `le`, `de...` Les alignements montrent que tous les phonèmes n'y contribuent pas de manière égale. Le tableau 2 montre les phonèmes réalisant les plus forts pourcentages de durées minimales dans le corpus de conversations téléphoniques. Des taux plus faibles sont observés pour le corpus journalistique, mais l'ordre des phonèmes impliqués reste sensiblement identique. Concernant les consonnes, on trouve en premier lieu le /l/ avec 43 % des segments à durée minimale, qui semble être une consonne très instable. Le /l/ est très fréquent en français, en particulier sur les mots-outils [4], et contribue souvent aux clusters consonantiques (`plus`, `triple`). De manière générale les liquides, les semi-voyelles, le /v/, le /d/ et le /n/ sont les consonnes les plus affectées. Le /d/ comme le /l/ servent fréquemment de support aux mots-outils. Pour les voyelles, on voit que le schwa arrive en première position avec pratiquement 50 % des occurrences. Ceci confirme, par mesures détournées, la nature instable du schwa. Ce qui est plus intriguant, c'est de voir des pourcentages élevés pour les voyelles ouvertes /ɛ/ et /a/. Concernant le /a/, nous pouvons observer, par exemple, que ce segment peut quasiment disparaître dans des mots ou séquences de mots comme `voilà`, `parce que` réalisés plutôt comme `v'là`, `p'ce que`. De même, les voyelles antérieures fermées sont alignées fréquemment avec une durée minimale. On peut remarquer que bon nombre de phonèmes qui ont parmi les plus forts pourcentages de durées minimales, sont également parmi les plus fréquents. La fréquence n'est cependant pas le seul critère, car, dans le tableau, il manque les consonnes non voisées comme /t/ et /s/, qui ont une durée intrinsèque plus longue.

Consonne	*% durées min.*		*voyelle*	*% durées min.*	
Phonème	*homme*	*femme*	*phonème*	*homme*	*femme*
/l/	43	44	/ə/	50	46
/ɳ/	34	43	/ɛ/	30	24
/v/	33	26	/a/	26	16
/j/	28	21	/i/	25	20
/ʁ/	26	23	/y/	24	27
/d/	23	17	/e/	24	16
/n/	20	12	/ø/	22	21

Tableau 2. *Phonèmes réalisant les plus forts pourcentages de durées minimales dans le corpus de conversations téléphoniques. Ces pourcentages sont donnés pour les voix d'hommes et les voix de femmes*

4. De nombreuses erreurs de transcription automatique incluent la confusion de `il` avec `y` ou des contractions incluant plusieurs mots comme `c'est les` en `ses`.

5.3. *Allongements, hésitations et accent de mot*

La segmentation automatique peut servir à vérifier dans quelle proportion les syllabes finales en français sont allongées. Cet allongement peut alors être mis en relation avec l'accent de mot, connu pour frapper la syllabe finale en français, surtout en fin de groupe prosodique ou groupe de sens. Il faut cependant garder à l'esprit que l'allongement peut également traduire une élocution hésitante. Plutôt que d'insérer des `euh` d'hésitation dans la parole, le locuteur peut ralentir la cadence, en particulier à des endroits où ces hésitations apparaissent fréquemment, *i.e.* autour des articles, ou de manière générale, des mots-outils.

Pour synthétiser les résultats concernant les durées et allongements, nous adoptons une représentation en escalier, utilisée par (Delattre, 1965) pour sa comparaison de la place de l'accent de mot (dans le groupe de sens) à travers différentes langues. Les marches de cet escalier correspondent à des mots à N syllabes (avec $N \leq 5$), chaque syllabe représentant un carré. Dans une première approximation nous posons que l'allongement syllabique équivaut à un allongement vocalique. Si dans une position syllabique donnée on trouve un pourcentage élevé de voyelles longues, ce carré sera d'autant plus sombre que le pourcentage mesuré est élevé. Utilisant la même représentation, nous ajoutons l'information complémentaire des segments à durée courte (inférieure à 50 ms). Cette information nous intéresse dans l'objectif d'expliciter des règles génériques pour les variantes de prononciation et d'améliorer la modélisation acoustique des mots.

À partir de nos corpus segmentés, nous mesurons simplement le pourcentage de voyelles dont la durée dépasse 100 ms. Nous en trouvons globalement entre 25 et 30 % suivant les corpus (ESTER journaux ou PFC entretiens). Ces mesures ont été faites suivant deux axes : (i) globalement *vs* sous-ensemble de mots-outils ou mots pleins ; (ii) les mots sont suivis ou non d'une voyelle schwa (cette voyelle schwa n'est pas comptabilisée dans la représentation en escalier). Notre hypothèse ici est que la réalisation d'un schwa en fin de mot est peut-être un marqueur d'allongement en français standard. Cette hypothèse sera ainsi examinée. En effet, le rôle du schwa, ne contribuant pratiquement pas à la discrimination de paires minimales en français standard, est essentiellement prosodique. La partie gauche de la figure 9 illustre les mesures de voyelles allongées sur l'ensemble des mots (hors ceux se terminant par un schwa). Cette représentation appelle quelques commentaires :

- d'abord le nombre d'occurrences de mots diminue avec la longueur syllabique (suivant une diminution quasi logarithmique) avec plus de 72 k d'occurrences de mots à une syllabe et un peu moins de mille occurrences de mots pentasyllabiques ;
- on peut voir que, conformément au schéma de Delattre, les syllabes allongées sont nettement plus fréquentes en position finale que dans n'importe quelle autre position. Cependant toutes les positions finales ne sont pas allongées dans la parole continue ;

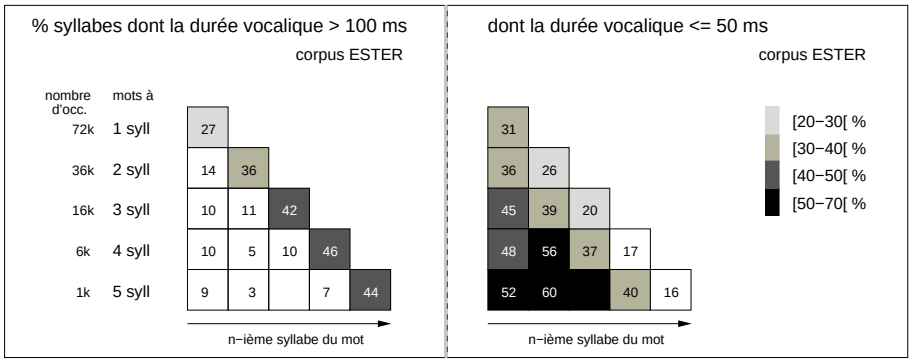

Figure 9. *Taux de segments vocaliques dont la durée est $>$ 100 ms (à gauche) et $\leq$ 50 ms (à droite). Les voyelles sont distinguées suivant leur rang syllabique dans le mot : première, deuxième, pénultième, dernière syllabe sur des mots de 1 à 5 syllabes*

– le taux de voyelles finales allongées augmente avec la longueur syllabique du mot. Ceci correspond aux attentes, pour deux raisons : le taux de mots-outils diminue dans la population des mots longs ; le nombre de fins de mots allongeables à l'intérieur d'un groupe prosodique d'une taille donnée est d'autant plus faible que les mots le composant sont longs ; ceci tend à augmenter la probabilité de voir ces fins de mot allongées ;
– le taux de voyelles allongées hors position finale tend à diminuer avec la longueur syllabique du mot ;
- le taux en position pénultième est semblable au taux en position initiale (*cf.* sous-section 5.1) ;
– les positions internes (hors pénultième, pour des mots d'au moins quatre syllabes) ont des taux particulièrement faibles.

Nos mesures ne nous permettent pas d'affirmer si la position initiale est mieux préservée en terme de durée que la position pénultième. La partie droite de la figure 9, avec les proportions de voyelles courtes ($\leq$ 50 ms) donne un éclairage complémentaire et permet d'apporter une réponse, au moins partielle, à cette question. Le schéma montre que le taux de voyelles courtes est particulièrement faible en position finale (diminuant de 31 à 16 % avec la longueur syllabique). Nous voyons que le taux de voyelles courtes en position pénultième (de 36 à 40 %) reste plus faible qu'en position initiale (de 36 à 52 %), les positions internes pouvant atteindre des taux jusqu'à 60 %. Ceci voudrait donc dire que les voyelles de syllabes pénultièmes se trouvent globalement moins raccourcies que les voyelles en syllabe initiale de mot.

Dans la figure 10, nous séparons les mots-outils (fonctionnels) – essentiellement des clitiques, mais également quelques mots lexicaux très fréquents – des mots pleins (lexicaux), afin de vérifier si l'allongement de la syllabe finale est observé davantage

sur les mots pleins que sur les mots-outils. On peut voir une différence nette entre les escaliers des mots pleins (à gauche) et ceux des mots-outils (à droite). Pour les mots

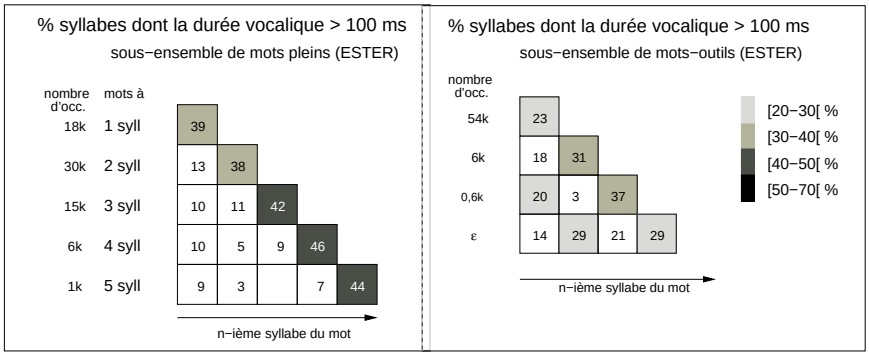

Figure 10. *Taux de segments vocaliques dont la durée est > 100 ms. Distinction entre mots pleins à gauche, mots-outils à droite*

pleins, les résultats changent un peu pour les mots mono- et bisyllabiques, dans la mesure où les mots-outils retirés affectent essentiellement ces catégories. De ce fait, le taux d'allongement final approche de 40 % pour les mots pleins courts. Les mots-outils (à droite), par effet de vases communiquants, obtiennent des taux plus faibles, avec, par exemple, 23 % d'allongements sur les mots monosyllabiques. Par ailleurs, pour les mots-outils, nous avons pu mesurer également des taux d'allongements relativement élevés en position initiale et en interne. Il faut cependant rappeler que le nombre d'occurrences de polysyllabes reste faible, empêchant des statistiques très fiables. Une interprétation des allongements observés sur les mots-outils consiste à dire qu'il peut s'agir d'allongements d'hésitation plutôt que d'accents de mot. Cette interprétation nécessite d'être vérifiée manuellement d'abord, et par un certain nombre de mesures complémentaires.

Nous revenons maintenant rapidement à notre hypothèse d'un schwa final comme éventuel marqueur d'une élocution allongée, soit globalement sur le mot, soit localement, du fait d'un segment schwa supplémentaire. Cependant notre critère de sélection (schwa final réalisé) est malheureusement relativement restrictif, au moins sur le corpus de parole journalistique. Ainsi, la figure 11 montre que le nombre d'occurrences de mots participant à cette analyse est faible. La réalisation ou non du schwa n'influe pas sur la représentation syllabique du mot dans les figures (par exemple le mot `force` est compté comme mot monosyllabique, même s'il y a un schwa final [fɔʁ.sə]). Les résultats des mesures pour les mots pleins et les mots-outils, par comparaison à ceux que nous venons de décrire (voir figure 10) évoluent de manière similaire. En effet, on peut voir, pour les mots pleins comme pour les mots-outils, que les taux d'allongement vocalique sont globalement plus élevés, à la fois sur la syllabe finale (proches de 50 %), sur la syllabe pénultième et en toute autre position. Ce résultat tend à montrer, au moins pour la parole journalistique, qu'il y a une corrélation importante entre la réalisation d'un schwa final et un allongement global des voyelles du mot en question.

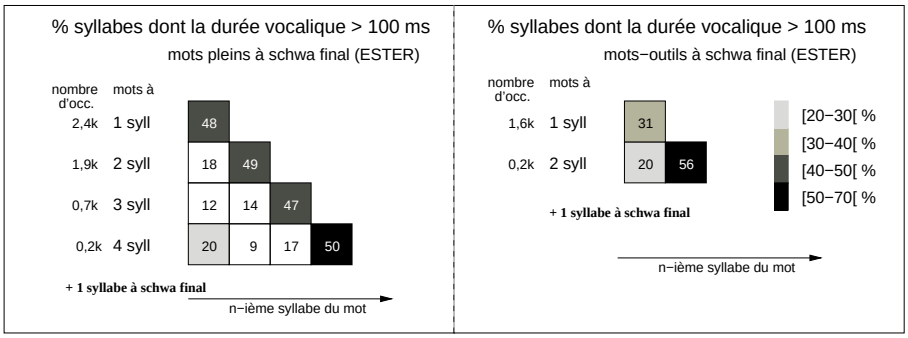

Figure 11. *Taux de segments vocaliques dont la durée est > 100 ms pour les mots réalisés avec un schwa final - mots pleins à gauche, mots-outils à droite*

Ce premier résultat nécessite des études supplémentaires sur plus de corpus différents. En particulier, une comparaison avec le français méridional peut être intéressante, dans la mesure où le schwa y a encore un rôle phonémique au niveau lexical.

Concernant les variétés régionales, nous nous contentons de comparer trois points d'enquête du corpus PFC : Brunoy en banlieue parisienne pour le français standard, Douzens (Languedoc) représentant d'une variété méridionale, et Nyon du canton de Vaud en Suisse romande pour un parler français de l'Est. Comme pour chaque point d'enquête, nous ne disposons que d'environ deux heures par point, nous faisons une analyse globale similaire à celle de la figure 9 (à gauche). Sans vouloir trop discuter le

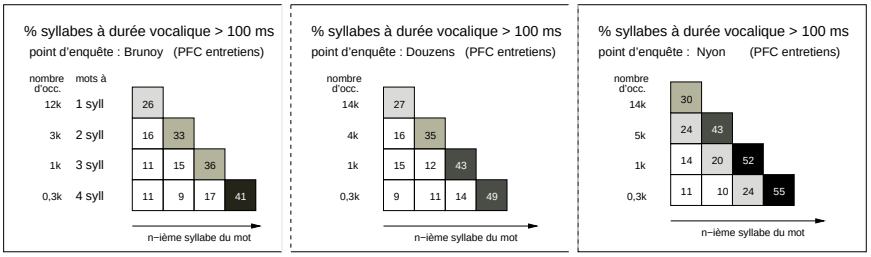

Figure 12. *Taux de segments vocaliques dont la durée est > 100 ms. Entretiens du corpus PFC : comparaison entre un point d'enquête du Nord (Brunoy, Île-de-France), du Sud (Douzens, Languedoc) et de l'Est (Nyon, Suisse romande)*

détail de ces mesures, nous pouvons observer que les entretiens spontanés produisent des taux d'allongement assez similaires à ceux de la parole journalistique. Nous pouvons cependant faire quelques remarques spécifiques, afin de guider des travaux futurs. Le taux d'allongement final est le plus faible pour le point du Nord, et le plus élevé pour l'Est. Ici, nous pouvons remarquer que la pénultième a également des taux d'allongement bien plus forts que les syllabes correspondantes du Nord ou du Sud.

5.4. *Fréquence fondamentale et formants*

Nous avons voulu mesurer une fréquence fondamentale intrinsèque pour les voyelles du français et examiner le lien entre hauteur et durée sur un grand volume de données. Ces mesures permettent en particulier de vérifier si le schwa se singularise dans ces mesures. En effet, le schwa est la seule voyelle du système français qui n'apparaît en général que comme noyau de syllabes inaccentuées. Les syllabes accentuées ont tendance à avoir une durée plus longue et une f_0 plus élevée.

Nous avons tiré profit des corpus alignés automatiquement pour mesurer la fréquence fondamentale des voyelles et en déduire des valeurs moyennes (qu'on peut considérer comme hauteur intrinsèque). La figure 13 montre pour les voyelles du français, la f_0 moyenne par type de voyelle et par genre du locuteur. Le même calcul de f_0 moyenne a ensuite été effectué sur le sous-ensemble de segments vocaliques ayant une durée supérieure à 100 ms (sous-ensemble à forte proportion de voyelles provenant de syllabes accentuées en première approximation). On peut observer une corrélation intéressante entre durée et f_0 pour toutes les voyelles, à l'exception du schwa. Pour le schwa, la durée n'a pas d'effet sur la hauteur moyenne. L'allongement du schwa évoque plutôt des phénomènes de pauses remplies (de type *euh*), pour lesquelles la f_0 reste en général basse par rapport à la parole active.

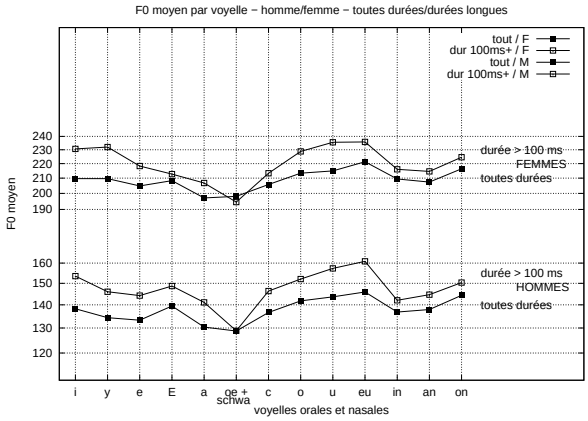

Figure 13. *Mesures de f_0 (fréquence fondamentale) moyenne par type de voyelle et par genre du locuteur. La moyenne est calculée sur l'ensemble des segments et sur le sous-ensemble des segments de longueur supérieure à 100 ms*

Nous avons pu mesurer, à partir de 70 heures de corpus journalistique, les formants des voyelles à partir d'une segmentation et d'un étiquetage phonémique automatiques (Gendrot et Adda-Decker, 2005). Les valeurs des formants, extraites automatiquement à l'aide de Praat, ont été calculées en fonction de durées décroissantes. La figure 14 montre un mouvement globalement centripète (centralisation) si la durée segmentale diminue. Trois catégories de durées ont été considérées :

$<$ 60 ms, entre 60 et 80 ms, $>$ 80 ms.
Ces résultats permettent à la fois de confirmer ceux connus sur la dépendance entre

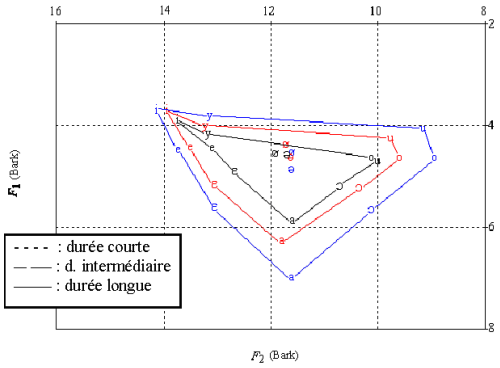

Figure 14. *Valeurs moyennes de F_1 et F_2 pour les voyelles orales du français en fonction de leur durée (normalisation en Bark) ; noir : durée courte ; rouge : durée intermédiaire ; bleu : durée longue*

durée et formants (Lindblom, 1963), et d'établir des valeurs de formants moyennes pour le français journalistique, qui peut être considéré comme un parler de référence.

Sur la base de ces résultats déjà proposés dans la littérature, mais développés sur un volume de données moins important, nous avons pu, dans un premier temps, montrer la validité de telles analyses basées sur l'alignement automatique. Dans un deuxième temps, nous avons également pu développer ces analyses sur des aspects jusqu'ici peu exploités, et ce pour l'ensemble des voyelles du français, comme par exemple la variation spectrale des voyelles en fonction de la position de la syllabe dans le mot ou bien en fonction de la position de la voyelle par rapport aux pauses (Gendrot et Adda-Decker, 2006). Sur le même principe, des comparaisons par rapport à d'autres langues ont pu être proposées (Gendrot et Adda-Decker, 2007a), mais aussi des variations spectrales des voyelles combinant le contexte phonémique et la durée de la voyelle (Gendrot et Adda-Decker, 2007b). Ces études ont permis de préciser certains aspects mentionnés par la littérature, notamment par l'utilisation d'un très grand nombre de locuteurs, montrant ainsi les tendances générales et permettant de gommer les stratégies individuelles développées par un nombre plus restreint de locuteurs et pouvant gêner la compréhension d'un phénomène linguistique.

6. Classification automatique de variantes

Dans cette section, nous nous servons du système d'alignement, et de ses capacités de classifieur, pour examiner la réalisation des voyelles mi-fermées, et plus particulièrement du /E/ en français journalistique (corpus ESTER), comme en français

conversationnel (entretiens PFC). Le but de ce travail est de montrer une nouvelle méthode pour explorer la variation segmentale, ne reposant sur aucune mesure acoustique explicite (formants, f_0, énergie, durée...), mais uniquement sur un alignement automatique, *via* des modèles acoustiques de mots proposant des variantes spécifiques. La variante retenue sera celle qui permet de maximiser la probabilité du modèle acoustique correspondant dans un cadre bayésien. La figure 15 schématise la modélisation acoustique des mots lors d'un alignement canonique (à gauche), *i.e.* obtenue à partir de la prononciation phonémique des mots. La partie droite illustre la modélisation

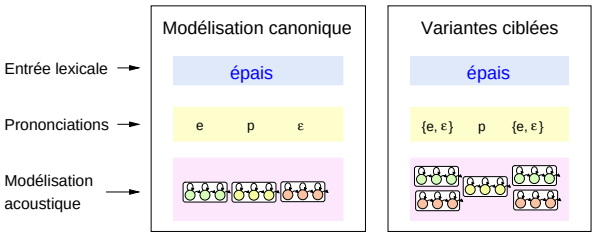

Figure 15. *Représentation schématique de la modélisation acoustique des mots pour le système d'alignement canonique (à gauche) et pour le système d'alignement à variantes ciblées pour le /E/ (à droite)*

acoustique des mots utilisée pour un alignement à variantes ciblées pour le /E/ (à droite). Dans cette dernière configuration toutes les voyelles /e/ (respectivement /ɛ/) des prononciations canoniques peuvent être remplacées par leur contrepartie à degré d'aperture plus ouvert (respectivement plus fermé). Cette dernière configuration du système d'alignement permet d'explorer la variation observée dans la zone concernée du triangle vocalique (voir figure 16) *via* les changements d'étiquettes phonétiques alignées par rapport à un alignement canonique.

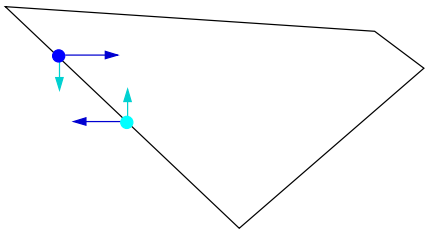

Figure 16. *Zone du triangle vocalique explorée via les variantes de prononciation ciblées pour le /E/*

Les données alignées sont ensuite comparées aux prononciations canoniques pour mesurer la variation *via* des taux de substitution. Par exemple, si une proportion importante de /ɛ/ est alignée comme [e], cela peut être interprété comme une tendance à

la fermeture et/ou à l'antériorisation. À l'inverse, un nombre élevé de /e/ canoniques étiquetés comme [ɛ], appelle une interprétation en faveur d'une ouverture et d'une centralisation. Ces mesures dépendent évidemment du dictionnaire de prononciation. Définir la prononciation canonique peut être simple pour un mot comme `idée` (/ide/). Pour le mot `aimer` la prononciation /ɛme/ peut déjà paraître un peu moins évidente, la prononciation [eme] étant majoritaire à l'oral dans l'usage courant. Quant à des mots comme `événement`, ils traduisent l'ambiguïté de prononciation jusque dans la graphie (variante d'écriture `évènement`). Avec cette mise en garde méthodologique en tête, il est clair que les résultats présentés sont informatifs surtout en relatif à travers les différentes configurations explorées, même si les mesures absolues sont intéressantes. Nous avons vérifié les /E/ du dictionnaire canonique pour les 3 000 mots les plus fréquents, qui permettent de tenir compte de 92 % des mots du corpus (de 100 heures). Le tableau 3 présente ces taux de substitution, en les détaillant suivant différents facteurs potentiellement intéressants : la fréquence du mot et la position du phonème dans le mot. Nous tenons à préciser ici que la position initiale/finale correspond bien au premier/dernier phonème du mot, et non à la première ou dernière voyelle du mot. Au niveau de la structure syllabique, la position initiale correspondrait donc pour le mot en question, à une syllabe à attaque vide et la position finale à une syllabe à coda vide. Globalement, le facteur fréquence entraînerait plutôt une tendance à la fermeture (en lien avec des durées plus courtes et des syllabes ouvertes pour les mots les plus fréquents). Au vu des résultats, on peut faire un certain nombre de remarques : le /e/ est relativement stable, ses taux de substitution restent proches de 10 % pour toutes les positions dans le mot (initiale, interne, finale), alors que le /ɛ/ présente des taux de substitution globaux autour de 20 %, avec des réalisations très différentes suivant la position dans le mot : relativement stable en position interne (8 %), très variable en frontière de mot (avec des taux autour de 45 %). Pour les dix mots les plus fréquents, les résultats traduisent le comportement des mots-outils comme `et` /e/, `les` /le(z)/, `des` /de(z)/, ainsi que `est` /ɛ(t)/. Ici, les voyelles internes correspondent au noyau d'une monosyllabe, la consonne de liaison fournissant le dernier phonème du mot. Par exemple, le /e/ dans `les` prononcé avec la liaison /lez/. Dans ce cas, seulement 5 % des /e/ sont alignés comme [ɛ] (240 sur 4 600). En revanche, dans la forme standard (sans liaison), 11 % des /e/ sont alignés comme [ɛ] (1 600 sur 14 300). Les résultats montrent que dans cette configuration (clitiques + /z/ de liaison), le /e/ reste particulièrement stable. Pour le /ɛ/, une partie importante de la variation mesurée, provient du verbe auxiliaire `est`, dont la prononciation est très largement réalisée comme [e], alors que sa forme canonique correspond à /ɛ/. Le taux de substitution très faible en position initiale pour les rangs de fréquence compris entre 11 et 100, provient des deux mots monosyllabes `elle`, `être`, où le /ɛ/ est le noyau de la syllabe fermée. Ceci suggère que le rang syllabique et/ou la structure de la syllabe (coda vide ou non) est un facteur important. Une analyse détaillée dans ce sens reste à faire. Les résultats montrent déjà que les syllabes ouvertes en position finale réalisent les taux de substitution les plus forts. En résumé, nous retenons que la position interne reste stable à la fois pour le /e/ et le /ɛ/, et qu'en revanche les frontières de mots sont particulièrement peu stables pour la voyelle ouverte.

Corpus journalistique					
Cas du /e/					
Rangs de fréquence	*glob.* (k)	*déb.* (k)	*int.* (k)	*fin* (k)	*Commentaires*
	Taux de substitution [ɛ]				
	% glob. (k)	*% déb.* (k)	*% int.* (k)	*% fin* (k)	
1-3 000	**11** **(124)**	**11** **(32)**	**10** **(41)**	**11** **(79)**	
1-10	9 (53)	10 (19)	5 (8)	10 (45)	`et`, `les`, `des`
11-100	10 (14)	14 (4)	5 (4)	12 (6)	/e/-interne, liaison
101-1 000	12 (38)	11 (7)	12 (17)	14 (15)	
1 001-3 000	13 (45)	12 (5)	13 (20)	13 (20)	
Cas du /ɛ/					
	Taux de substitution [e]				
	% glob. (k)	*% déb.* (k)	*% int.* (k)	*% fin* (k)	
1-3 000	**20** **(113)**	**47** **(24)**	**8** **(69)**	**44** **(33)**	
1-10	66 (14)	66 (14)	-	65 (11)	`est` prononcé [e]
11-100	15 (26)	6 (3)	8 (14)	27 (9)	`elle`, `être` : stable
101-1 000	12 (47)	22 (3)	6 (36)	35 (9)	
1 001-3 000	15 (40)	24 (4)	9 (30)	38 (6)	

Tableau 3. *Pourcentages de substitution du /E/ calculés sur la parole journalistique, pour des sous-ensembles de mots suivant leur fréquence, globalement et en fonction des positions (début, interne et fin) dans le mot. Pour chaque condition est précisé, entre parenthèses, le nombre total d'occurrences du phonème (en milliers (k))*

Nous nous intéressons ensuite à l'évolution de ces mesures sur de l'oral spontané et moins formel que la parole journalistique. Pour cela nous effectuons le même type d'analyse sur les entretiens (guidés et libres) d'une dizaine de points d'enquête du corpus PFC. Nous pouvons également examiner les taux en fonction de différentes variétés régionales, même si les quantités de données deviennent plus limitées ici. Le tableau 4 montre les résultats obtenus. Une première observation concerne les taux de substitution globaux, qui sont nettement plus élevés ici que pour la parole journalistique. En examinant ces taux en fonction de la place dans le mot, nous pouvons faire plusieurs constats : en position interne les taux restent comparables (proche de 30 %) entre nos deux voyelles /e/ et /ɛ/, taux qui étaient proches de 10 % pour la parole journalistique. Ensuite, on peut voir qu'en position de frontière de mot, le /e/ a des taux autour de 20 %. Ces taux ont certes doublé par rapport à la parole journalistique, mais ils restent cependant plus faibles qu'en position interne (30 %) et ils s'écartent très largement des taux observés en position de frontière de mot pour le /ɛ/. Ce dernier obtient des taux de substitution qui dépassent 60 % à la fois en début et en fin de mot. Pour la parole spontanée, de type plutôt vernaculaire, nos observations générales faites sur le corpus journalistique, restent valables : stabilité relative de la position interne ; en frontière de mot, stabilité pour le /e/ et grande instabilité pour le /ɛ/.

Entretiens PFC					
Cas du /e/					
	Taux de substitution [ɛ]				
	% global (k)	*% début* (k)	*% int.* (k)	*% fin* (k)	*Localisation*
Ensemble	**19** **(14)**	**18** **(4)**	**28** **(3)**	**17** **(10)**	
Brunoy	19 (1)	16 (0,4)	28 (0,2)	17 (0,7)	nord
Vendée	16 (1)	16 (0,2)	22 (0,2)	14 (0,6)	ouest
Lacaune	15 (1)	14 (0,3)	24 (0,1)	12 (0,6)	sud-ouest
Douzens	18 (2)	15 (0,4)	26 (0,3)	16 (1,2)	sud-ouest
Dijon	18 (1)	22 (0,4)	22 (0,3)	16 (0,9)	nord-est
Lyon	28 (1)	18 (0,3)	44 (0,2)	25 (0,7)	est
Nyon	30 (1)	44 (0,3)	47 (0,2)	23 (0,8)	est, Suisse
Cas du /ɛ/					
	Taux de substitution [e]				
	% global (k)	*% début* (k)	*% int.* (k)	*% fin* (k)	
Ensemble	**51** **(21)**	**62** **(6)**	**31** **(8)**	**65** **(12)**	
Brunoy	45 (2)	56 (0,5)	29 (0,6)	57 (0,8)	nord
Vendée	55 (1)	68 (0,4)	38 (0,4)	65 (0,7)	ouest
Lacaune	60 (1)	69 (0,3)	41 (0,5)	77 (0,6)	sud-ouest
Douzens	52 (2)	61 (0,7)	28 (0,9)	76 (1,1)	sud-ouest
Dijon	51 (2)	68 (0,5)	38 (0,8)	60 (1,2)	nord-est
Lyon	40 (2)	61 (0,6)	24 (0,6)	47 (1,1)	est
Nyon	35 (2)	48 (0,6)	20 (0,6)	45 (1,0)	est, Suisse

Tableau 4. *Pourcentages de substitution du /E/ sur le corpus PFC-entretiens, avec détails pour quelques points d'enquête, calculés globalement et en fonction des positions (début, interne et fin) dans le mot. Pour chaque condition est précisé, entre parenthèses, le nombre total d'occurrences du phonème (en milliers (k))*

En examinant les taux observés pour les différents points d'enquête, on voit globalement des comportements similaires pour le /e/, à part pour les points se situant à l'est (Lyon, Nyon), pour lesquels les taux de substitution sont plus élevés. Ceci indique une tendance à ouvrir la voyelle fermée /e/. En particulier le point en Suisse romande, se singularise par des taux de substitution particulièrement élevés en position initiale de mot. Par exemple, des mots comme `était`, `élection`, `écoute` ont été alignés majoritairement avec un [ɛ] en position initiale. Il s'agit ici d'un trait contribuant à identifier l'accent suisse (Woehrling *et al.*, 2008). De manière symétrique, pour le /ɛ/, les points de l'est obtiennent des taux de substitution plus faibles que les autres points du nord et du sud de la France. La forte tendance à la fermeture observée globalement en frontière de mot, est moins marquée pour le parler de l'est de la France et pour la Suisse. Concernant les points du sud (Lacaune, Douzens), ils se démarquent par des taux de substitution particulièrement élevés (dépassant 75 %) en position finale

de mot, alors que les autres points ont des taux de substitution qui restent plutôt plus faibles en position finale par rapport à la position initiale.

Les résultats de ce travail sur le /E/ mettent en évidence que la variation observée pour les voyelles antérieures mi-fermées est importante, et plus accentuée en parole spontanée qu'en élocution journalistique. Elle peut contribuer à caractériser des variétés régionales du français. Afin de mieux comprendre cette variation, et éventuellement la formaliser de manière plus claire, elle nécessite d'être approfondie en lien avec d'autres facteurs : l'harmonie vocalique, la loi de position, la structure syllabique, l'accent de mot et la prosodie.

7. Conclusion et perspectives

Un but de ce travail est de promouvoir les nouvelles technologies issues des recherches en traitement automatique de la parole, comme instruments d'analyse en phonétique et en phonologie. En effet, les systèmes de transcription automatique permettent de produire des corpus oraux annotés en quantité potentiellement illimitée, et ils peuvent être adaptés afin d'étudier des phénomènes précis, comme la réalisation des voyelles mi-ouvertes, de la liaison, du schwa et de l'assimilation, ou d'éclairer des questions d'ordre général, incluant la f_0 intrinsèque des voyelles, leurs formants en fonction de la durée segmentale. L'utilisation de grandes quantités de données permet de décliner les analyses suivant différents axes de variabilité, comme le rang syllabique de la voyelle dans le mot ou le groupe prosodique. Les instruments issus du traitement automatique de la parole, permettent à de nombreux champs de la linguistique de s'enraciner davantage dans les sciences expérimentales. Il paraît alors important de cultiver également, pour les domaines concernés, une rigueur expérimentale, combinant une connaissance de l'instrument et du matériau traité, avec une pratique objective de l'observation et de la mesure.

Nous avons commencé par motiver le travail présenté par des variantes de prononciation observées sur des mots (et groupes de mots), variantes qui dévient fortement d'une prononciation canonique. Ce phénomène de variantes réduites, particulièrement sensible en parole spontanée, pose de vrais défis à la reconnaissance automatique de la parole et met en cause la représentation phonologique adoptée pour la modélisation acoustique des mots. À partir de ces constats, nous avons d'abord développé des réflexions méthodologiques sur l'alignement automatique, incluant les aspects de segmentation et d'étiquetage, ainsi que sur les mesures de formants et de f_0. Grâce à l'utilisation d'alignements provenant de différents systèmes, nous avons pu vérifier l'indépendance des résultats moyens par rapport au système utilisé, ce qui renforce la validité de tels résultats obtenus *via* les instruments d'alignement automatique et de détection de formants. La stabilité des résultats a ainsi été montrée pour de grandes tendances comme l'effet centripète pour les formants en fonction d'une durée vocalique décroissante, les durées des syllabes initiales, internes, pénultièmes, finales. En revanche, les triangles vocaliques issus de mesures de formants prises au premier tiers des segments vocaliques présentent des perturbations liées aux effets de coarti-

culation et/ou à la précision de segmentation. Des investigations supplémentaires sont nécessaires ici, et le résultat montre que des précautions méthodologiques sont indispensables.

Les corpus utilisés totalisent environ 200 heures de parole journalistique, d'entretiens en face-à-face et de conversations téléphoniques. Le style de parole a un effet important sur la réalisation des mots et des segments les composant. Ceci a été mis en évidence simplement par des distributions de durée segmentale. Suivant le style, le pourcentage de segments de durée minimale varie de moins de 10 à plus de 20 %. Ces segments pointent sur des variantes de prononciation, dont la modélisation demande à être améliorée pour la reconnaissance automatique de la parole. Une meilleure connaissance de ces variantes permettra à la fois d'augmenter nos connaissances du fonctionnement de l'oral, et d'enrichir les représentations et modélisations pour le traitement automatique. Ainsi nous avons pu mesurer qu'en plus du schwa, d'autres voyelles comme le /i/, le /ɛ/ ou le /a/ étaient très souvent alignées avec une durée minimale en parole spontanée. Quelques consonnes voisées, comme le /l/ et le /v/, sont également très fréquemment mesurées avec une durée minimale.

La durée segmentale a ensuite été utilisée comme mesure potentiellement liée à l'accent de mot et, à un degré moindre, à l'hésitation. Des analyses de la durée vocalique en fonction de la position syllabique de la voyelle dans le mot ont permis d'abord de reproduire globalement les schémas d'accent de mot faits par Delattre avec des quantités de données plus réduites et une méthodologie peu explicite. Nos analyses confirment une corrélation entre durée et accent de mot et très probablement accent prosodique. Les mesures montrent une régularité statistique des phénomènes analysés, et donnent une quantification des réductions temporelles et des allongements, permettant d'ouvrir des perspectives nouvelles pour la modélisation acoustique des mots en traitement automatique. Nous avons pu mettre en évidence que la réalisation d'une voyelle schwa en fin de mot est le marqueur d'un allongement non seulement du mot en question (par ce segment schwa supplémentaire), mais également des voyelles précédant le schwa. La séparation du corpus en deux sous-ensembles de mots pleins et de mots-outils a permis de noter des différences intéressantes dans les durées vocaliques composant ces mots. Alors que les mots pleins reproduisent fidèlement les représentations de Delattre, les mots-outils montrent un fonctionnement un peu différent, peut-être en lien avec des allongements d'hésitation. Ces premiers résultats nous encouragent à développer davantage ces travaux, en tenant compte de groupes de mots ou autres entités prosodiques.

Enfin, nous avons consacré la dernière partie à l'utilisation de l'instrument d'alignement pour l'étude de la réalisation des voyelles mi-fermées /e/ et /ɛ/, en introduisant des variantes ciblées sur le /E/ dans le dictionnaire de prononciation. Les alignements résultants ont été analysés en tenant compte des facteurs suivants : fréquence lexicale, position de la voyelle dans le mot, structure syllabique, style de parole et variété régionale. Les résultats mettent en évidence que la variation observée pour les voyelles antérieures mi-fermées est importante, et plus accentuée en parole spontanée qu'en élocution journalistique. Elle peut contribuer à caractériser des variétés régio-

nales du français. Afin d'être mieux comprise, et d'être éventuellement formalisée de manière plus claire, cette variation nécessite d'être approfondie en lien avec d'autres facteurs : l'harmonie vocalique, la loi de position, la structure syllabique, l'accent de mot et la prosodie.

Le traitement automatique de la parole est appelé à jouer un rôle de plus en plus important dans l'étude de la variation de l'oral en fonction de divers paramètres. Il contribuera fortement à l'essor d'une phonologie de corpus, dont le souci est la confrontation des modèles avec des observations. L'apport du traitement automatique à la phonétique et à la phonologie n'est plus à démontrer. Notre conviction est qu'en retour, une meilleure formalisation des mécanismes à l'œuvre dans la variabilité de la parole contribuera en définitive à améliorer la modélisation des prononciations et les performances des systèmes de transcription automatique.

8. Bibliographie

Adda-Decker M., « Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole », *Actes des 7e rencontres jeunes chercheurs en parole*, Paris, p. 211-216, 5-6 juillet, 2007.

Adda-Decker M., Boula de Mareüil P., Adda G., Lamel L., « Investigating syllabic structures and their variation in spontaneous French », *Speech Communication*, vol. 46, p. 119-139, 2005.

Adda-Decker M., Lamel L., « Pronunciation variants across system configuration, language and speaking style », *Speech Communication*, vol. 29, p. 83-98, 1999.

Adda-Decker M., Lamel L., « Systèmes d'alignement automatique et études de variantes de prononciation », *Actes des XXIIIèmes Journées d'Étude sur la Parole*, Aussois, p. 189-192, 19-23 juin, 2000.

Auran C., Bouzon C., « Phonotactique prédictive et alignement automatique : application au corpus MARSEC et perspectives », *Travaux Interdisciplinaires du Laboratoire Parole et Langage*, vol. 22, p. 33-63, 2003.

Barras C., Geoffrois E., Wu Z., Liberman M., « Transcriber : development and use of a tool for assisting speech corpora production », *Speech Communication*, vol. 33, n° 1-2, p. 5-22, 2001.

Blanche-Benveniste C., « Constitution et exploitation d'un grand corpus », *Revue Française de linguistique appliquée*, vol. IV, n° 1, p. 65-74, 1999.

Boula de Mareüil P., Adda-Decker M., « Studying pronunciation variants in French by using alignment techniques », *Proceedings of Interspeech 2002*, 16-20 septembre, p. 2273-2276, 2002.

Bürki A., Fougeron C., Gendrot C., Frauenfelder U., « Chute du schwa en français : un processus sans ambiguïté ? », *Actes des 5e Journées d'Études Linguistiques*, Nantes, p. 83-88, 27-28 juin, 2007.

Bürki A., Gendrot C., « Reconnaissance automatique et analyse linguistique : l'exemple du schwa », *Actes des 7e Rencontres Jeunes Chercheurs en Parole*, Paris, p. 40-43, 5-6 juillet, 2007.

Delattre P., *Comparing the phonetic features of English, Spanish, German and French*, Julius Gross Verlag, Heidelberg, 1965.

Duez D., « Modelling aspects of reduction and assimilation in spontaneous French speech », *Proceedings of the IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, University of Tokyo, Tokyo, p. 120-124, 12-15 avril, 2003.

Durand J., Laks B., Lyche C., « Le projet *Phonologie du français contemporain* (PFC) », *La Tribune Internationale des Langues Vivantes*, vol. 33, p. 3-9, 2003.

Fougeron C., Gendrot C., Bürki A., « Le schwa : une voyelle comme les autres ? », *Actes des 5e Journées d'Études Linguistiques*, Nantes, p. 191-198, 27-28 juin, 2007.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-P., Gravier G., « The ESTER Phase II evaluation campaign for the rich transcription of French broadcast news », *Proceedings of Eurospeech-Interspeech*, Lisbonne, p. 1149-1152, septembre, 2005.

Gauvain J., Adda G., Adda-Decker M., Allauzen A., Gendner V., Lamel L., Schwenk H., « Where Are We in Transcribing French Broadcast News ? », *Proceedings of Eurospeech-Interspeech*, Lisbonne, p. 1665-1668, septembre, 2005a.

Gauvain J., Adda G., Lamel L., Lefèvre F., Schwenk H., « Transcription de la parole conversationnelle », *TAL*, 2005b.

Gendrot C., Adda-Decker M., « Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande », *Actes du colloque MIDL 2004*, Paris, p. 7-12, 29-30 novembre, 2004.

Gendrot C., Adda-Decker M., « Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in French and German », *Proceedings of Eurospeech 2005*, Lisbonne, Portugal, p. 2453-2456, 4-8 septembre, 2005.

Gendrot C., Adda-Decker M., « Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique », *Actes des XXVIe Journées d'Études sur la parole*, Dinard, p. 407-410, 12-16 juin, 2006.

Gendrot C., Adda-Decker M., « Impact of duration and vowel inventory size on formant values of oral vowels : an automated formant analysis from eight languages », *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saabrücken, p. 1417-1420, 6-10 août, 2007a.

Gendrot C., Adda-Decker M., « Influence of consonantal context and duration on F1/F2 centralization of oral vowels : an automatic analysis of large broadcast news corpora in French », *Proceedings of the Workshop on Coarticulation : Cues, direction, and representation*, Montpellier, p. 44-47, 7 décembre, 2007b.

Habert B., « Portrait de linguiste(s) à l'instrument », *Texto ! Textes et cultures*, 2005.

Lindblom B., « Spectrographic study of vowel reduction », *Journal of the Acoustical Society of America*, vol. 35, p. 1773-1781, 1963.

Meunier C., « Invariants et variabilité en phonétique », *in* N. Nguyen, S. Wauquier, J. Durand (eds), *Phonologie et phonétique : Forme et substance*, Hermès, Paris, p. 349-374, 2005.

Nguyen N., Espesser R., « Méthodes et outils pour l'analyse des systèmes vocaliques », *Bulletin Phonologie du français contemporain*, vol. 3, p. 77-85, 2004.

Troubetzkoy N., *Grundzüge der Phonologie / Principes de phonologie*, Vandenhoek & Ruprecht / Klincksieck, Göttingen / Paris, 1939-1976.

Woehrling C., Boula de Mareüil P., Adda-Decker M., « Aspects prosodiques du français parlé en Alsace, Belgique et Suisse », *Actes des XXVII^e^ Journées d'Études sur la parole*, Avignon, 9-13 juin, 2008.