Réduction des segments en français spontané :apports des grands corpus et du traitement automatique de la parole

Yaru Wu, Martine Adda-Decker

Réduction des segments en français spontané : apports des grands corpus et du traitement automatique de la parole

Yaru Wu[1,2], Martine Adda-Decker[1,2]
[1]Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
[2]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France}
yaru.wu@sorbonne-nouvelle.fr, madda@limsi.fr

Ce travail sur la réduction segmentale (c.-à-d. suppression ou réduction temporelle de segments) en français spontané nous a permis de proposer une méthode de recherche pour les études en linguistique, ainsi que d'apporter des connaissances sur la propension à la réduction des segments à l'oral. Cette méthode, appelée méthode ascendante, nous permet de travailler sans hypothèse spécifique sur la réduction. Les résultats suggèrent que les liquides, les glides et la fricative voisée /v/ sont plus facilement réduites que les autres consonnes et que les voyelles nasales résistent mieux à la réduction que les voyelles orales. Parmi les voyelles orales, les voyelles orales arrondies ont tendance à être plus souvent réduites que les autres voyelles orales.

Abstract

This study on segmental reduction (i.e. deletion or temporal reduction of segments) in spontaneous French allows us to propose a research method for linguistic studies on large corpora and to bring new insights on the propensity of segmental reduction. We applied the so-called bottom-up method while we do not have specific hypotheses. Results suggest that liquids, glides and /v/ fricative tend to be more often reduced than other consonants whereas nasal vowels are less prone to reduction than oral vowels. Among the latter ones, rounded oral vowels tend to be reduced more often than other oral vowels.

Mots-clés
réduction, élision, parole spontanée, grand corpus oraux, alignement forcé, segments courts

Keywords
reduction, elision, spontaneous speech, large speech corpora, forced alignment, short segments

1. Introduction

La variation de la parole est souvent observée en parole continue (Kohler, 1990 ; Duez, 1997 ; Ernestus, 2000 ; Johnson, 2004 ; Adda-Decker *et al*., 2005 ; Adda-Decker *et al.*, 2007 ; Dilley et Pitt, 2010 ; Meunier et Espesser, 2011 ; Nguyen et Adda-Decker, 2013 ; Meunier et Bigi, 2016). Manifestée par des changements phonétiques, la variation phonologique soulève des questions linguistique et extralinguistique intéressantes. Grâce aux travaux en reconnaissance automatique de la parole, nous avons à disposition de grands corpus de parole transcrite et ces corpus peuvent être exploités à des fins de recherche en phonétique et en linguistique de l'oral de manière plus générale. Avec les outils technologiques, nous avons aujourd'hui la possibilité d'étudier cette variation à grande échelle et d'examiner des phénomènes de réduction, qui ont été peu étudiés jusqu'à présent, à l'aide de grands corpus.

L'objectif de cette étude est d'examiner la variation de la parole à l'aide de nouvelles méthodes qui proviennent de la reconnaissance automatique de la parole, et d'apporter de nouvelles connaissances sur la variation des prononciations, et en particulier les phénomènes de réduction, en parole continue. Cette étude comporte deux volets : un volet méthodologique dans le but de répondre à la question « Comment étudier la variation de la parole naturelle à partir de grands corpus oraux ? » et un volet plus linguistique motivé par la question : « Qu'observons-nous dans le signal de la parole en ce qui concerne la réduction ? ».

2. Méthodologie

Dans cette section, nous présenterons la méthode de recherche qui servira à étudier différents phénomènes en linguistique en utilisant l'alignement automatique issu de la reconnaissance automatique de la parole. Nous allons également y résumer quelques détails techniques concernant cette méthode, ainsi que décrire le corpus utilisé pour notre étude.

2.1. Alignement forcé

La méthodologie utilisée dans notre étude repose sur l'alignement forcé entre le signal de parole et sa transcription manuelle. Lors de l'alignement forcé, la suite de mots à mettre en face du signal est connue, imposée (d'où le terme « forcé ») par opposition à la reconnaissance automatique où la suite de mots est inconnue et doit être déterminée par le système qui se sert dans ce cas-là du modèle de langue. L'alignement forcé permet de segmenter automatiquement le signal acoustique en mots et en phones[1] composant ces mots. Les frontières des segments[2] sont obtenus par le meilleur appariement entre le signal de parole et sa transcription à l'aide de modèles ou de références acoustiques correspondant à cette transcription et d'un algorithme de mise en correspondance.

La modélisation acoustique de la parole consiste à établir des représentations statistiques du signal sous forme de séquences de vecteur de paramètres – typiquement des paramètres MFCC[3] (Bridle *et al.*, 1974 ; Davis & Mermelstein, 1980) ou PLP[4] (Hermansky, 1990) – calculés à un pas régulier (en général toutes les 10 ms) à partir du signal acoustique. Ces représentations simulent la sensibilité de la perception humaine. Dans cette étude, nous ne parlerons pas des vecteurs issus des approches d'apprentissage profond (*deep learning*) qui ont permis de réaliser un saut qualitatif important dans la modélisation acoustique des systèmes de reconnaissance automatique de la parole depuis le début des années 2000 (Bengio, 2009 ; Lecun *et al.*, 2015). Pour les travaux impliquant l'alignement forcé, nous resterons avec le formalisme des modèles acoustiques de phones par les modèles de Markov cachés (Rabiner, 1989) qui a prévalu dans les systèmes de reconnaissance automatique à grand vocabulaire (~ 100 000 mots) autour des années 1990-2010. N'importe quel mot de la langue se trouve facilement modélisé d'un point de vue acoustique dès lors que sa prononciation est spécifiée dans le dictionnaire du système : il suffit de concaténer les modèles HMM de phones correspondant à cette prononciation. Un modèle HMM de phone comporte typiquement trois états pour rendre compte de l'évolution du son au cours du temps (début, milieu, fin) : le début est influencé par le contexte gauche (les sons précédents), la fin par le contexte droit (les sons suivants) et le milieu est considéré comme l'état stable le plus

[1] Réalisation d'un phonème.
[2] Dans la suite de l'article, nous utiliserons le terme « segment » pour désigner un segment phonétique, comme c'est l'usage en phonétique.
[3] Mel frequency cepstral coefficients
[4] Perceptual linear predictive

spécifique du son modélisé. Chaque état d'un HMM peut boucler sur lui-même. Cette boucle peut être vue comme un point d'orgue (en notation musicale) sur chaque état : un état dans le modèle peut correspondre à un seul vecteur ou une séquence plus ou moins longue de vecteurs dans le signal de parole à aligner, avant de le quitter pour l'état voisin. Ainsi, les modèles HMM combinés à l'algorithme de Viterbi (Forney, 1973 ; Rabiner, 1989) permettent de rendre compte des déformations temporelles dans la parole qui sont inévitables en raison des variations de débit et de rythme. L'instant de passage du dernier état d'un modèle de phone au premier état du modèle de phone suivant détermine la position de la frontière segmentale. De manière analogue, la frontière de mot est déterminée par l'instant de passage du dernier état du modèle du mot au premier état du mot suivant.

La figure 1 illustre l'implémentation des différents niveaux de modélisation impliqués pour aligner le mot « cinéma » sur le signal représenté sous forme de séquence de vecteurs acoustiques. Le mot en forme orthographique obtient une représentation phonémique via le dictionnaire de prononciation. Chaque symbole phonémique est associé à un modèle acoustique (un modèle HMM ou Hidden Markov Model à 3 états) qui synthétise les caractéristiques des sons telles que observées dans les corpus d'apprentissage. Lors de l'alignement, chaque état doit générer (ou absorber suivant le point de vue) au moins un vecteur acoustique, ce qui va correspondre à une durée minimale de 10 ms par état et donc à une durée minimale de 30 ms pour un modèle de phone. Les frontières de phone (de mot) correspondent aux endroits de passage d'un modèle de phone (de mot) à l'autre. Nous obtenons ainsi en sortie un signal segmenté en mots et en phones avec des informations sur la durée des mots et des phones. Il faut cependant garder à l'esprit que l'alignement forcé ne peut produire que des étiquetages en phones qui sont prévus par le dictionnaire de prononciation.

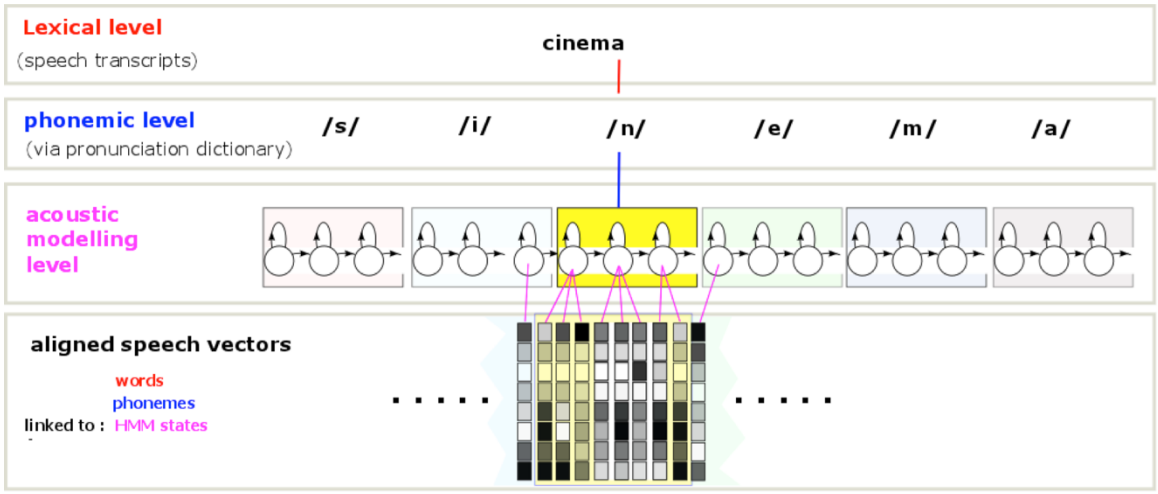

Figure 1. Illustration de la modélisation de la parole à travers les différents niveaux de représentation dans un système d'alignement automatique de la parole (Adda-Decker et Lamel, 2018).

L'alignement forcé et le dictionnaire de prononciation peuvent être utilisés pour étudier différentes hypothèses linguistiques et pour analyser de grands corpus (Adda-Decker *et al.*, 1999 ; Boula de Mareüil et Adda-Decker, 2002 ; Van Bael *et al.*, 2007 ; Schuppler *et al.*, 2014 ; Wu *et al.*, 2017 ; Tahon *et al.*, 2018). Avec cette méthode, l'absence ou la présence du segment en question est décidée automatiquement par l'alignement forcé. Même si la décision automatique de présence/absence ne possède pas la finesse d'une oreille phonétique experte, le fait de pouvoir exploiter facilement de grands corpus permet de dégager des tendances, si

possible en fonction de différents facteurs, comme le style de parole. Des études comparatives de ces tendances sont très instructives et permettent souvent des interprétations linguistiques.

2.2. Prononciation de référence et variantes

D'un point de vue méthodologique, nous voulions retenir comme prononciation de référence des prononciations telles que définies et utilisées par les chercheurs en linguistique et psycholinguistique. Ainsi, nous avons choisi le dictionnaire de prononciation de référence Lexique380 (New *et al.*, 2007), afin d'examiner nos résultats indépendamment du dictionnaire de prononciation d'un système d'alignement spécifique (en l'occurrence celui du LIMSI). Ce choix entraîne cependant de ne considérer que les mots qui sont présents à la fois dans Lexique380 et dans le dictionnaire de prononciation du système d'alignement. Ce dernier a une couverture de 100% par rapport aux corpus de parole traités (par construction) alors que Lexique380 ne couvre pas l'ensemble, et en particulier, il n'inclut pas les noms propres. Ainsi, pour notre étude sur la propension à la réduction des segments, 21% des mots-tokens ont été exclus, concernant notamment des noms propres qui n'ont pas été répertoriés dans Lexique380.

Le dictionnaire de prononciation du système d'alignement contient ces prononciations de référence, appelées aussi formes canoniques (ou formes sous-jacentes suivant le contexte). Si un mot n'est représenté que par sa prononciation canonique, l'étiquetage du système d'alignement ne sera pas capable de révéler des différences phonétiques. En effet, l'alignement automatique découpera le signal du mot en autant de segments que de phonèmes dans la forme canonique. Les segments résultants, nommés d'après les étiquettes phonémiques de la forme canonique, sont de durées nécessairement plus courtes que si cette réalisation pouvait être alignée avec une variante réduite (impliquant moins de phonèmes). Afin de pouvoir produire des étiquetages automatiques révélant des différences de réalisation, il est nécessaire d'introduire des variantes (Boula de Mareüil et Adda-Decker, 2002 ; Schuppler *et al.*, 2008 ; Schuppler *et al.*, 2014). Pour la langue française, les variantes ajoutées dans le dictionnaire du système concernent essentiellement la présence optionnelle de schwa et de consonnes de liaison. Certaines variantes supplémentaires ont été introduites pour prendre en compte des phénomènes de réduction connus et fréquents, comme par exemple la réalisation du mot « il » comme [i] permettant ainsi la présence ou absence du segment [l] en fonction du signal de parole à aligner.

Pour examiner nos résultats d'alignement quant à la présence/absence d'un segment, nous comparerons la prononciation alignée (et qui reflète au mieux la production des locuteurs) avec la prononciation de référence (celle donnée dans Lexique380). Ainsi, nous aurons une prononciation alignée (forme de surface) et une prononciation de référence (forme sous-jacente) pour chaque mot examiné. Par exemple, le mot « quatre » /katʁ/ admet comme variante réduite [kat] dans le dictionnaire du système. Si une occurrence de ce mot est alignée comme [kat], nous pourrons détecter que le segment /ʁ/ est « absent » par comparaison des deux formes de prononciation (Figure 2).

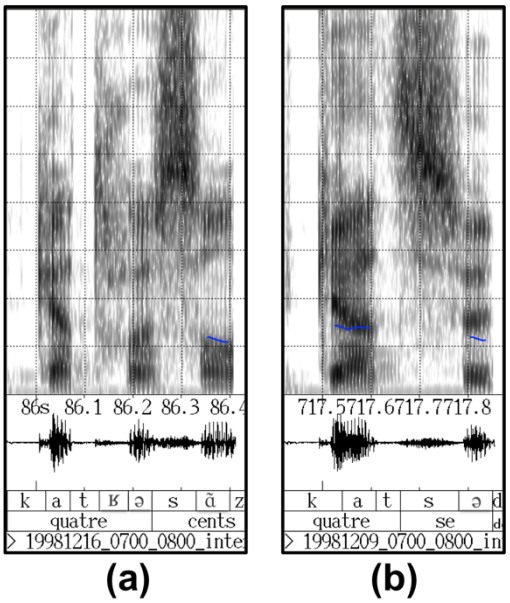

Figure 2. Le mot « quatre » /katʁ/ avec (a) et sans (b) /ʁ/ ou schwa, aligné par le système de transcription du LIMSI.

2.3. Méthode de recherche proposée : méthode ascendante

La démarche traditionnelle en linguistique consiste à formuler une hypothèse précise et ensuite à vérifier si cette hypothèse est confirmée. Cette démarche est tout à fait pertinente, mais se limite de fait aux hypothèses que nous sommes capables de formuler. Dans ce cas, on peut introduire des variantes spécifiques dans le dictionnaire de prononciation du système d'alignement afin de tester leur pertinence dans les corpus de parole à disposition. La réalisation (ou non) du schwa, de la liaison ou du /ʁ/ dans différents contextes rentrent bien dans ce paradigme d'analyse, que nous nommons méthode descendante (Wu *et al.*, 2017, 2019 ; Wu, 2018 ; Boula de Mareüil *et al.*, 2003 ; Adda-Decker *et al.*, 2012).

Nous pensons que, en ce qui concerne la réduction en parole spontanée, il reste de nombreuses zones d'ombre pour lesquelles nous sommes en mal de formuler des hypothèses claires. En effet, en parole spontanée, beaucoup de facteurs pourraient être en interaction et influencer la production réelle des locuteurs. Par conséquent, il est difficile de prévoir précisément les moments et les endroits où les réductions pourraient avoir lieu. Pour cette situation, nous proposons la « méthode ascendante » qui vise à exploiter ce qui est en général considéré comme un point faible de l'alignement automatique. En effet, comme nous l'avons décrit plus haut, à l'issu de l'alignement automatique, les zones de parole réduite se caractérisent par des séquences de segments de courte durée (typiquement de 30 ou 40 ms) et par un manque de précision phonétique, à la fois concernant l'identité du segment et la localisation des frontières.

La méthode ascendante se veut être une méthode généraliste capable de détecter n'importe quelle séquence de parole temporellement réduite par rapport à une prononciation canonique (ou complète). En fait, Meunier et Bigi (2016) ont appliqué une méthode similaire pour l'étude de segments courts, mais sans ajouter une contrainte sur le nombre de segments réduits consécutifs comme nous le proposons ici. Nous l'utilisons ici pour non seulement localiser les segments courts, mais aussi pour détecter des suites de segments courts. Nous avons évoqué précédemment le fait que la parole spontanée comprend de nombreuses occurrences

d'articulation affaiblie, réduite incluant éventuellement des prononciations raccourcies. Si, lors de l'alignement forcé, nous obtenons des suites de segment de phones de 30 (ou 40) ms, de telles séquences révèlent un « mismatch », c'est-à-dire une inadéquation entre le modèle acoustique prévu par le dictionnaire et l'articulation en effet réalisée. Il se peut qu'une variante de prononciation réduite typique ne soit pas incluse dans le dictionnaire de prononciation. Dans tous les cas, cette situation de « mismatch » est suspecte, soit du point de vue du système d'alignement et des techniques de modélisation, soit du point de vue linguistique : une séquence de segments courts pointe potentiellement sur des zones de réduction temporelle avec des productions peu décrites dans la littérature.

Un exemple typique en parole spontanée concerne la suite de mots « je (ne) sais pas » souvent prononcée par une séquence monosyllabique à peu près comme [ʃpa]. Dans cet exemple, la réduction englobe une suite de mots courts avec un même segment pouvant chevaucher deux ou plusieurs mots : la fricative [ʃ] produite peut s'expliquer à la fois par la fricative théoriquement voisée du mot « je » dont elle garde le lieu d'articulation que par la fricative [s] du mot « sais » dont elle garde le trait de voisement – on est en présence d'une assimilation régressive de voisement et progressive du lieu d'articulation. Pour la reconnaissance automatique de la parole, il est envisageable d'agglutiner les mots concernés afin de former une nouvelle entrée lexicale dans le dictionnaire « je_sais » à laquelle on peut attribuer les prononciations réduites [ʃɛ] et [ʃ]. Cependant, de telles formes réduites ne sont en général pas modélisées correctement dans les dictionnaires des systèmes, et le résultat de l'alignement forcé sera dans ce cas une suite de segments de durée minimale de 30 ms afin de placer tous les phones de son modèle acoustique trop long correspondant à la prononciation complète [ʒənəsɛpa] ou [ʒəsɛpa] en omettant la particule de négation « ne ».

En dehors d'hypothèses linguistiques précises, nous pourrons profiter du fait que des suites de segments courts sont produites automatiquement par l'alignement forcé dans les zones de forte réduction, peu importe leur origine ou leur nature précise. Grâce à l'alignement forcé, nous obtenons des suites de segments courts de 30 (ou 40) ms. L'approche ascendante permet ainsi de filtrer les données en deux parties : une partie avec des séquences de segments considérés de durée « normale » et une partie avec des séquences de segments courts considérés comme pointant potentiellement sur des phénomènes de réduction. Contrairement à l'approche descendante où nous pouvons espérer une segmentation et un étiquetage plus précis, avec l'approche ascendante, nous focalisons notre intérêt sur des zones dont les qualités de segmentation et d'étiquetage sont fortement suspectées de poser problème et ainsi, de nous révéler des phénomènes de réduction peu connus et décrits dans la littérature.

2.4. Corpus

Le corpus *Nijmegen Corpus of Casual French* (NCCFr) (Torreira et Ernestus, 2010) a été utilisé pour cette étude. Il contient 35 heures de conversations familières entre amis, incluant au total 46 locuteurs répartis en 24 femmes et 22 hommes. Tous les locuteurs sont des étudiants à l'université âgés d'environ 20 ans, sauf deux locutrices de 40 et 50 ans. Les enregistrements ont été effectués dans le studio d'enregistrement du Laboratoire de Phonétique et Phonologie (UMR7018, CNRS – Sorbonne Nouvelle) à Paris. L'alignement forcé automatique a été effectué à l'aide du système de reconnaissance automatique de la parole du LIMSI (Gauvain *et al.*, 2005).

3. Analyses et résultats sur la propension à la réduction des segments

Dans cette section, nous nous focaliserons sur les résultats obtenus en utilisant la méthode ascendante définie dans la section 2.3 qui consiste à localiser les séquences d'au moins trois segments consécutifs de 30 ou 40 ms et d'étudier l'identité des segments impliqués.

Afin de tenir compte du fait que notre dictionnaire utilisé pour l'alignement peut inclure des variantes réduites, nous discuterons la propension à la réduction des segments en prenant en compte non seulement les segments réduits tels que mis en évidence par l'alignement automatique dans la forme de surface, mais également les segments absents (en comparant la prononciation de référence ou sous-jacente avec la prononciation alignée ou de surface). Au-delà d'une analyse au niveau des segments, nous aimerions apporter de nouvelles connaissances sur la propension à la réduction de séquences de segments en analysant des « suites » (>= 3 segments) de segments courts et des segments qui sont absents dans l'alignement. Nous aimerions identifier les phones qui ont le plus tendance à être réduits et inversement ceux qui résistent le plus à la réduction.

La figure 3 illustre la distribution de la durée des segments dans le corpus NCCFr. Plus de 20% des segments ont une durée de 30 ms dans ce corpus, ce qui est la durée minimum permise par l'alignement. Il est cependant intéressant de noter que ce taux sous-estime la proportion de segments sujets à réduction, dans la mesure où un certain nombre de réductions est directement pris en compte par l'alignement d'une variante de prononciation plus courte. Outre ces segments de durée minimale, le sommet de la distribution se trouve à 50 ms. Cette distribution du corpus NCCFr composé de conversations entre amis est similaire à celle du corpus téléphonique en parole spontanée illustrée par Adda-Decker et Lamel (2018), avec les mêmes caractéristiques de réduction temporelle (fort taux de segments de 30 ms > 20%, valeur modale de la distribution localisée très à gauche, à 50 ms).

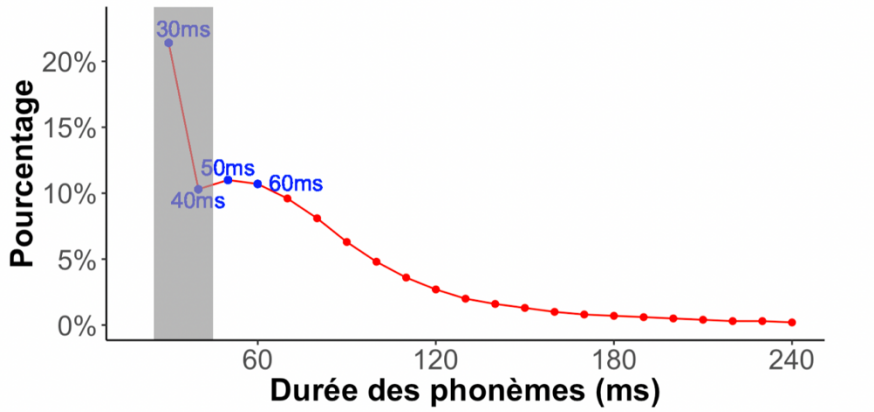

Figure 3. Distribution de la durée des segments dans le corpus conversationnel NCCFr (Torreira et Ernestus, 2010). L'abscisse concernant la durée segmentale est donnée en secondes. L'ordonnée indique le pourcentage de cette durée dans le corpus.

Le dictionnaire de prononciation inclut des variantes réduites pour des phénomènes fréquents et récurrents, comme la prononciation de surface [i] pour le mot « il » (chute du -l final). De ce fait, si nous voulons identifier si un mot est réduit en nous basant uniquement sur les suites de segments courts, nous risquons d'ignorer les mots pour lesquels la prononciation réduite a été utilisée lors de l'alignement forcé (cf. figure 4).

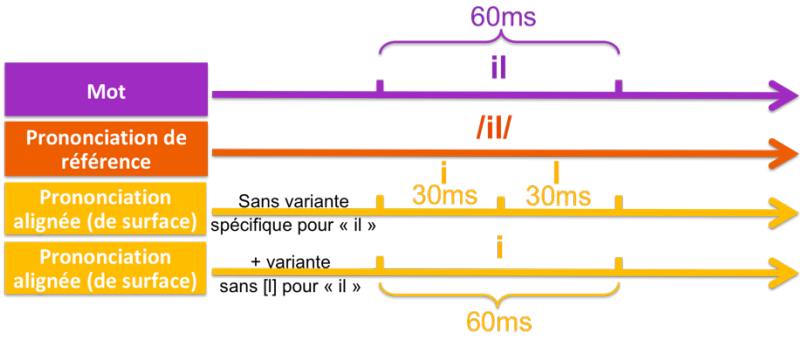

Figure 4. Illustration de l'effet de l'alignement sans ou avec variantes de prononciation sur le résultat de segmentation. Si le dictionnaire inclut des prononciations réduites (ici [i] en plus de [il] pour « il »), l'alignement forcé pourra utiliser ces variantes réduites et aura moins tendance à produire des segments courts (ici 60 ms pour [i], au lieu de 30 ms avec prononciation de référence [il]).

Dans la prononciation de surface obtenue par l'alignement automatique, certains segments de la prononciation de référence peuvent être absents (« Abs ») et d'autres de durée qu'on considère comme normale ou longue (> 40ms, que nous nommons « Nrm ») ou courts que nous nommons « Alrt » (s'il s'agit de suites d'au moins trois segments consécutifs de 30 ou 40 ms). Ces différents étiquetages ont été attribués automatiquement. Notons que pour les suites limitées à seulement un ou deux segments courts, les segments en question ne sont pas retenus dans notre catégorie « Alrt », mais sont également catégorisés comme « Nrm ». Ce choix est motivé par le fait que l'alignement forcé peut avoir de multiples raisons de générer de temps en temps un segment court (notamment lorsque le modèle acoustique ne correspond pas bien au signal de parole en présence, ce qui peut être lié à divers bruits de bouche, bruits de fonds, parole superposée…). En revanche, plus le nombre de segments dans la séquence de segments courts est élevé, plus il y a de raisons de s'intéresser à la zone ainsi localisée : soit il y a une erreur due à l'alignement, et il est intéressant d'en connaître la cause ; soit la zone en question pointe sur un cas de décalage entre la prononciation de référence et la réalisation de surface. C'est ce dernier cas qui nous intéresse plus particulièrement. Ainsi, afin de ne retenir que des zones de parole où la présomption de réduction est élevée, notre critère de sélection vise les séquences d'au moins trois segments courts consécutifs, soit à l'intérieur d'un mot, soit au-delà des frontières de mot.

Le tableau 1 illustre un exemple selon différents cas que nous pouvons rencontrer à l'issue de l'alignement.

Ex. /stʁ/ du mot « ministre » /ministʁ/	
– Si les segments [s], [t] et [ʁ] (qui se suivent) sont alignés chacun avec une durée courte (30 ou 40ms)	→ [s] segment en alerte : « Alrt » → [t] segment en alerte : « Alrt » → [ʁ] segment en alerte : « Alrt »
– Si les segments [s] et [t] sont alignés chacun avec une durée courte (30 ou 40ms) et le [ʁ] est aligné avec une durée de 50ms	→ [s] segment sans alerte : « Nrm » → [t] segment sans alerte : « Nrm » → [ʁ] segment sans alerte : « Nrm »

Tableau 1. Exemple du mot « ministre » sur la catégorisation des segments dans cette étude.

Les segments « Abs » et « Alrt » seront utilisés comme indice pour identifier les segments qui ont une propension à la réduction et les segments « Nrm » seront utilisés comme indicateurs de segments « stables ».

Afin de découvrir quels sont les segments qui ont tendance à mieux résister à la réduction (en tenant compte de la chute des segments préalablement détectés par le système[5] et la réduction potentielle non prévue par le système), nous avons décidé de recourir à la prononciation de référence (Lexique380) de New *et al*., 2007. La comparaison entre prononciation de référence et prononciation alignée nous permet d'affiner la mesure objective du taux de segments réduits.

Nous avons également établi une *stop list*[6] qui inclut les mots fréquents ayant 2000 mots-tokens ou plus dans le corpus (voir tableau 2). Cette *stop list* inclut plus de 55% des mots-tokens du corpus NCCFr (207309 occurrences sur 378515). Elle contient 41 mots-types. Nous comparerons nos résultats avant et après la suppression des mots de la *stop list*, ce qui permet d'illustrer l'importance relative des mots les plus fréquents sur la propension à la réduction.

Nous présenterons tout d'abord nos résultats sur la propension à la réduction des segments sans considérer le fait que certains segments soient absents (« Abs ») dans l'alignement. Dans un deuxième temps, nous présenterons nos résultats sur la propension à la réduction des segments en regroupant les segments « Alrt » et les segments « Abs », et nous comparerons ces segments avec les segments « Nrm ». Nous comparerons également les résultats avant et après la suppression des mots de la *stop list*.

[5] Les variantes de production sur les mots extrêmement fréquents tels qu' « il », qui peut être produit comme [i] tout court sans le /l/ en parole spontanée, sont incluses dans le système, comme mentionné ci-dessus.
[6] Liste de mots ou d'autres éléments qui devraient être ignorés dans le traitement des données pour une raison spécifique. Ici, il s'agit des mots extrêmement fréquents en parole continue.

Orthographe	est	je	tu	que	pas	de	ça
Prononciation de référence	ɛ	ʒə	ty	kə	pa	də	sa
Occurrences	16381	10178	9685	8804	8439	8345	8228

Orthographe	mais	et	il	le	ouais	la	a
Prononciation de référence	mɛ	e	il	lə	wɛ	la	a
Occurrences	8135	7114	6524	6175	5981	5820	5654

Orthographe	les	on	non	à	un	sais	des
Prononciation de référence	le	ɔ̃	nɔ̃	a	ɛ̃	sɛ	de
Occurrences	5305	4856	4839	4794	4607	4185	3986

Orthographe	quoi	fait	l'	en	elle	moi	y
Prononciation de référence	kwa	fɛ	l	ɑ̃	ɛl	mwa	i
Occurrences	3915	3671	3637	3634	3547	3521	3519

Orthographe	qui	ils	une	oui	enfin	ai	là
Prononciation de référence	ki	il	yn	wi	ɑ̃fɛ̃	ɛ	la
Occurrences	3302	3282	2982	2902	2688	2591	2494

Orthographe	dans	pour	t'	si	plus	vois
Prononciation de référence	dɑ̃	puʁ	t	si	ply	vwa
Occurrences	2477	2376	2365	2278	2062	2031

Tableau 2. Mots d'au moins 2000 occurrences inclus dans la *stop list*.

Dans ce qui suit, nous tenterons d'identifier quels segments ont le plus tendance à disparaître en parole spontanée familière. Pour cela nous calculons le taux de segments réduits par type de segment. Par exemple, le taux de segments réduits pour le phonème /t/ est donné par le rapport entre le nombre de segments de la consonne /t/ ayant été étiquetés comme « Alrt » (et éventuellement « Abs ») et le nombre total de segments de la consonne /t/. Toutes les figures montrant les taux de segments réduits par voyelles ou par consonnes gardent une échelle fixe sur l'axe des ordonnées (entre 0 et 40% de réduction).

La figure 5 illustre les taux de segments réduits de chaque voyelle sans prendre en compte l'absence des segments dans l'alignement[7] et la figure 6 illustre les taux de segments réduits de chaque consonne. Notons que nous avons regroupé les /e/ et les /ɛ/ en raison de l'alternance parfois « libre » entre ces deux phonèmes en français[8]. Nous avons également procédé au regroupement des /o/ et des /ɔ/ pour la même raison. Les schwas et les /œ/ ont été regroupés car ces deux segments sont représentés par le même symbole dans l'alignement (c.-à-d. prononciation de surface). Ici, le taux de segments réduits représente les segments qui font

[7] Segments non-alignés en raison des variantes spécifiques introduites dans le dictionnaire de prononciation du système de reconnaissance.

[8] Ex. Le mot « sérieux » /seʁjø/ peut être prononcé [sɛʁjø].

partie des suites de segments courts (segment « Alrt ») par rapport à tous les segments de la forme de référence :

$$\text{Taux de segments réduits (\%)} = \frac{\text{Segments « Alrt »}}{\text{Total des segments de la forme de référence}} \times 100$$

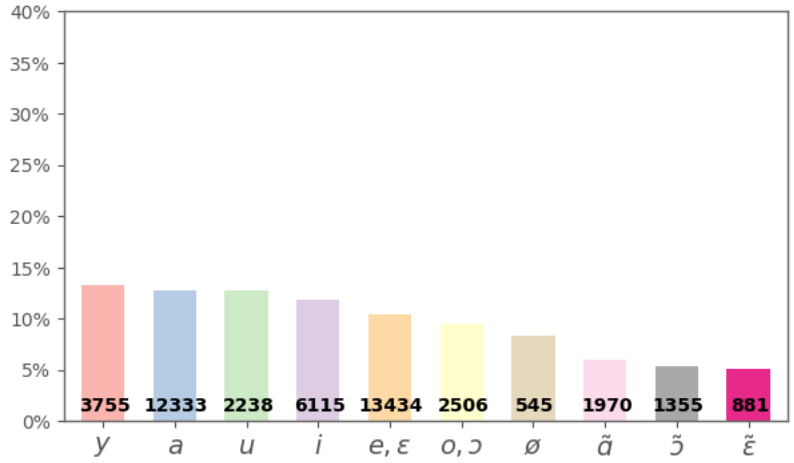

Figure 5. Taux de segments réduits pour les voyelles (schwa exclu) sans prendre en compte l'absence des segments dans l'alignement. Le nombre d'occurrences de ces segments est illustré sur les barres.

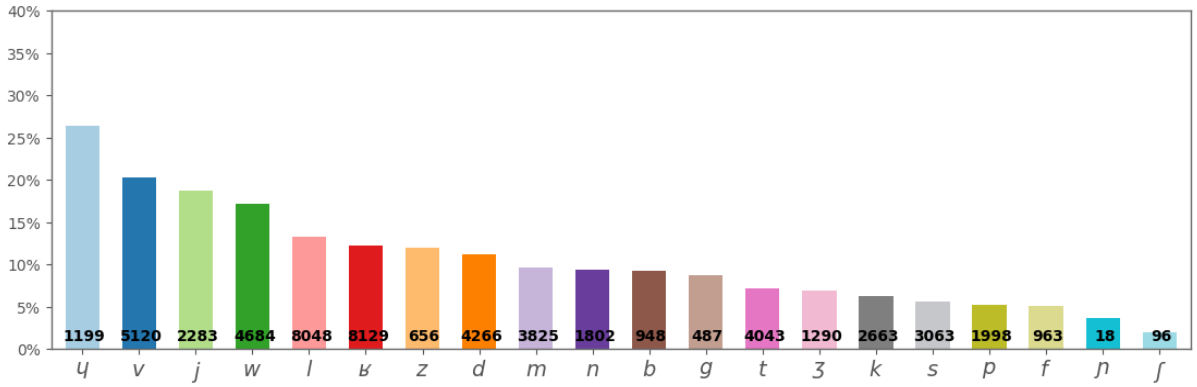

Figure 6. Taux de segments réduits pour les consonnes sans prendre en compte l'absence des segments dans l'alignement. Le nombre d'occurrences de ces segments est illustré sur les barres.

Dans la figure 5 les voyelles sont triées par taux de segments réduits décroissant. Tous les taux restent inférieurs à 15% et il n'y a pas de voyelle dont le taux soit remarquablement élevé. Les voyelles orales ont des taux plutôt autour de 10%. Nous remarquons que les voyelles nasales ont tendance à être moins réduites que les voyelles orales, avec des taux autour de 5%. D'après cette figure, il n'y a pas de tendance spécifique remarquable parmi les voyelles orales.

Contrairement à la figure des voyelles, la figure des consonnes (figure 6) montre quelques taux dépassant les 15%. Ainsi, nous remarquons que les segments qui résistent le moins à la réduction sont les semi-consonnes /ɥ, j, w/ et la fricative labiale voisée /v/. Arrivent ensuite

les liquides dont les taux de segments réduits sont cependant inférieurs à ceux observés pour les semi-consonnes et le /v/, et qui sont plutôt proches des consonnes obstruantes alvéolaires /z/ et /d/.

Ces premiers résultats ne comptabilisent pas les réductions prises en charge lors de l'alignement forcé par les variantes réduites. Il nous paraît donc plus juste d'inclure dans les taux de segments réduits également les segments omis lors de l'alignement du fait que la prononciation réduite est déjà connue du système (typiquement, [i] pour « il ») et qu'elle a été utilisée (cf. figure 4). Dans la suite, nous allons recalculer les taux de segments réduits par type de segment en prenant en compte à la fois les segments localisés dans les suites de segments courts et les segments qui n'ont pas été alignés en raison de la présence de variantes. Pour ce faire, nous devrons tout d'abord calculer la différence entre les occurrences observées de la forme de référence et celle de la forme de surface. Le taux de segments réduits représentera ainsi les segments « Abs » et ceux qui font partie des suites de segments courts (segment « Alrt ») par rapport à tous les segments de la forme de référence :

$$\text{Taux de segments réduits (\%)} = \frac{\text{Segments « Abs »} + \text{segments « Alrt »}}{\text{Total des segments de la forme de référence}} \times 100$$

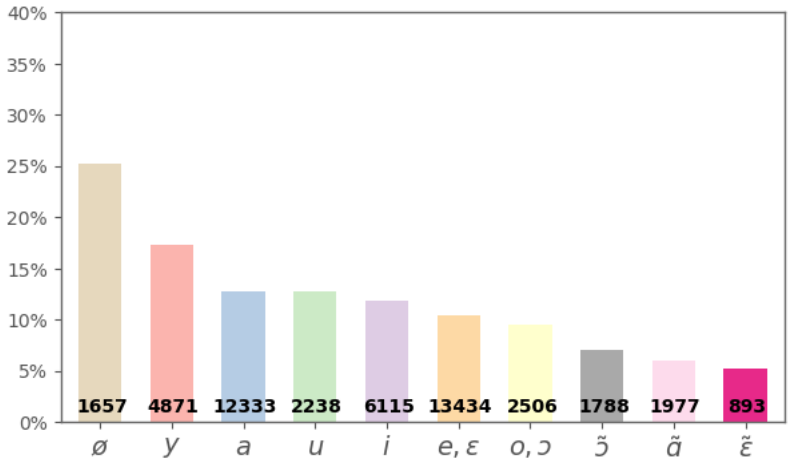

Figure 7. Taux de segments réduits pour les voyelles (schwa exclu) en prenant en compte l'absence des segments dans l'alignement.

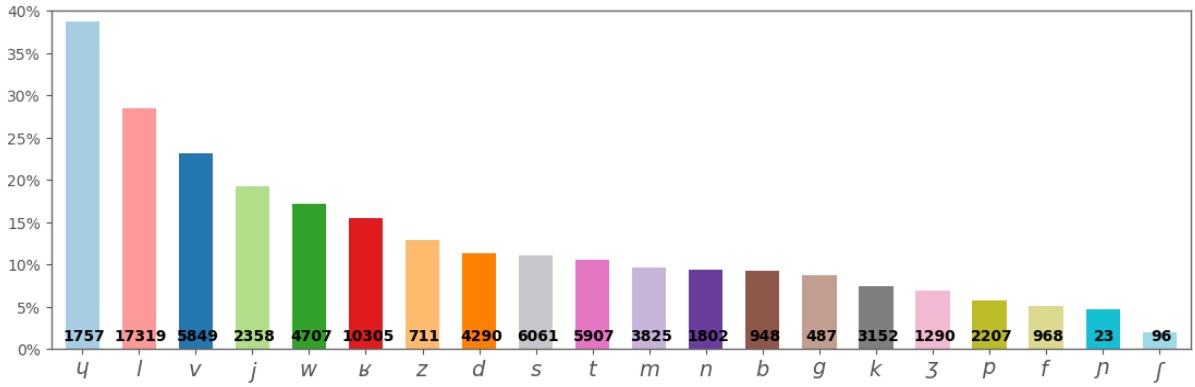

Figure 8. Taux de segments réduits pour les consonnes en prenant en compte l'absence des segments dans l'alignement.

La figure 7 reprend le cas des voyelles comme dans la figure 5. Mais dans le cas présent, nous comptabilisons non seulement les séquences « Alrt » comme relevées uniquement par l'alignement de séquences de segments courts, mais on tient compte également du cas des segments manquants (par le biais des prononciations réduites comme expliqué dans la figure 4). On remarque clairement une différence entre l'ordre des voyelles dans les deux figures 5 et 7 : le /y/ perd sa première position au profit du /ø/. Nous observons une augmentation considérable des taux de segments réduits pour ces deux voyelles, dépassant maintenant 15% par rapport à ce qui a été observé dans la figure 5. Cela est lié au fait que des variantes existent pour quelques mots fréquents dans le dictionnaire de prononciation, comme par exemple, « peut-être » [ptɛtʁ] sans /ø/.

La figure 8 illustre le taux de segments réduits pour les consonnes en prenant en compte l'absence des segments dans l'alignement. Nous remarquons que l'allure de la courbe s'est déformée en accentuant les taux pour les consonnes les plus sujettes à réduction. Les semi-consonnes /ɥ, j, w/ et le /v/ ont toujours des taux très élevés. Néanmoins, nous observons une augmentation considérable des taux de segments réduits pour le /ɥ/, le /l/ et le /v/ ; le /l/ ayant un taux plus élevé que le /v/ et les semi-consonnes /j,w/ cette fois-ci. Le fait que les /ɥ/ et /l/ aient un taux de segments réduits nettement plus élevé dans cette figure (par rapport à ce qui a été observé dans la figure 6) suggère que les mots ayant un /ɥ/ ou un /l/ dans la forme de référence sont souvent alignés avec la variante réduite par l'alignement automatique, dès lors que cette variante existe.

Pour que nos résultats sur la propension à la réduction des segments soient moins influencés par les mots extrêmement fréquents dans le corpus (c.-à-d. les mots de la *stop list*), nous avons décidé d'exclure ces mots de nos données et de voir si cela change les résultats observés dans les figures 7 et 8.

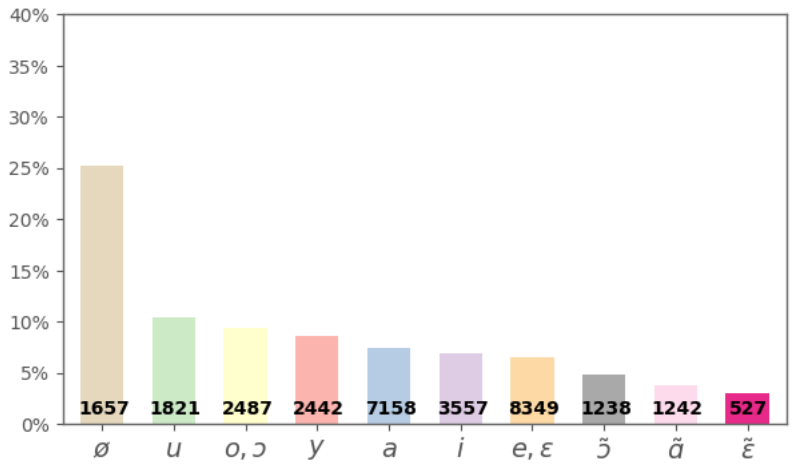

Figure 9. Taux de segments réduits pour les voyelles (schwa exclu) en prenant en compte l'absence des segments dans l'alignement et en excluant les voyelles provenant des mots de la *stop list*.

La figure 9 illustre le taux de segments réduits pour les voyelles en prenant en compte l'absence des segments dans l'alignement et en éliminant les mots de la *stop list*. Nous remarquons sur la figure 9 que le taux de segments réduits des /y/ et celui des /a/ ont baissé

davantage après la suppression des occurrences qui concerne les mots de la *stop list* par rapport à ce qui a été observé dans la figure 7. Cela indique que le taux de segments réduits pour les voyelles /y/ et /a/, illustré dans la figure 9, a été fortement influencé par les mots de la *stop list* ; notamment le mot « tu » pour la voyelle /y/ et le mot « la » pour la voyelle /a/. Le taux de segments réduits pour la voyelle /ø/ reste le plus élevé dans la figure 9 (comme dans la figure 7). Cela provient fréquemment des /ø/ réduits dans les mots tels que « peut-être » et « veut » en parole spontanée. Nous nous apercevons que les phonèmes ayant les taux de segments réduits les plus élevés (c.-à-d. qui résistent le moins à la réduction) sont des voyelles orales arrondies /ø/, /u/, /o, ɔ/ et /y/ ($\chi^2 = 3173{,}3$; df = 1 ; $p < 0{,}001$). Les taux de segments réduits les moins élevés restent toujours les voyelles nasales ($\chi^2 = 15{,}446$; df = 1 ; $p < 0{,}001$).

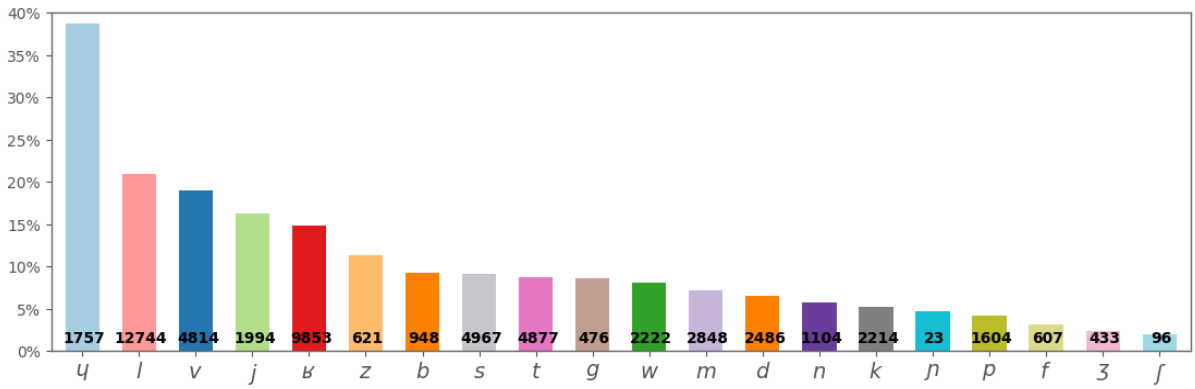

Figure 10. Taux de segments réduits pour les consonnes en prenant en compte l'absence des segments dans l'alignement et en excluant les consonnes provenant des mots de la *stop list*.

La figure 10 illustre le taux de segments réduits pour les consonnes en prenant en compte l'absence des segments dans l'alignement et en éliminant les mots de la *stop list*. La différence entre la figure 8 et la figure 10 se trouve surtout sur le taux de segments réduits de la semi-consonne /w/. Cela est majoritairement dû aux mots « quoi », « moi », « oui » et « vois » de la *stop list*. D'après les différentes figures de consonnes, les glides (dans une moindre mesure le /w/) et les liquides sont enclines à la réduction ($p < 0{,}001$), et la consonne fricative post-alvéolaire non-voisée /ʃ/ a le « taux d'alerte » le moins élevé parmi les consonnes. La fricative labiale voisée /v/ a un statut spécial : le taux de segments réduits du phonème (≈ 20%) est beaucoup plus élevé que ceux des autres fricatives (entre 2 et 13%).

Nous avons d'ailleurs effectué des tests statistiques avancés (avec des modèles linéaires mixtes généralisés) afin de tester si la fréquence de mots était un facteur qui pourrait favoriser la réduction. Nous avons observé effectivement une influence significative de la fréquence de mots sur la réduction ($p < 0{,}01$).

4. Discussion

Cette étude sur la propension à la réduction des segments montre qu'il est possible d'utiliser des grands corpus de parole et des systèmes de traitement automatique de la parole pour y rechercher et quantifier des phénomènes linguistiques, comme la réduction des segments, peu décrits dans la littérature. Avec la méthode ascendante proposée, il n'est pas nécessaire d'avoir des hypothèses très précises sur le phénomène retenu. Il est cependant important de savoir le caractériser objectivement de manière à pouvoir utiliser l'instrument de mesure (ici le système d'alignement forcé) afin de rechercher des extraits de parole dont les mesures objectives sont

particulièrement intéressantes (par exemple, déviantes ou au contraire dans la norme) pour le phénomène étudié. Dans notre cas, nous utilisons la caractérisation de durée des segments alignés des mots ou de suites de mots. Des durées jugées « trop courtes » par rapport à la durée attendue (étant donnée la forme phonologique sous-jacente) révèlent très probablement des phénomènes de réduction.

Nous avons considéré comme séquence réduite non seulement des suites d'au moins 3 segments consécutifs alignés avec des durées courtes de 30 ou 40 ms, mais nous avons également pris en compte des segments de durée effective de 0 ms correspondant à des phonèmes présents dans la prononciation de référence (la forme sous-jacente) mais absents de l'alignement à cause de l'utilisation d'une variante de prononciation réduite lors de l'alignement (forme de surface). Ceci nous a permis de mettre en évidence l'importance des segments absents/non-alignés lors de l'utilisation de cette méthode.

Grâce à la méthode ascendante, nous avons pu observer que les liquides, les semi-consonnes (glides) et la fricative voisée /v/ sont particulièrement sujettes à réduction, ce qui est cohérent avec le fait que leurs durées intrinsèques sont relativement courtes. Ceci pourrait être lié au fait que leurs caractéristiques acoustiques ressemblent davantage à celles des voyelles que celles des obstruantes (hors /v/) par exemple. Au contact des voyelles, les liquides et les glides auront tendance à fusionner avec les voyelles environnantes, particulièrement en parole rapide ou peu articulée et pour des syllabes non-accentuées. Concernant les voyelles, les voyelles nasales résistent mieux à la réduction que les voyelles orales. Les voyelles nasales ont une durée intrinsèque plus longue. Un segment de voyelle nasale est composé typiquement d'une première partie orale suivi d'une partie nasalisée. Les voyelles nasales ont à leur disposition la nasalité en plus et, de ce fait, elles seraient intrinsèquement renforcées. Les voyelles orales arrondies, quant à elles, résistent moins à la réduction que d'autres voyelles orales. Ceci pourrait éventuellement être lié au fait que le trait arrondi qui est très saillant entraîne une coarticulation forte avec les consonnes qui l'entourent. Ces voyelles, quand elles deviennent très courtes, se trouvent souvent dévoisées ou peuvent être vues comme partie intégrante de fricatives ou du burst d'occlusives qui les précèdent. L'absence d'un segment autonome vocalique ne gêne en général pas l'intelligibilité du mot, notamment le mot dans son contexte. Dans le futur, des tests perceptifs peuvent être envisagés sur des échantillons de parole étiquetée comme réduite afin de tester l'effet de cette réduction sur la perception humaine.

Nos résultats en ce qui concerne la réduction des consonnes sont conformes à ce qui a été observé dans les données de Meunier et Bigi (2016) : les liquides et les glides ont plus tendance à être réduites que d'autres consonnes. En plus de ce qui a été montré par Meunier et Bigi (2016), nous remarquons que la consonne fricative voisée /v/ a un « taux d'alerte » très élevé, au niveau de celui des liquides et des glides.

Au-delà de ce qui a été montré par Meunier et Bigi (2016) en ce qui concerne la propension à la réduction des voyelles[9], nous observons que les voyelles orales ont davantage tendance à être réduites que les voyelles nasales. Parmi les voyelles orales, les voyelles orales arrondies (/ø/, /u/, /o, /ɔ/ et /y/) ont plus tendance à être réduites que les autres voyelles orales dans notre corpus.

[9] Meunier et Bigi (2016) montrent que les voyelles fermées ont tendance à être réduites.

La tendance de réduction illustrée dans cette étude met en évidence une grande variabilité dans les productions orales, non seulement au niveau paradigmatique (segmental) mais également au niveau syntagmatique (séquences de segments) : concernant la production de mots, on observe des différences importantes entre les formes de surface et les formes sous-jacentes, qui peuvent aller bien au-delà des phénomènes communément décrits pour le français (schwa, liaison, simplification de clusters obstruante-liquide). L'écart entre forme sous-jacente et forme de surface semble d'autant plus facilement toléré qu'il n'altère pas l'intelligibilité de l'information linguistique en cours de transmission. Par exemple, si le mot « plus » (/ply/ ou /plys/) est prononcé sans le /l/ ([py] ou [pys]), il y a certes de nombreux homophones en français possibles pour un mot prononcé [py] ou [pys], mais il n'y a pas d'autres mots de la même fonction syntaxique et sémantique qui pourraient interférer avec l'intelligibilité du mot : les mots « pu » (participe passé du verbe « pouvoir ») ou « pus » (nom commun) /py/ n'ont pas les mêmes fonctions syntaxiques que le mot « plus » /p͟ly/ prononcé [py] en parole continue. De même le mot « puce » /pys/ n'a pas la même fonction syntaxique que le mot « plus » /p͟lys/ prononcé [pys].

Nos résultats sur la propension à la réduction des segments peuvent aider à développer davantage les dictionnaires de prononciation spécifiques à la parole spontanée, utiles à la fois pour les systèmes de reconnaissance et de synthèse automatiques comme pour l'apprentissage du français langue étrangère. Ils peuvent également offrir des pistes intéressantes à tester pour des recherches expérimentales de laboratoire. Enfin, nos résultats posent des questionnements sur les différents processus phonologiques et cognitifs à l'œuvre dans la communication verbale, permettant en production la réalisation de formes raccourcies (formes de surface) et perçues complètes ou restaurées (forme sous-jacente) par l'auditeur.

Bibliographie

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

Adda-Decker, M., Boula de Mareüil, P., & Lamel, L. (1999, August). « Pronunciation variants in French: schwa & liaison », *XIVth International Congress of Phonetic Sciences* : 2239-2242.

Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E., & Lamel, L. (2012). « French Liaison in Casually Spoken French, as Investigated in a Large Corpus of Casual French Speech », *Revue française de linguistique appliquée* 17(1) : 113-128.

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

Adda-Decker, M., & Lamel, L. (2018). « Discovering speech reductions across speaking styles and languages », *Rethinking reduction - Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation* : 101-128 .

Bengio, Y. (2009). « Learning deep architectures for AI ». *Foundations and trends® in Machine Learning* 2(1) : 1-127.

Boula de Mareüil, P., & Adda-Decker, M. (2002). « Studying pronunciation variants in French by using alignment techniques », *Seventh International Conference on Spoken Language Processing 2002*.

Boula de Mareüil, P. B., Adda-Decker, M., & Gendner, V. (2003). « Liaisons in French: a corpus-based study using morpho-syntactic information », *ICPhS 2003*.

Bridle, J. S., & Brown, M. D. (1974). « An experimental automatic word recognition system ». *JSRU report* 1003(5) : 33.

Davis, S., & Mermelstein, P. (1980). « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences ». *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28(4) : 357-366.

Dilley, L. C., & Pitt, M. A. (2010). « Altering context speech rate can cause words to appear or disappear », *Psychological Science* 21(11) : 1664-1670.

Duez, D. (1997). « Acoustic markers of political power », *Journal of Psycholinguistic Research* 26(6) : 641-654.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*, thèse de doctorat, Vrije Universiteit Amsterdam, Utrecht : LOT.

Forney, G. D. (1973). « The Viterbi algorithm ». *IEEE 1973*, 61(3) : 268-278.

Gauvain, J. L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., & Schwenk, H. (2005). « Where are we in transcribing French broadcast news? », *Ninth European conference on speech communication and technology, Interspeech 2005*.

Hermansky, H. (1990). « Perceptual linear predictive (PLP) analysis of speech ». *Journal of the Acoustical Society of America* 87(4) : 1738-1752.

Johnson, K. (2004). « Massive reduction in conversational American English », *Spontaneous speech: Data and analysis. 1st session of the 10th international symposium* : 29-54.

Kohler, K. J. (1990). « Segmental reduction in connected speech in German: Phonological facts and phonetic explanations », *Speech production and speech modelling* 55: 69-92.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). « Deep learning ». *Nature* 521(7553) : 436-444.

Meunier, C., & Bigi, B. (2016). « Répartition des phonèmes réduits en parole conversationnelle. Approche quantitative par extraction automatique ». *Actes de la conférence conjointe JEP-TALN-RECITAL 2016* : 615-623.

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

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). « The use of film subtitles to estimate word frequencies », *Applied psycholinguistics* 28(4) : 661-677.

Nguyen N., & Adda-Decker, M. (2013). *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*. Hermès-Lavoisier.

Rabiner, L. R. (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». *IEEE 1989* : 257-286.

Schuppler, B., Adda-Decker, M., & Morales-Cordovilla, J. A. (2014). « Pronunciation variation in read and conversational austrian german ». *Interspeech 2014* : 1453-1457.

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2008). « Preparing a corpus of Dutch spontaneous dialogues for automatic phonetic analysis », *Interspeech 2008* : 1638-1641.

Tahon, M., Lecorvé, G., & Lolive, D. (2018). « Can we Generate Emotional Pronunciations for Expressive Speech Synthesis? », *IEEE Transactions on Affective Computing, Institute of Electrical and Electronics Engineers 2018*.

Torreira, F., & Ernestus, M. (2010). « The Nijmegen Corpus of Casual Spanish », *LREC 2010* : 2981-2985.

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

Wu, Y. (2018). *Étude de la réduction segmentale en français parlé à travers différents styles : apports des grands corpus et du traitement automatique de la parole à l'étude du schwa, du /ʁ/et des réductions à segments multiples*, thèse de doctorat, Université Sorbonne Nouvelle – Paris 3.

Wu, Y., Adda-Decker, M., Fougeron, C., & Lamel, L. (2017). « Schwa Realization in French: Using Automatic Speech Processing to Study Phonological and Socio-Linguistic Factors in Large Corpora », *Interspeech 2017*.

Wu, Y., Gendrot, C., Adda-Decker, M, & Fougeron, C. (2019). « Post-consonantal Word-final /ʁ/ Realization in French: Contributions of Large Corpora », *ICPhS 2019*.