Méthodes et outils pour l'analyse phonétique des grands corpus oraux

Gilles Adda
Martine Adda-Decker
Philippe Boula de Mareüil
Julien Eychenne
Cédric Gendrot
Lori Lamel
Christine Meunier
Yohann Meynadier
Noël Nguyen
Atanas Tchobanov
Cécile Woehrling

3 décembre 2012

Préface

Après avoir longtemps suivi des chemins séparés, les chercheurs en phonétique et en phonologie, d'une part, et les spécialistes du traitement automatique du langage oral, d'autre part, en sont récemment venus à se rapprocher. Ce rapprochement est notamment lié au considérable essor des grands corpus oraux. En phonétique et en phonologie, nombreux sont les chercheurs qui considèrent aujourd'hui que leurs hypothèses sur la forme sonore du langage doivent s'appuyer sur des analyses quantitatives, appliquées à des données empiriques recueillies de telle manière que des généralisations soient possibles au-delà de l'échantillon étudié. De plus, une tendance forte aujourd'hui consiste à s'écarter de la parole lue par des locuteurs enregistrés individuellement, pour s'engager sur le sol à la fois plus riche et plus complexe des interactions conversationnelles entre deux ou plusieurs locuteurs. On a vu ainsi se multiplier les corpus oraux à grande échelle, dont l'exploitation a été ouverte à l'ensemble des chercheurs concernés, et cela a eu un impact majeur sur les recherches en phonétique et en phonologie. Le passage des « petits » aux « grands » corpus n'a pas simplement entraîné une extension quantitative de notre champ d'investigation : il a permis de faire émerger des méthodes d'analyse nouvelles. Par exemple, à cause de leur forte richesse, les grands corpus nous offrent la possibilité d'examiner de manière *a posteriori* l'influence de différents facteurs (phonétiques, lexicaux, grammaticaux, etc.), pris chacun séparément, sur la forme prononcée de telle voyelle ou consonne. Cette analyse décompositionnelle des facteurs susceptibles d'avoir une influence sur la forme sonore du langage, n'était auparavant possible qu'en construisant de toutes pièces un corpus dans lequel ces facteurs étaient manipulés de manière *a priori* et qui était ensuite donné à lire à différents locuteurs. Par ailleurs, le fait de rendre les grands corpus oraux librement accessibles à la communauté scientifique dans son ensemble, n'a pas simplement un intérêt économique lié à ce que des ressources sont partagées : il renforce considérablement la validité scientifique de nos analyses, en permettant à chacun de les répliquer, et de les étendre à des corpus plus larges encore.

Chapitre rédigé par Noël NGUYEN et Martine ADDA-DECKER.

Pour les spécialistes du traitement automatique du langage oral, les grands corpus oraux constituent un élément essentiel au développement et à la validation de leurs outils. Ces grands corpus leur permettent en particulier de se confronter à toutes les *sources de variation* dont la forme sonore du langage est le siège, dans l'espace géographique, dans l'espace social, et dans le temps. Ils offrent aux chercheurs la possibilité d'étendre le champ d'action de leurs outils au-delà de la parole dite standard vers des formes produites par des locuteurs de différentes générations et d'une large variété sur le plan géographique et social. Ils servent de support empirique pour la construction de techniques et d'outils aux performances accrues, des aligneurs automatiques entre texte et signal de parole, aux systèmes de synthèse et de reconnaissance automatique du langage oral.

Le projet international *Phonologie du français contemporain : usages, variétés et structure* (PFC, [DUR 02a, DUR 09])[1] incarne l'un des meilleurs exemples de ces grands corpus oraux servant de point de convergence pour les recherches en phonétique et phonologie, d'une part, et le traitement automatique du langage oral, d'autre part. Lancé il y a plus de dix ans par Jacques Durand, Bernard Laks et Chantal Lyche, PFC s'est fixé pour objectif de caractériser la prononciation du français contemporain dans toute sa diversité géographique, stylistique et sociale. Le très large corpus dont il est à l'origine (plusieurs centaines de locuteurs et des dizaines de points d'enquête dans l'espace francophone) est également destiné à mettre à l'épreuve les modèles théoriques développés aujourd'hui en phonologie et phonétique. Dans l'exploitation de ces données d'une richesse inégalée auparavant, s'est très vite imposée la nécessité de faire appel à des outils de traitement automatique ou semi-automatique du meilleur niveau. Cet ouvrage s'est construit dans le cadre du projet PFC et il en forme l'un des multiples résultats.

L'objectif de cet ouvrage est de présenter un panorama des méthodes et des outils utilisables pour les analyses phonétiques réalisées sur de grands corpus oraux. L'accent sera placé ici sur l'analyse du français, et les données recueillies dans PFC servent de support à nombre des exemples présentés. Les questions abordées concernent notamment le prétraitement des données acoustiques, l'extraction (semi-)automatique d'un ensemble de paramètres d'analyse, la mise en relation entre les données acoustiques et les catégories phonologiques. L'ouvrage accorde également une large place à la contribution apportée par les technologies de la parole, et notamment les systèmes de reconnaissance automatique de la parole, à l'analyse phonétique des grands corpus oraux.

Le chapitre 1, « Éléments de phonétique acoustique », rédigé par Yohann Meynadier, contient différents éléments de référence pour l'analyse acoustique des sons de la parole en français. Yohann Meynadier commence par définir les principales notions

1. Site web : `www.projet-pfc.net`.

de base en acoustique : vibrations périodiques simples et complexes, fréquence fondamentale, vibrations apériodiques, analyse spectrale, représentation temps-fréquence (spectrogramme), concepts de filtre et de résonateur. Il expose ensuite les principes de base de la théorie source-filtre et les liens que celle-ci permet d'établir entre configurations articulatoires et formes sonores. La suite du chapitre est consacrée à une présentation des sons du français rangés en quelques grandes classes : voyelles orales et nasales, consonnes obstruantes et consonnes sonantes, dont Yohann Meynadier décrit les principales caractéristiques articulatoires et acoustiques. Des valeurs de référence (ex. : fréquences formantiques relatives aux différentes voyelles du français) sont fournies. Un aperçu des principaux effets de coarticulation, d'assimilation et de réduction en français est également apporté.

Le chapitre 2, « Traitement et analyse du signal de parole », par Christine Meunier et Noël Nguyen, commence par quelques éléments de base sur les paramètres utilisés dans l'acquisition et l'analyse spectrale du signal de parole. Il offre ensuite une discussion détaillée sur la manière dont peuvent être définies les opérations de segmentation, d'alignement et d'annotation, et sur les questions soulevées par ces opérations. Les différences, mais aussi la complémentarité, entre segmentation manuelle et alignement automatique, sont soulignées. Les principaux critères de segmentation utilisables pour les macro-classes du français (occlusives, fricatives, consonnes vocaliques, voyelles) sont passés en revue. Le chapitre se termine par des éléments de discussion sur les solutions pouvant être apportées aux problèmes qui se rencontrent dans la segmentation dans la parole dite non préparée, au-delà de la parole de laboratoire.

Dans le chapitre 3, « Les bases de données parole », Julien Eychenne et Atanas Tchobanov abordent les questions concernant la construction d'une base de données parole de grande taille. Ces questions couvrent un vaste domaine allant du recueil des enregistrements à la consultation en ligne des données/métadonnées sur internet. Le chapitre offre un aperçu des méthodes utilisables pour la réalisation et le prétraitement des enregistrements sonores, et il compare différentes bases de données de référence disponibles aujourd'hui. Il expose ensuite l'architecture de la base PFC : les conventions employées dans le codage et la transcription de ces enregistrements sont passées en revue, ainsi que les outils développés pour faciliter le traitement, la consultation et l'analyse des données. L'accent est placé sur les questions d'interopérabilité et de pérennité des données, métadonnées et outils, questions qui forment autant d'enjeux majeurs pour l'accessibilité et la conservation de la base à long terme.

Dans le chapitre 4, « Les systèmes de transcription automatique de la parole comme instruments d'exploration de grands corpus oraux », Martine Adda-Decker, Lori Lamel et Gilles Adda montrent de quelle manière ces systèmes nous offrent aujourd'hui la possibilité d'explorer des corpus oraux dont la taille est virtuellement illimitée, en ouvrant un champ d'investigation hors de notre portée auparavant. Les auteurs présentent en premier lieu les principes de modélisation statistique de la parole permettant de réaliser des étiquetages et segmentations temporelles des données (« annotations »).

La qualité et la précision des annotations sont discutées en fonction de la configuration du système, et notamment du type des modèles acoustiques et des dictionnaires de prononciation. Selon la configuration choisie, les annotations produites peuvent représenter des prononciations canoniques (segmentation phonémique) ou bien refléter les variantes de prononciation (segmentation plutôt phonétique). La cohérence et la précision de la segmentation temporelle sont examinées, dans la mesure où de nombreuses études linguistiques prennent appui sur cette segmentation. Les capacités d'analyse de ces instruments sont mises en lumière à travers le cas des variantes de prononciation. Les capacités de mesure sont illustrées par quelques grandeurs simples, comme les comptes d'occurrence et les durées de phonèmes, et par des grandeurs plus complexes concernant la réalisation des variantes de prononciation.

Dans le chapitre 5, « Traitement de la variation régionale », Philippe Boula de Mareüil, Cécile Woehrling et Martine Adda-Decker commencent par souligner que la variation, en tant qu'objet de recherche en linguistique, nous oblige à manipuler d'importantes quantités de données. Les instruments de mesure dérivés à partir du traitement automatique de la parole sont donc particulièrement appropriés pour quantifier des tendances connues et moins connues en phonétique / phonologie. Après un aperçu des différentes dimensions dans laquelle peut se déployer la variation, ce chapitre se focalise sur la variation diatopique (i.e. régionale) entre le Nord et le Sud de la France. Des expériences perceptives sont brièvement présentées, qui font apparaître la difficulté d'identifier plus finement les accents de locuteurs pourtant bien ancrés géographiquement (issus du projet PFC). Différentes méthodes et techniques sont éprouvées pour caractériser ces accents : analyse discriminante, *clustering*, échelonnement multi-dimensionnel, arbres de décision… À partir d'alignements automatiques en phonèmes, les auteurs entreprennent notamment de montrer de quelle manière il est possible de hiérarchiser les traits de prononciation les plus discriminants parmi l'avancée du /ɔ/ ouvert vers [œ] dans le Nord, le maintien du schwa et la dénasalisation des voyelles nasales dans le Sud. Ils examinent également comment des changements linguistiques en cours peuvent affecter ces voyelles, qui sont gouvernées par le contexte gauche / droit, et mettent en lumière des différences entre lecture et parole spontanée.

Le chapitre 6, « De la normalisation formantique des voyelles », par Cédric Gendrot, est consacré aux méthodes utilisables pour traiter différentes sources de variabilité inter-individuelle dans la forme acoustique des sons de la parole, et notamment des voyelles. Cédric Gendrot commence par donner un aperçu de ce que nous savons aujourd'hui sur la façon dont le système auditif fait face à cette variabilité. Il présente ensuite un panorama des méthodes de normalisation inter-individuelle employées dans l'analyse de la forme spectrale des voyelles, et dont l'objectif est de nous permettre de faire abstraction des caractéristiques anatomiques individuelles des locuteurs. Chacune des ces méthodes offre à la fois des avantages et des inconvénients dont Cédric Gendrot offre une discussion détaillée.

Le chapitre 7, « Apport du traitement automatique à l'étude des voyelles », rédigé par Martine Adda-Decker, Cédric Gendrot et Noël Nguyen, vise à mettre en avant la contribution apportée par le traitement automatique de la parole à l'étude de la variation en français parlé, et il est plus spécifiquement centré sur les voyelles. Différentes sources de variation sont décrites, telles qu'il est possible de les identifier grâce aux systèmes de transcription automatique. Les auteurs décrivent ensuite différentes manières d'utiliser les systèmes de transcription, et notamment des systèmes d'alignement, en tant qu'instruments de mesure pour les recherches en linguistique. La validité des mesures est discutée en lien avec la fiabilité des frontières de segment obtenues par alignement. La durée vocalique est analysée en fonction de différents alignements, de différents styles de parole, de la position de la voyelle dans le mot lexical ou prosodique, de la variante régionale. Enfin, la réalisation des voyelles moyennes est explorée par le biais du traitement automatique.

Table des matières

3.6. Conclusion

Dans ce chapitre, nous avons discuté la constitution et le déploiement d'une grande base de données orales à la lumière de l'expérience acquise au sein du projet PFC. L'un des attraits de ce projet est qu'il a su, dans ses choix méthodologiques, répondre à ses besoins propres tout en rendant possible l'exploitation des données pour des besoins initialement non prévus. Le choix de technologies et formats ouverts et pérennes s'avère ici un élément absolument crucial. Les outils développés dans ce projet permettent une exploration fine et à grande échelle des données.

Il est important de souligner que le processus de constitution d'une base de données, quelle qu'elle soit, se doit d'être un processus itératif : il faut prendre en compte les retours des utilisateurs et participants de manière à maintenir et faire évoluer les conventions. De plus, la nécessité de disposer de données homogènes ne doit jamais être perdue de vue, sans quoi les données peuvent rapidement devenir inexploitables. Le développement d'outils de contrôle, s'il s'avère de prime abord coûteux en ressources humaines, est une étape nécessaire dans un projet de grande taille.

En dernier lieu, il nous semble opportun de souligner que les outils informatiques, aussi puissants soient-ils, ne sont destinés qu'à aider le linguiste et non à s'y substituer. Par exemple, les outils d'analyse de la liaison et du schwa ne livrent pas des réalisations de schwas et de liaisons, mais bien des codages schwa et liaison. Ces codages doivent toujours être interprétés et compris à la lumière des conventions qui les sous-tendent, dans le cadre des limites qu'elles imposent. Malgré ces contraintes, il est clair que les bases de données ont un rôle important à jouer dans le renouvellement des données empiriques et la mise à l'épreuve des modèles théoriques.

Chapitre 4

Les systèmes de transcription automatique de la parole comme instruments de mesure sur les grands corpus oraux

4.1. Introduction

Dans ce chapitre nous allons présenter quelques pistes pour utiliser les systèmes de transcription automatique de la parole comme instruments visant l'acquisition de connaissances phonétiques et phonologiques. Les systèmes de transcription s'appuient sur de très grands corpus de parole et de textes afin de modéliser la langue parlée. Les grands corpus servent à estimer des propriétés moyennes concernant la réalisation acoustique des phonèmes ainsi que des fréquences d'occurrence de mots et de suites de mots avec leurs prononciations. Ces statistiques, incluses dans le modèle de parole, permettent de guider la reconnaissance automatique. Reconnaître un mot revient simultanément à le localiser dans le signal acoustique, à identifier les sons élémentaires qui le composent et à décider de l'identité exacte du mot. Au-delà de leur vocation première de transcription automatique, les systèmes de reconnaissance de la parole peuvent servir, comme instruments pour aider à examiner et à mesurer les réalisations des sons et d'entités plus larges comme les syllabes et les mots. Ainsi, les systèmes de transcription peuvent être adaptés afin d'explorer des corpus virtuellement illimités et d'y étudier des phénomènes précis en ouvrant un champ d'investigations jusque-là inenvisageable. Ils permettent d'extraire des connaissances à partir de corpus contrôlés, tel que le corpus PFC (Phonologie du Français Contemporain) [DUR 03] conçu pour l'étude de la variation régionale. Les systèmes de transcription automatique peuvent

Chapitre rédigé par Martine ADDA-DECKER et Gilles ADDA et Lori LAMEL.

également servir à explorer avec précision ces grands corpus qui ont été collectés pour leur mise au point et leur évaluation. Ces derniers corpus peuvent être qualifiés de faiblement contrôlés, dans la mesure où ils sont composés de parole produite indépendamment de tout objectif d'étude linguistique précis. Ils sont cependant contrôlés par rapport à une situation de communication et à un contexte de production donnés, comme par exemple des corpus de parole radio- ou télédiffusée avec choix de tel ou tel type d'émissions. L'estimation des modèles pour la transcription automatique requiert des corpus les plus larges possibles afin de refléter aussi complètement que possible la langue orale dans la situation de communication visée.

Les instruments d'alignement issus de la transcription automatique de la parole permettent de localiser automatiquement les mots prononcés dans un enregistrement audio transcrit. Ils permettent également d'en déterminer leur prononciation avec les segments de phones correspondants, dans la limite des variantes de prononciation proposées dans le dictionnaire de prononciation, l'une des briques essentielles d'un système de transcription automatique. Ceci constitue le point de départ pour un éventail de mesures globales incluant l'ensemble du corpus ou des sous-ensembles d'échantillons sélectionnés suivant des critères linguistiques précis. Par exemple, on peut mesurer la durée moyenne des réalisations des consonnes sur l'ensemble du corpus, ou juste d'une consonne précise, par exemple le /z/ ou uniquement des segments correspondant au /z/ de liaison. Les mesures peuvent porter, suivant le focus de l'étude, sur différentes grandeurs : segmentales (par exemple durées et taux de voisement des segments, formants de voyelles) ou suprasegmentales (par exemple contours d'intonation et patrons rythmiques).

Avant de se lancer dans des analyses phonétiques à grande échelle, une question cruciale concerne la validité des prononciations produites par alignement automatique. Il est important de réfléchir au statut des étiquettes produites automatiquement au niveau segmental. S'agit-il d'une transcription phonémique ou d'une transcription phonétique ? Que pouvons-nous apprendre de telles transcriptions automatiques ? Est-il nécessaire de les valider par des experts humains ? Si oui, dans quelle situation ? Les réponses dépendent au moins partiellement de la configuration du système d'alignement, de la précision des modèles acoustiques et des prononciations autorisées par le système. En général, les mots de la langue sont décrits dans le dictionnaire de prononciation par leur prononciation canonique (c'est-à-dire incluant tous les phonèmes qui auraient été produits lors d'une articulation soignée). L'instrument produit alors une transcription qu'on peut qualifier de phonémique. En tout cas, les étiquettes produites ne peuvent refléter les variations fines qui seraient observables à un niveau phonétique. Mais alors, quelle est leur validité par rapport au signal observé ? Que valent les frontières de segments ? Si le dictionnaire inclut les variantes phonologiques correspondant aux réalisations phonétiques majoritairement observables dans la langue, les étiquettes peuvent traduire des variations en fonction du locuteur. Dans ce cas, on peut éventuellement parler de transcription quasi-phonétique au moins quand on se propose d'étudier plus particulièrement ces variables phonologiques. Les instruments, suivant

leur réglage, permettent ainsi différents types d'exploration, comme par exemple l'assimilation de voisement des obstruentes, la réalisation du schwa et de la liaison, la dénasalisation des voyelles nasales.

Dans ce chapitre, nous allons nous efforcer d'apporter des réponses aux questions ci-dessus et de présenter quelques exemples d'études à base de grands corpus et d'instruments. Les chapitres suivants de ce volume illustreront plus amplement l'utilisation de systèmes de transcription automatique pour l'étude de la variation régionale (chapitre 5) et l'étude des voyelles (chapitres 6 et 7) . Dans la section 4.2, nous allons donner un survol historique des progrès en reconnaissance automatique de la parole, via la description de quelques corpus-types utilisés au cours du temps. Nous allons décrire l'approche statistique, en développant ensuite les parties essentielles pour l'alignement automatique. Dans une dernière partie 4.3, nous proposons quelques exemples d'utilisation des systèmes de transcription automatique en tant qu'instruments pour des études phonétiques. Nous nous appuyons sur différents corpus oraux, en particulier les corpus PFC [DUR 03] et ESTER [GAL 05]. La précision des segments et de leur étiquetage sera discutée en fonction de la configuration du système, en particulier en fonction du type des modèles acoustiques et des dictionnaires de prononciation. Suivant la configuration, les étiquettes produites peuvent représenter des prononciations canoniques (segmentation phonémique) ou bien refléter des variantes de prononciation (segmentation quasi-phonétique). Nous allons finir ce chapitre par l'illustration des capacités de mesure de l'instrument concernant quelques grandeurs simples, comme la durée des sons et leur fréquence ainsi que des grandeurs plus complexes comme la quantification de la réalisation de variantes de prononciation. La cohérence et la précision de la segmentation temporelle seront également discutées. Le chapitre 7 continuera l'examen de ces questions, en particulier à travers la comparaison de systèmes d'alignement de différents laboratoires francophones.

4.2. Transcription automatique de l'oral

Les progrès en transcription automatique de la parole se sont accompagnés (et même ont été précédés) du développement des corpus de parole. Ces corpus ont acquis au cours du temps une complexité croissante à tous points de vue, y compris linguistique.

En partant de corpus de sons et de mots isolés, les corpus ont progressivement inclus de la parole plus fluide sous forme de lecture (écrit oralisé), d'oral journalistique plus ou moins prompté pour aller actuellement vers des corpus d'oral spontané et interactif.

4.2.1. *Aperçu historique : de l'écrit oralisé vers l'oral*

Les premiers travaux en transcription automatique de la parole peuvent être situés vers le milieu du siècle dernier [DRE 50, DAV 52, WIR 52, TUB 70, DRE 72]. Ces travaux visaient à identifier quelques sons élémentaires à partir de formants et divers traits distinctifs ou bien à reconnaître à partir du signal acoustique un petit nombre de mots différents, prononcés de manière isolée. Jusqu'aux années 70-80, différentes approches étaient proposées, plus ou moins holistiques ou analytiques. La reconnaissance de la parole était vue essentiellement soit comme un problème de *reconnaissance des formes*, soit comme un problème de *système expert*, avec comme questions sous-jacentes : quels paramètres acoustiques pour représenter la parole, quelles unités pour la modéliser (côté reconnaissance des formes), quelles connaissances utiliser pour identifier les réalisations des phonèmes (côté système expert), comment tenir compte du problème des « distorsions spectrales et temporelles » observées entre deux répétitions d'un même énoncé par un même locuteur ? On se rend facilement compte de l'intérêt potentiel de l'approche analytique pour de nombreuses disciplines des sciences du langage et de la parole s'intéressant à la variation et aux invariants : phonétique, phonologie, psychoacoustique et psycholinguistique. Les premiers travaux, loin de viser la transcription automatique de phrases complexes, se limitaient à quelques énoncés simples. Un locuteur, tout au mieux deux ou trois, enregistraient quelques mots ou phrases simples en laboratoire, avec une prédilection pour les chiffres, les lettres et les nombres. En effet la reconnaissance des chiffres et des lettres, tout en se limitant à des vocabulaires très petits, pose les défis scientifiques pertinents de discrimination entre paires minimales et permet déjà d'envisager un certain nombre d'applications comme la reconnaissance automatique des codes postaux ou des numéros de téléphone. La production de corpus de parole restait essentiellement dépendante d'initiatives individuelles de chercheurs.

Au début des années 70 fut lancé le programme américain ARPA-SUR *Advanced Research Project Agency - Speech Understanding Research*. Il est intéressant de rappeler que la compréhension de la parole n'était qu'un volet des programmes de recherche sur la langue qui ont visés en premier lieu la traduction automatique. Les grands laboratoires américains (notamment CMU et BBN) ont participé à ce programme qui visait la compréhension de phrases simples construites avec un vocabulaire d'environ 1000 mots sans trop de spécifications. Alors que les résultats étaient globalement décevants, des conclusions importantes pour la direction des recherches futures au niveau international en ont découlé, même si leur mise en place a pris environ dix ans. Ces conclusions furent (i) qu'il faut séparer la reconnaissance ou la transcription automatique de la compréhension, eu égard à la complexité du problème ; (ii) qu'il faut travailler sur des tâches et des corpus partagés avec une évaluation commune afin de produire des résultats comparables et interprétables. Ces deux conclusions ont profondément influencé l'évolution du domaine. C'est donc à partir des années quatre-vingts, que les premières grandes initiatives de collectes de corpus ont vu le jour. En

1984, l'agence américaine ARPA [1] lance un programme s'attaquant aux problèmes majeurs de la coarticulation et de la variabilité interlocuteur. Il s'agit d'enregistrer des centaines de personnes, hommes et femmes, lisant des phrases garantissant une bonne couverture phonémique. Le but du corpus TIMIT [ZUE 90, GAR 93] est d'étudier cette coarticulation à grande échelle et de permettre d'estimer des modèles acoustiques de phones génériques, capables de représenter n'importe quel vocabulaire et n'importe quel locuteur. Des efforts semblables sont lancés en Europe et notamment en France avec les projets BDSONS [CAR 84] et BREF [LAM 91]. Ce dernier, avec la lecture d'articles du journal *Le Monde*, était déjà clairement tourné vers le domaine journalistique. La table 4.1 rappelle quelques-unes de ces initiatives. Les corpus lus

date	*style*	parole	nom	langue	vol. (h)
		chiffres			
1982	lecture	élicitée (labo)	TI-digits	anglais	12
		phrases phonétiquement équilibrées			
1988	lecture	élicitée (labo)	EUROM	multilingue	10×L
1989	lecture	élicitée (labo)	TIMIT	anglais	8
1989	lecture	élicitée (labo)	BDSONS	français	10
		information trafic aérien			
1990	spontané	élicitée (terrain/labo)	ATIS	anglais	20
		lecture de journaux écrits			
1990	lecture	élicitée (labo)	BREF	français	100
1990	lecture	élicitée (labo)	WSJ	anglais	162
		émissions radio- et télédiffusés			
1996	oral préparé	non-élicitée (terrain)	BN-hub4	anglais	100
2005	oral préparé	non-élicitée (terrain)	ESTER	français	100
2010	oral conversationnel	non-élicitée (terrain)	EPAC, ETAPE	français	>130

Tableau 4.1 – Exemples de corpus de parole créés pour le traitement automatique (TA) de l'oral. La parole des corpus générée spécialement pour le TA, est notée "élicitée"

présentent pour le traitement automatique l'intérêt de rester conforme à l'écrit et de disposer a priori d'une transcription correcte (modulo quelques écarts lors de la lecture) ; on n'est pas ici dans un "vrai" genre oral, mais simplement dans de l'écrit oralisé. Les premiers grands corpus oraux sans modalité écrite préexistante sont enregistrés sur des tâches limitées, visant à implémenter des services téléphoniques de renseignements (ATIS *Air Traffic Information System* aux États-Unis [PRI 90], projets avec la SNCF en France [LAM 99], avec le CNET [COR 97]). De telles tâches

1. ARPA : Advanced Research Program Agency, l'agence américaine de recherche dépendant du département de défense.

impliquent de fait l'usage d'une phraséologie restreinte, même si les locuteurs peuvent s'exprimer librement. En particulier, le vocabulaire reste limité en dehors des entités spécifiques (noms de personne ou noms de ville par exemple). Ces corpus posent le problème de création de transcriptions manuelles de référence à grande échelle, et les travaux des linguistes de l'oral ont été précieux pour mettre au point des conventions de transcription [BLA 99]. Les corpus de parole journalistique, qu'on qualifie d'oral préparé et qui restent proches de la lecture sont les premiers grands corpus de terrain (c'est-à-dire non enregistrés dans les laboratoires de recherche) collectés depuis 1996 aux États-Unis [GRA 02] dont la parole n'a pas été élicitée par des chercheurs à des fins de recherche. Ces types de corpus sont exempts des principaux biais dus à l'implication de l'expérimentateur.

En parallèle, les méthodes statistiques [JEL 76, BAK 75, JEL 98] ont pris leur essor depuis les années soixante-dix et ont acquis une place prépondérante dès la fin des années quatre-vingts pour la reconnaissance automatique de la parole continue. La transcription automatique de la parole a abandonné l'approche analytique ascendante au profit d'une approche globale statistique. L'approche statistique peut s'appuyer sur des corpus de parole et de textes toujours croissants afin d'estimer des modèles de la langue orale. Aujourd'hui, des ordinateurs en réseau extrêmement performants avec des capacités de stockage gigantesques permettent de traiter des corpus volumineux incluant des centaines voire des milliers d'heures de parole. Les avancées théoriques proposées par le cadre statistique ont ainsi pu être mises en œuvre à grande échelle grâce au progrès des ordinateurs et des quantités croissantes de données disponibles. Le paradigme d'évaluation visant à comparer les performances des différents systèmes à travers des campagnes nationales [DOL 97, GAL 05, GAL 09], européennes [YOU 97, MAR 05] et internationales [PAL 03] a fortement contribué à orienter les travaux de recherche dans les directions les plus prometteuses et à stimuler les progrès des systèmes de transcription, grâce notamment à la production de corpus partagés par la communauté scientifique.

La production de corpus pour le traitement automatique s'est faite à grande échelle aux États-Unis et à degrés divers dans les différents pays européens (France, Angleterre, Allemagne, Pays-Bas, Italie, Espagne, Portugal, Grèce... et plus récemment les pays asiatiques). Pour les langues européennes cette production a souvent été soutenue par des projets européens [YOU 97] et transnationaux (corpus CGN pour les néerlandophones de Belgique et des Pays-Bas). Une partie importante de la production de corpus est soutenue par la défense (ARPA *Advanced Research Project Agency* aux États-Unis, DGA*Délégation Générale à l'Armement* en France). Les efforts importants déployés pour la production de corpus, et de manière générale, pour les ressources linguistiques (dictionnaires de prononciation, corpus étiquetés et enrichis avec des classes grammaticales, des entités nommés, corpus alignés multilingues) ont entraîné la naissance d'agences de soutien au développement et à la distribution de corpus. Ainsi en 1992 le LDC (*Linguistic Data Consortium*) a été créé à l'Université de Pennsylvanie, avec le soutien de ARPA et de la NSF (*National Science Foundation*). En 1995 la France a été motrice dans la création

de ELRA (*European Linguistic Resources Association*) à Paris [MAR 99], visant à la validation, la gestion et la distribution de ressources de parole, texte et terminologie. Grâce à ces agences, des centaines, voire des milliers d'heures d'oral préparé journalistique ont pu être transcrites manuellement pour alimenter les recherches. Des logiciels dédiés à la transcription manuelle (p.ex. TRANSCRIBER [BAR 01]), ont été développés en collaboration entre les États-Unis et la France. De nos jours, ce logiciel libre est utilisé bien au-delà du traitement automatique de l'oral par des linguistes de l'oral ou des psycholinguistes s'intéressant par exemple à l'acquisition du langage [MAC 00].

Nous avons donné un aperçu historique des recherches en reconnaissance automatique de l'oral en faisant une part importante au développement de corpus oraux. Nous avons pu remarquer que la démarche dans la collecte des données est différente de celle des disciplines relevant des sciences humaines et sociales s'intéressant à l'oral (phonétique, phonologie, sociolinguistique, psycholinguistique…) [BAU 06]. Pour la reconnaissance automatique de la parole, les corpus sont conçus d'abord afin d'inclure l'ensemble des sources de variabilité de la parole représentatif d'un type de situation de communication et d'un type d'application (par exemple, la transcription automatique d'émissions journalistiques en vue d'un sous-titrage et/ou d'une traduction). Ils ne sont pas conçus pour étudier un phénomène linguistique précis a priori, comme par exemple la réalisation des voyelles moyennes ou l'assimilation de voisement des obstruentes en français, ou de manière plus vaste, les variétés régionales du français. Étant donnés les volumes de parole transcrits et annotés disponibles, il devient cependant souvent possible aujourd'hui d'appliquer à ces corpus issus du traitement automatique de la parole une démarche de linguiste a posteriori : pour une hypothèse à vérifier, on peut construire des sous-corpus en sélectionnant toutes les occurrences intéressantes pour l'étude en question. Ceci ouvre de nouvelles perspectives de recherche à partir de corpus reflétant le phénomène étudié dans l'usage de la langue et sans biais lié au contrôle du phénomène étudié. Au fur et à mesure des progrès accomplis par les systèmes de transcription automatique, les corpus produits revêtent un intérêt grandissant pour des études à caractère linguistique, dans la mesure où le type de parole étudiée rejoint de plus en plus l'oral naturel. À l'inverse, certains grands corpus conçus et collectés par les linguistes sont venus enrichir les ressources disponibles pour le traitement automatique de l'oral [DUR 03, BAU 06].

4.2.2. *Modélisation statistique de la parole*

Pourquoi la conversion du signal acoustique en un flux d'écrit est-elle si difficile et pourquoi une approche analytique, "bottom-up" mettant en oeuvre toutes nos connaissances phonétiques et phonologiques n'a pas permis d'aboutir à des résultats satisfaisants ? Quels problèmes pose donc la parole, ce signal physique continu, quant à sa transformation en flux de mots écrits, signes linguistiques discrets ?

La variabilité du signal de parole contient une partie de la réponse à cette question. Les différents facteurs responsables de cette variabilité sont à peu près connus.
(i) De manière générale il n'y a pas de frontières clairement détectables entre les mots, et les frontières segmentales sont souvent peu évidentes. La réalisation acoustique d'un phonème dépend fortement de son contexte phonémique gauche-droite, ainsi que d'autres facteurs incluant prosodie et fréquence, non seulement du phonème, mais aussi de la syllabe englobante, du mot etc.
(ii) Le signal de parole varie en fonction du locuteur (sexe, âge, santé, émotions, accent, catégorie socioprofessionnelle. . .).
(iii) Les conditions d'enregistrement et le bruit de fond se mélangent au signal de parole dans l'enregistrement.
(iv) Le style d'élocution (lu, préparé, spontané) influe fortement sur le débit, la prosodie, la précision de l'articulation, les variantes de prononciations. De même, la situation (parole publique ou privée ; monologue, dialogue, réunion ; familier ou formel) et l'information partagée entre les protagonistes et/ou portée par le contexte jouent un rôle important sur le choix des mots et leur réalisation articulatoire et acoustique. Les lecteurs, qui se sont exercés à la lecture de spectrogrammes, ont pu se rendre compte à quel point un bout de signal de parole peut être ambigu et que son interprétation dépend non seulement de ses qualités propres mais aussi du contexte dans lequel ce bout de signal s'insère. Nous allons voir comment ces deux termes : *qualités propres* et *contexte* sont pris en compte de manière théoriquement élégante dans l'approche statistique.

La transcription automatique de la parole repose sur une modélisation statistique de la langue développée depuis les années soixante-dix [BAK 75, JEL 76, JEL 98] . Transcrire un signal de parole consiste à trouver la suite de mots la plus probable $\hat{m}$ étant donnés le signal ou l'observation acoustique x. Ceci peut s'écrire :

$$\hat{m} = \arg\max_{\{m\}} P(m/x)$$

où $\{m\}$ désigne l'ensemble des séquences de mots m possibles et $\arg\max$ désigne l'opération qui choisit dans cet ensemble l'argument (c'est-à-dire la séquence) ayant la probabilité maximale. Cette première équation, qu'on ne sait résoudre directement, se réécrit grâce à la **formule de Bayes** :

$$\hat{m} = \arg\max_{\{m\}} P(m/x) = \arg\max_{\{m\}} P(x/m)P(m)$$

Nous avons à droite de notre équation une formule contenant trois parties, résumant la transcription automatique : $P(x/m)$ concerne la modélisation acoustique des mots de la langue, $P(m)$ le modèle permettant de générer la langue (sous forme de séquences de mots) et $\arg\max_{\{m\}}$, le décodeur.

Le terme $P(x/m)$ (la probabilité du signal de parole x étant donnée une suite de mots m) correspond à la modélisation acoustique et peut être vu comme la contribution des *qualités propres* de x étant donnée l'hypothèse m. Le modèle acoustique

permet d'évaluer la vraisemblance d'observer le signal acoustique x, en faisant l'hypothèse que les mots m ont été prononcés. Cette probabilité est estimée grâce au modèle acoustique sous forme de modèles de Markov cachés ou HMM (*Hidden Markov Models*) [RAB 89]. $P(x/m)$ est évaluée pour toutes les hypothèses m envisageables. Celles-ci dépendent du vocabulaire de mots connus du système. Le terme de modélisation acoustique $P(x/m)$ met en lien l'observation acoustique x et le (ou les) mot(s) m. À ce stade, il ne fait pas apparaître les prononciations des mots. $P(x/m)$ peut être développé afin d'expliciter les prononciations Φ :

$$\hat{m} = \arg \max_{\{m,\Phi\}} P(x/\Phi)P(\Phi/m)P(m)$$

Les prononciations Φ sont spécifiées dans le dictionnaire de prononciation (voir section 4.2.4).

Le terme $P(m)$ permet d'estimer la probabilité a priori de la séquence de mots m, On parle de modèle de langue (ou de modèle de langage), qui se présente le plus souvent sous forme de collection de n-grammes de mots $P(m_t/m_{t-1}m_{t-2}\ldots m_{t-n+1})$ [BAH 89, JEL 91] : la probabilité du mot à l'instant t (m_t) est prédit étant donné son *contexte* (son historique), c'est-à-dire les $n-1$ mots qui le précèdent ($m_{t-1}m_{t-2}\ldots m_{t-n+1}$). Le modèle de langue permet de tenir compte du *contexte* linguistique lors du décodage acoustique : il fournit des probabilités de cooccurrence de mots pour chaque n-uplet de mots. Plus une séquence est probable, et moins elle nécessite une réalisation acoustique précise pour être correctement reconnue par un système de transcription automatique. Considérons comme exemple deux séquences phonétiquement très proches, `moyens du bord` et `moyens du port`. D'un point de vue acoustique, elles ne diffèrent que par le trait de voisement de la consonne plosive labiale (/b/ au lieu de /p/) du deuxième nom. La première séquence étant a priori nettement plus probable que la deuxième dans la langue française (une simple recherche GOOGLE en novembre 2012 donne 700 000 occurrences pour `moyens du bord` contre 3000 pour `moyens du port`), un locuteur pourra prononcer le /b/ comme un [p] sans être nécessairement sanctionné par le système de transcription automatique. En revanche, un manque de précision de phonation en sens inverse (un /p/ prononcé comme [b]), sera rapidement sanctionné. Cet exemple sert à illustrer l'importance du contexte large dans l'interprétation d'un son élémentaire : il n'y a pas que ses qualités intrinsèques (présence/absence de voisement) qui contribuent à l'identifier (/b/ ou /k/), mais tout le contexte lexical est mis à contribution. Les modèles n-grammes permettent a priori de générer l'ensemble des phrases de la langue (si on fait l'hypothèse que le vocabulaire de la langue est fini et connu du système), mais malheureusement bien au-delà, une infinité de phrases incorrectes d'un pont de vue humain. Leur point fort est de donner des probabilités nettement plus élevées à des séquences de mots fréquentes, qu'à des séquences rares. La fréquence favorise globalement la construction de phrases jugées correctes (ou localement correctes), sans pour autant garantir la génération de phrases correctes. Ces modèles n-grammes (chaînes de Markov, le plus souvent avec n=3 ou 4) correspondent à des grammaires locales reflétant implicitement des

niveaux syntaxiques, sémantiques et pragmatiques de la langue, ou plus précisément la syntaxe, sémantique, pragmatique empiriquement contenues dans les corpus à partir desquels ils sont estimés. Ainsi, des suites de mots très fréquentes dans la langue, comme par exemple `de la` (comme dans `loin de la Bretagne, ministère de la Défense, annexe de la Sorbonne, de la même façon, j'écoute de la musique`...) auront une probabilité d'apparition extrêmement élevée, alors que des séquences a priori très peu probables comme  resteront néanmoins possible (comme dans `donner le la à l'instrumentiste`) avec une probabilité très faible. Plus une séquence de mots est fréquente, plus sa probabilité donnée par le modèle de langue est élevée. Plus le modèle de langue favorise un mot ou une séquence de mots, plus sa réalisation acoustique peut s'écarter de la "référence", sans pour autant entraîner une erreur de décodage. Le modèle de langue ne donne pas l'ensemble des suites de mots possibles ou impossibles, mais la *prééminence* d'une suite de mots sur une autre suite de mot. Cette propriété se montre très appropriée pour la reconnaissance de la parole, dans la mesure où on peut observer plus de variation sur les mots les plus fréquents.

L'opérateur $\arg\max$ de l'équation correspond au décodeur dont le rôle est d'évaluer la probabilité de l'ensemble des suites de mots m étant donné le signal acoustique inconnu x (et de retenir la meilleure solution $\hat{m}$) en combinant les deux termes $P(x/m)$, la contribution liée aux *qualités propres* du signal acoustique, et $P(m)$, la contribution liée au *contexte lexical*. Cette combinaison permet d'éviter des décisions locales prématurées en faveur d'un phonème ou d'un mot et permet de décider des phonèmes et des frontières de segment en tenant compte du contexte lexical. Il est important de retenir que le sens du mot "contexte" est ici bien plus large que le contexte pour les modèles acoustiques (p.ex. triphones, où "contexte" se limite aux phonèmes voisins du segment modélisé). La recherche de la meilleure solution repose sur le principe de programmation dynamique [BEL 57, VIN 68, FOR 73] en général implémenté dans une approche temps-synchrone [NEY 84]. La segmentation en mots et en phones sera un sous-produit de la recherche de la meilleure solution globale $\hat{m}$ dans l'espace de recherche correspondant à l'ensemble des hypothèses m possibles. Les systèmes d'alignement peuvent être vus comme des décodeurs dégénérés où l'espace de recherche est limité à la seule suite de mots m_{ref} correspondant à la transcription manuelle de référence. Dans la transcription imposée, le modèle de langue est inutile, la séquence de mots étant imposée.

L'estimation des modèles statistiques, à la fois acoustique et linguistique, est faite à partir de corpus de parole acoustique et de corpus textuels, ces derniers provenant en grande partie de sources écrites, mais également de parole transcrite. Les modèles censés représenter la langue de manière générale, représentent avant tout le contenu des corpus utilisés lors de leur estimation.

4.2.3. *Modélisation acoustique*

Nous allons examiner quelques aspects de l'analyse et de la modélisation acoustiques permettant de mieux comprendre les possibilités et les limites des systèmes de transcription et d'alignement concernant en particulier la localisation de frontières segmentales.

(i) Le signal de parole, typiquement échantillonné à 16 kHz, est transformé en une suite de vecteurs, où un vecteur est généré toutes les n millisecondes. La plupart des systèmes travaillent avec un pas de $n = 10$ ms (une seconde de parole correspond alors à 100 vecteurs), même s'il existe des systèmes calculant un vecteur toutes les 8 ou 5 ms [LUD 12]. Les vecteurs acoustiques sont calculés sur des fenêtres temporelles d'environ 30 ms, afin de capter des propriétés stationnaires des sons de parole, et en particulier, d'inclure plusieurs cycles de la fréquence fondamentale pour les sons voisés (voir figure 4.1.

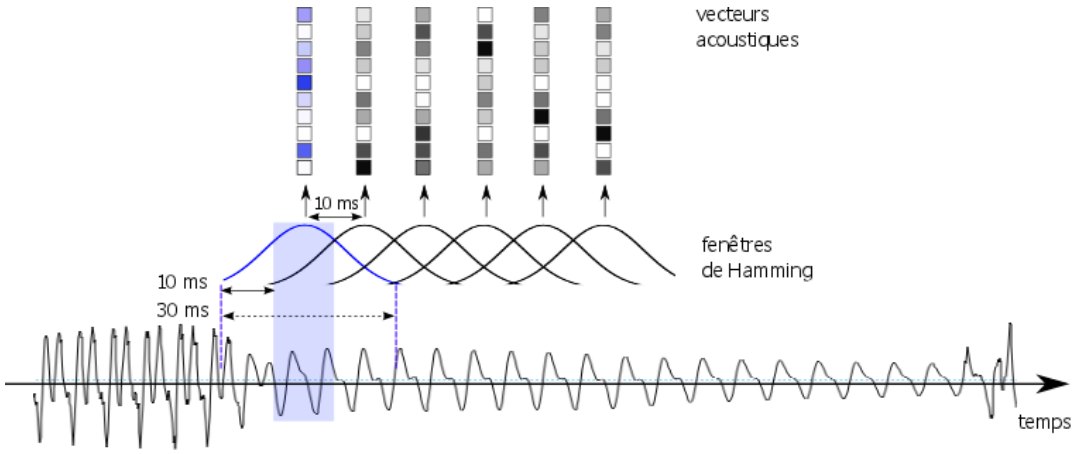

Figure 4.1 – Paramétrisation du signal acoustique : un vecteur acoustique est calculé toutes les 10 ms à partir de portions de signal découpées par des fenêtres de Hamming de 30 ms. Les vecteurs représentent essentiellement la partie centrale de la fenêtre (en grisée).

(ii) La paramétrisation généralement retenue correspond à des vecteurs de coefficients cepstraux distribués suivant une échelle quasi-logarithmique (échelle Mel) qui traduit la sensibilité de l'oreille humaine. Les coefficients cepstraux ou coefficients MFCC (*Mel Frequency Cepstral Coefficients*) [DAV 89] reflètent le spectre de puissance du son à un instant donné. À ces paramètres « statiques » sont rajoutés des

paramètres dynamiques (coefficient delta d'ordre 1 et 2), qui mesurent la vitesse et l'accélération des changements du conduit vocal dans le temps.

(iii) les modèles acoustiques ne représentent pas directement les mots, mais des sons élémentaires (voir figure 4.2) ou phonèmes [2]. Les phonèmes sont modélisés par des modèles de Markov cachés à trois états. Chaque état est décrit par un ensemble probabilisé de vecteurs de coefficients cepstraux (via des densités gaussiennes) décrivant les réalisations possibles pour cette partie de phonème. La durée minimale d'un segment de phone 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 (c'est-à-dire un vecteur acoustique de 10 ms). Les gaussiennes sont estimées à partir de corpus d'apprentissage, segmenté au préalable utilisant le dictionnaire de prononciation. Afin de ne pas mélanger toutes les sources de variabilité acoustique énumérées auparavant, des jeux de HMM spécifiques sont estimés en fonction des caractéristiques des locuteurs (par exemple hommes, femmes) et en fonction de la qualité des enregistrements (bande large, bande téléphonique) [GAU 94, GAL 95, MAT 09, WAN 11].

Les modèles acoustiques de sons élémentaires vont servir comme dans un jeu de construction à assembler les modèles de mots. Pour cela, il faut commencer à définir un inventaire phonémique, qui peut s'écarter légèrement du jeu de phonèmes classiquement admis dans la langue. Pour le français, par exemple, l'opposition entre les deux nasales /ɛ̃/ et /œ̃/ n'est plus/peu tenue. Concernant les voyelles centrales /ə/, /œ/ et /ø/, le schwa, voyelle centrale à réalisation optionnelle, peut-être fusionné avec l'une ou l'autre des voyelles /œ/ et /ø/. On peut ou non garder l'opposition entre les deux /o/ et /ɔ/. Ces simplifications n'entraînent pas de changements significatifs dans les résultats de transcription. En revanche, si on veut faire des analyses phonétiques à partir d'un alignement automatique, il faut connaître les choix opérés pour l'inventaire du système et éventuellement procéder à des corrections. Le choix de l'inventaire phonémique peut dépendre de la fréquence des différents sons dans la langue ou dans les corpus disponibles pour l'apprentissage et de leur rôle dans les prononciations. Si deux sons proches peuvent alterner dans les prononciations en fonction du locuteur ou de la position prosodique par exemple, on peut envisager de fusionner ces deux sons dans l'inventaire. De même, si un phonème est peu fréquent dans le données d'apprentissage, il serait modélisé de manière peu fiable et dans ce cas, il peut être utile de le remplacer par un phonème proche.

Afin de mieux tenir compte de la variabilité des segments en fonction de leur contexte de réalisation, des modèles acoustiques sont estimés non seulement pour

2. Terminologie : phonème vs phone vs allophone : un modèle acoustique élémentaire contribue à modéliser un phonème de la langue. Les réalisations acoustiques correspondant à ce phonème sont appelés phones. Les modèles contextuels (par exemple triphones) peuvent être vus comme des modèles allophones du phonème en question.

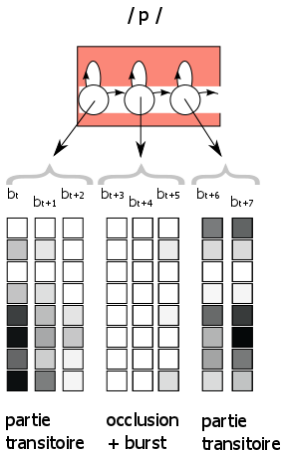

Figure 4.2 – Modèle du phonème /p/ (modèle de Markov caché à trois états). Exemple de génération d'un segment [p] composé d'une partie transitoire liée au contexte gauche, d'une partie centrale (occlusion et burst) et d'une partie transitoire liée au contexte droit.

chaque phonème de la langue, mais aussi pour chaque phonème en contexte gauche et droite. On parle alors de modèles triphones (leur empan temporel reste homogène à des segments de phones). Les jeux de modèles acoustiques triphones incluant souvent des (dizaines de) milliers de modèles élémentaires, visent une modélisation précise de l'ensemble de l'espace acoustique de la langue. Ils permettent de représenter les allophones d'un même phonème. Les modèles acoustiques de mot sont obtenus par combinaison de modèles élémentaires en fonction des prononciations spécifiées dans le dictionnaire de prononciation.

4.2.4. *Dictionnaires de prononciation*

Le dictionnaire de prononciation sert à déterminer, pour les mots qui y sont inclus, leur modélisation acoustique. Le dictionnaire indique pour chaque mot une prononciation qualifiée de standard ou prononciation canonique complète telle que produite par une articulation soignée en mode isolé. Un modèle acoustique de mot est alors obtenu en concaténant des modèles HMM élémentaires correspondant aux phonèmes de cette prononciation (voir figure 4.3). Chaque mot peut éventuellement être complété par des variantes de prononciation (par exemple tenant compte d'élisions de schwa, de liaisons, de formes réduites...). La table 4.2 donne un extrait de dictionnaire de prononciation avec des exemples de variantes. Pour mieux tenir compte de phénomènes

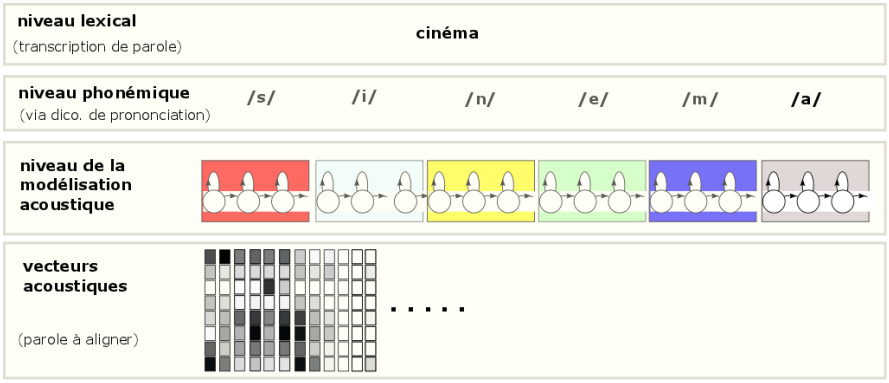

Figure 4.3 – Modèle acoustique du mot `cinéma` obtenu par concaténation de HMM élémentaires correspondant aux phonèmes de la prononciation du mot telle que décrite dans le dictionnaire de prononciation.

Entrée lexicale	Prononciation canonique	Variantes	Commentaires
le	lə	l	élision du schwa
trouble	tʁublə	tʁubl tʁub(C)	simplification de cluster consonantique
dix	dis	di(C) diz(Vn)	finale muette, finale assimilée
est	ɛ	ɛt(V) e et(V)	liaison, fermeture de la voy. mi-fermée
multi-mots			
je_suis	ʒəsɥi	ʃɥi	élision schwa+assimilation
je_crois_que	ʒəkʁwakə	ʃkʁwak ʃʁwak	métaplasme

Tableau 4.2 – Extrait d'un dictionnaire de prononciation avec dans la première colonne de gauche la forme lexicale fléchie, dans la deuxième colonne la prononciation canonique, dans la troisième colonne des variantes de prononciation. Entre parenthèses sont indiquées des contraintes contextuelles : C=consonne, V=voyelle.

de parole spontanée, telles que les *métaplasmes* [ADD 08, MEU 12b] avec des réductions fortes impliquant souvent plusieurs mots, on peut introduire des *multi-mots* avec des pronociations spécifiques (cf. tableau 4.2 en bas).

Comme nous venons de le voir, la variabilité liée à la coarticulation est gérée par des unités contextuelles : chaque phonème de la langue sera modélisé par un ensemble de HMM distincts représentant des allophones de ce phonème dans différents

contextes phonémiques gauche-droite (triphones : un phonème à gauche et à droite ; quinphones : deux phonèmes à gauche et à droite).

La figure 4.3 illustre comment le modèle acoustique du mot `cinéma` est construit par concaténation de n=6 modèles HMM élémentaires pour les 6 phonèmes consécutifs /s/, /i/, /n/, /e/, /m/, /a/. Cette modélisation impose une durée minimale aux mots ainsi qu'aux phones que le système peut localiser dans le signal acoustique : une prononciation à n phonèmes sera représentée au niveau acoustique par une chaîne à $3 \times n$ états imposant une durée minimale de $3 \times n \times t$ ms à chaque observation du mot, et une durée minimale de $3 \times t$ ms pour chaque segment de phone. Avec t=10 ms, la durée minimale d'un phone est de 30 ms et du mot `cinéma` 180 ms.

Différentes options peuvent être prises pour le **dictionnaire de prononciation** : canonique ou incluant des variantes. Si le dictionnaire de prononciation contient des variantes, le modèle acoustique de mot, linéaire dans le cas de la figure 4.3, devient un graphe contenant toutes les variantes. Pour la reconnaissance automatique de la parole, les dictionnaires de prononciations sont plutôt de type phonémique, n'incluant que peu de variantes. Dans cette configuration, on ne peut pas espérer effectuer automatiquement un étiquetage phonétique précis. La variabilité acoustique observée est alors modélisée de manière implicite via les mélanges de gaussiennes dans les modèles de Markov cachés ; de ce fait, la segmentation et l'étiquetage correspondants sont à interpréter à un niveau phonémique plus que phonétique.

Prenons comme exemple de variante de prononciation, un changement de timbre d'une voyelle (lié par exemple à l'harmonie vocalique ou aux accents régionaux). La variation observée sera modélisée de manière implicite par un sous-ensemble de gaussiennes du HMM en question. Sans variante de prononciation prévue dans le dictionnaire, un changement de timbre dans le signal ne pourra pas induire un changement d'étiquette lors de l'alignement du mot avec le signal. L'exemple du changement de timbre peut être considéré comme variante parallèle : elle n'affecte pas la structure temporelle (ou la topologie) du modèle acoustique. Intéressons-nous maintenant à une variante séquentielle, c'est-à-dire impliquant un changement de la structure temporelle de la prononciation. Le *e-muet* ou *schwa* en français en est l'exemple-type. Sa nature instable (voir section 4.3.3, propriété (1)) voudrait que tous les schwas soient rendus optionnels dans le dictionnaire de prononciation. Cependant, afin de limiter le risque d'erreur de reconnaissance, la chute du schwa n'est en général pas prévue dans les mots monosyllabiques dont le noyau est un schwa (essentiellement les mots grammaticaux *de, le, que, ne, se, ce, je, me, te*, cf. Fig. 4.5) qui se réduiraient sinon à une seule consonne (homophone des entrées lexicales *d', l', n'*...), trop facilement insérables sans contrainte forte du modèle de langue n-gramme (pour *d', l', n'*..., la prononciation du mot suivant doit commencer par une voyelle, ce qui n'est pas le cas pour *de, le, ne*...). Les variantes peuvent également être limitées pour de simples raisons de complexité de calculs, en particulier si la variante concerne la frontière de mot.

Ainsi le schwa optionnel en fin de mot polysyllabique est pour cette raison traditionnellement omis. Pour résumer, 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. Concernant le schwa, cela implique qu'un modèle acoustique de consonne en fin de mot peut modéliser non seulement la consonne en question, mais également une voyelle épenthétique (cf. figure 4.8 pour le /d/ final de *Bagdad* réalisé comme [bagdadə]). De manière générale, plus la réalisation acoustique des mots s'écarte de la chaîne linéaire proposée par le dictionnaire de prononciation lors de l'apprentissage des modèles, moins les modèles acoustiques reflèteront 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 causes qui perturbent la modélisation acoustico-phonémique recherchée.

4.2.5. *Variantes de prononciation & Loi de Zipf*

Pourquoi ne pas proposer un grand nombre de variantes de prononciation dans le dictionnaire de prononciation ? Comme nous l'avons déjà évoqué, un nombre élevé de variantes augmente le risque d'homophonie entre différents mots de la langue. Face à des options homophones, le système de transcription automatique tend à produire systématiquement le mot-homophone le plus fréquent, celui qui est favorisé par le modèle de langue. Pour la transcription automatique, l'ajout de variantes de prononciation dans le dictionnaire doit être conditionné sur sa fréquence d'occurrence dans de grands corpus. Or, avant de s'intéresser à des fréquences d'occurrences des prononciations des mots, on peut déjà s'interroger sur la fréquence d'occurrence des mots tout courts. Concernant la fréquence d'apparition des mots d'une langue, la loi de Zipf nous apprend que la fréquence d'occurrence d'un mot, en fonction de son rang de fréquence est une loi exponentielle. Plus simplement, ceci revient à dire que la langue possède peu de mots très fréquents et un très grand nombre de mots rares. La loi de Zipf peut se vérifier sur différentes langues. La figure 4.4 montre, pour le français, l'allemand et l'anglais, les comptes d'occurrence des cent mille mots (types en formes fléchies) les plus fréquents triés par rang de fréquence (figure à gauche). L'axe x correspond donc au lexique des 100k mots-types les plus fréquents de la langue et l'axe y (en échelle logarithmique) nous donne le nombre d'occurrences pour chaque mot-type. Les comptes proviennent de corpus de 30 millions de mots (*tokens*) pour chaque langue. La figure à droite, avec une échelle logarithmique sur x, illustre la loi de Zipf pour ces trois langues. Cette loi nous apprend donc que sur des corpus de parole correspondants à 30 millions de mots (environ 3000 heures en appliquant la règle approximative de 10k mots dans une heure de parole bien remplie) on pourrait avoir au moins cent répétitions de chaque mot pour les dix mille mots les plus fréquents, et ainsi estimer des probabilités d'observation des variantes de prononciation.

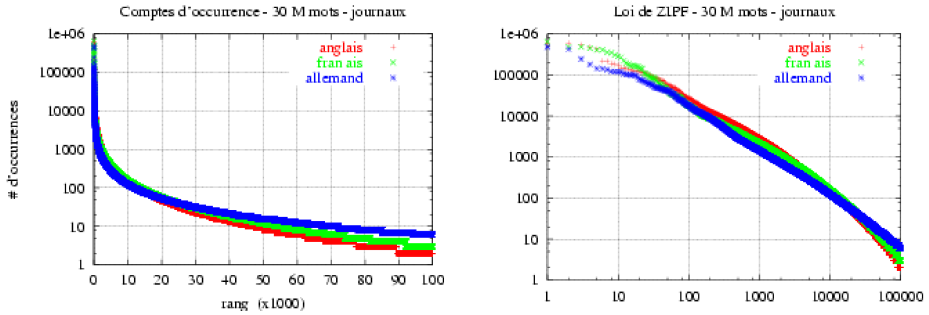

Figure 4.4 – La distribution des mots dans la langue suit une loi de Zipf.

Cet exemple à partir de corpus de textes illustre la difficulté d'estimer des probabilités pour les variantes de prononciation pour tous les mots de la langue à l'exception des n premiers, où n se limite en général à quelques centaines de mots pour quelques dizaines d'heures d'oral transcrit.

Nous avons profité de la disponibilité de corpus dans différentes situations de communication, pour examiner l'ordre relatif des mots fréquents en prenant comme cas d'étude les mots monosyllabiques se terminant par schwa. La fréquence des mots est montrée dans la figure 4.5 pour un corpus de transcriptions de journaux et des transcriptions d'entretiens et conversations provenant du corpus PFC [DUR 03]. Nous avons ajouté sur les deux graphiques les mots grammaticaux monosyllabiques ayant un schwa comme noyau vocalique. Ces mots sont très fréquents dans les deux types de corpus, même si leur ordre d'apparition est marqué par le type de corpus. Ainsi, on peut observer que le mot *je* a un rang de fréquence inférieur à 10 pour le corpus d'entretiens PFC, alors que pour les données journalistiques, le *je* est nettement moins fréquent. Cet écart de rang entre styles de parole est encore bien plus important pour le mot *te*. Les mots grammaticaux comme *de, le, que* sont très fréquents dans les deux corpus, *de* restant le plus fréquent pour les deux situations examinées.

La loi de Zipf nous enseigne que modèles et études linguistiques fondés sur grands corpus permettent de rendre compte en premier lieu des mots les plus fréquents. Par exemple, il sera toujours difficile d'estimer des probabilités de variantes de prononciation pour la majorité des mots de la langue. En effet, les graphiques de la figure 4.5 montrent qu'avec un corpus de 100 heures de parole (à gauche), tous les mots au-delà du rang 10000, sont observés moins de 10 fois. En particulier, pour les entretiens et conversations du corpus PFC exploités ici (à droite), avec 25 heures d'une dizaine de points d'enquête, moins de 2000 mots présentent plus de 10 répétitions dans le corpus et seulement 200 mots admettent plus de 100 occurrences. La grande majorité des

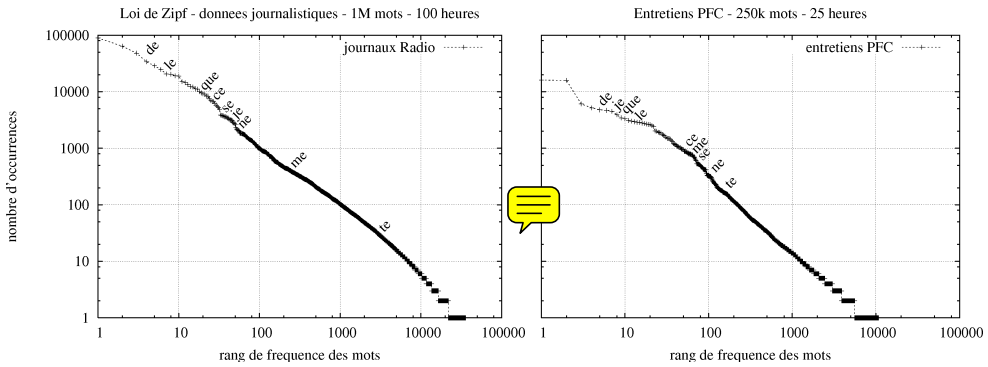

Figure 4.5 – Loi de Zipf - Fréquence d'occurrence des mots en fonction de leur rang de fréquence à partir de corpus de données journalistiques de radio (gauche) et un sous-ensemble d'entretiens et conversations du corpus PFC (droite). Les échelles sont logarithmiques.

mots reste peu observée. Mais avec la taille des corpus toujours croissante, on peut cependant envisager de filtrer ou sélectionner des sous-ensembles. Des approches et techniques spécifiques doivent être développés afin d'étudier des événements rares ou des phénomènes particuliers, sans les noyer dans la masse des événements prédominants dans la langue.

4.2.6. *Précision de la transcription automatique*

Nous allons présenter ce que nous entendons par "précision de l'instrument" de transcription automatique, précision que nous allons discuter en particulier en termes d'erreurs de transcription de mots. Nous allons essayer de relier, le cas échéant, ces erreurs à des manques de précision dans les prononciations prévues ou/et dans les modèles acoustiques de la parole.

Les principales avancées de la recherche en transcription automatique de la parole sont rythmées par des campagnes d'évaluation. Lors des évaluations internationales ou nationales, il s'agit de minimiser des taux d'erreurs sur des données communes envoyées dans les différents laboratoires participants, par des organismes indépendants : le NIST (*National Institute of Standards and Technology*) pour les évaluations ARPA américaines depuis plus de 15 ans, le TNO (*Technical National Office*, Soesterberg, Pays-Bas) dans les évaluations du projet européen LE-*Sqale* autour de 1995, l'ICP de Grenoble lors de la première évaluation francophone en 1997 (avec le soutien de l'*Aupelf*), la Direction Générale à l'Armement (DGA) lors des évaluations ESTER (en 2005 et 2008)

dans le cadre de l'appel à projets *Technolangues* du Ministère de la Recherche avec le concours d'ELDA (Evaluations and Language resources Distribution Agency). Depuis quelque temps, le LNE (Laboratoire National de métrologie et d'Essais) est devenu l'acteur national pour l'évaluation des technologies de la langue. Il s'agit, d'évaluation en évaluation, de démontrer les progrès obtenus dans les méthodes de décodage et de modélisation de la parole, par des taux d'erreurs de plus en plus faibles, sur des styles de parole de plus en plus complexes. L'étude des erreurs de transcription, qui accompagne ces évaluations, vise à identifier des points faibles de l'ensemble {parole, système de transcription}. Les erreurs de transcription sont souvent dues à des écarts entre l'observation et les modèles : l'écart peut provenir d'erreurs de production de la part du locuteur, mais le plus souvent les erreurs de transcription automatique sont dues à une modélisation incomplète. Ainsi, les erreurs de transcription pointent souvent sur des phénomènes linguistiques pour lesquels les modélisations sont insuffisantes ou pour lesquels les connaissances associées ne sont que partiellement décrites et quantifiées. La table 4.3 donne un exemple d'erreurs de transcription automatique de parole journalistique, illustrant des problèmes de transcription typiques du français. Beaucoup d'erreurs en français sont simplement dues à des mots ou séquences de mots

	Exemple de parole transcrite									
transcription manuelle de référence :	*l'*	*aggrave*		*et*	*peut*		*le*	*tuer*		
transcription du système (hypothèse) :	*l'*	*aggraver*			*Paul*			*tués*		
types d'erreur :	-	S		O	S		O	S		

Tableau 4.3 – Exemple d'erreurs de transcription automatique extrait du corpus ESTER illustrant des erreurs liées à l'homophonie. (S : substitution ; O : omission).

homophones (notamment l'accord en genre et en nombre, la conjugaison des verbes), pour lesquels la modélisation acoustique n'est pas vraiment à mettre en cause. Pour éviter ce genre d'erreur il faudra améliorer la composante $P(m)$ (modèle de langue n-gramme). Par exemple, l'introduction d'informations morpho-syntaxiques permet de produire de légères réductions du taux d'erreur [HUE 10]. D'autres erreurs proviennent d'un décalage entre le modèle acoustique et la prononciation, décalage qui peut être dû à la non-production de la part du locuteur de segments acoustiques attendus par le modèle et vice-versa.

Parmi les erreurs de transcription liées au schwa, certaines peuvent s'expliquer par un schwa prévu dans la prononciation, mais non réalisé (comme par exemple la chute du schwa dans la séquence de mots grammaticaux `de la` réalisée comme `d'la`) mais également par un schwa réalisé et non prévu par le modèle acoustique du mot en question, entraînant soit une erreur d'insertion ou une confusion. Il s'agit essentiellement de voyelles épenthétiques produites en fin de mot pour des raisons phonotactiques (dont la loi des trois consonnes) ou de relâchement épenthétique après un

appui final. Les exemples d'erreurs de la table 4.4 provenant de la campagne ESTER donnent d'abord l'extrait de référence avec la partie problématique soulignée, suivi de la transcription automatique correspondante, avec les erreurs en gras. Un schwa pré-

Référence manuelle	**Transcription automatique**	**Commentaire**
I. Schwa absent de la modélisation **mais produit par le locuteur**		
en fait	en fait **de**	relâchement...
le week-end pascal	le week-end pascal **le**	... en fin de groupe
Marc Blondel	**marque** Blondel	loi 3 consonnes
quatrième round prévu	quatrième **rang de** prévu	+ mot étranger
II. Schwa présent dans la modélisation **mais non produit par le locuteur**		
tout le temps	tout _ temps	locution fréquente
temps de leur installation	temps _ leur installation	locution composée
quai de Seine	**quête saine**	+ assimilation
c'était le même marasme	c'**est elle** même marasme	homophone
cette période de troubles	cette période _ trouble	homophone'
appréciable le tandem	appréciable _ tandem	homophone'

Tableau 4.4 – Exemple d'erreurs de transcription automatique sur des données journalistiques. Pour les exemples de la partie I., il n'y a pas de modèle de schwa dans la modélisation alors qu'il a été produit par le locuteur, et l'inverse pour la partie II. La notation « homophone' » désigne des prononciations raccourcies qui restent homophone via des variantes.

sent dans le dictionnaire mais non-réalisé engendre en général une erreur d'omission ou une confusion. Ce dernier type d'erreur concerne essentiellement les mots grammaticaux en position intra-groupe (prosodique, sémantique). La consonne du mot grammatical concerné est elle-même très souvent altérée par assimilation ou carrément absente. D'autres erreurs, comme la confusion entre les séquences `cette période trouble` et `cette période de troubles`, pouvant être réalisées comme homophones, peuvent s'expliquer par des prononciations rapides (au sens que tous les segments prévus par le modèle ne sont pas nécessairement articulés) et le système préfère la variante avec le moins d'états dans le modèle. plutôt par les fréquences a priori des suites de mots qui vont en faveur de la dernière solution. Enfin, très peu d'erreurs sont dues à des schwas en position interne de mots.

Sur des données journalistiques, les taux d'erreur de mot sont souvent inférieurs à 10%. Ce genre de parole, généralement préparé, reste proche d'un écrit oralisé. Pour les conversations téléphoniques en français, qui à l'inverse sont un vrai genre oral, les taux d'erreur sont autour de 30%. Certes les conditions acoustiques sont moins

bonnes et contribuent à augmenter les erreurs, mais les problèmes essentiels pour la parole conversationnelle concernent à la fois l'estimation d'un modèle de langue approprié au genre traité, et les prononciations des mots avec la modélisation acoustique associée. Des problèmes supplémentaires concernent l'établissement d'une transcription de référence dans des zones de parole disfluente ou simplement mal articulée, et le cas de locuteurs multiples (parole superposée).

4.3. Instrument de mesure

Afin d'illustrer les possibilités des systèmes de transcription comme instruments de mesure [ADD 06, HAB 05], nous présentons quelques études s'appuyant sur les corpus oraux collectés pour le traitement automatique de l'oral. Même si les corpus ont pu être collectés avec des critères où les exigences technologiques ont primé sur des considérations linguistiques, des études quantitatives issues de ces corpus peuvent venir compléter et affiner nos connaissances en phonétique, phonologie, prosodie et dresser un tableau plus précis de la variation à l'oral. Si de tels grands corpus d'oral transcrit et balisé temporellement, permettent de faire de nombreux types d'études en phonétique expérimentale afin de valider des théories et modèles existants, il est important d'inclure les corpus conçus et collectés par des linguistes. Notamment, le corpus PFC [DUR 03], qui rassemble des centaines d'heures de parole (lecture, entretiens et conversations) collectés 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 [Bou 02, WOE 09] (voir chapitre 5).

Comme évoqué dans les sections précédentes, les erreurs de transcription automatique peuvent pointer sur des insuffisances de la modélisation des prononciations : oubli dans le dictionnaire de prononciation d'une liaison, d'une forme contractée usuelle (par exemple [ot] pour `autres`) ou de variantes liées à un accent régional ou étranger. Les travaux en traitement automatique de l'oral menés sur différents corpus (présentation de journaux, interviews, entretiens, conversations) montrent que des modèles acoustiques performants pour un type de données, peuvent perdre de leur acuité face à des données d'un autre type. Ces observations suggèrent que les réalisations des sons changent non seulement en fonction de contextes phonémiques gauche et droit, mais aussi suivant d'autres facteurs que nous qualifions de manière approximative de *styles* de parole. L'estimation de nouveaux modèles acoustiques permet alors d'inclure implicitement une grande partie de la variation systématique, et de réduire des taux d'erreurs en transcription automatique. Même si ceci est une méthode efficace pour la transcription automatique, il est cependant important de mieux connaître les mécanismes de variation sous-jacents : on peut supposer que les prononciations changent avec le style de parole (lecture ou spontané, parole publique ou privée). Comme premier pas dans cette direction, nous allons nous intéresser aux changements des durées segmentales obtenus par alignement automatique. Afin d'avoir une idée de la précision de la mesure, nous allons d'abord nous intéresser aux durées segmentales moyennes

en fonction de différentes configurations du système pour un corpus de données journalistiques avant d'examiner ces valeurs sur d'autres corpus.

4.3.1. *Segmentation, frontières et durées*

Les systèmes d'alignement automatique sont souvent critiqués pour leur localisation approximative de frontières de segments. Cette critique est fondée, car les critères utilisés pour poser des frontières ne reposent pas sur des caractéristiques acoustiques locales précises telles celles exploitées par un expert humain, mais sur une optimisation globale du modèle acoustique de l'énoncé face à l'observation acoustique correspondant à cet énoncé. Il en résulte des décalages entre segmentation automatique et segmentation manuelle, voire entre segmentations automatiques issues de différents systèmes (mettant en œuvre différents modèles acoustiques et différents dictionnaires de prononciation). Ces décalages entre frontières de phonème manuelles et automatiques restent en général inférieurs à 20 ms [DIC 12].

Configuration du système d'alignement

Dans cette partie, nous allons essayer de quantifier l'impact de la configuration du système du LIMSI [GAU 05b] sur des durées moyennes de segments. La table 4.5 montre ainsi des durées moyennes des voyelles et des consonnes obtenues à partir de 270 heures de données journalistiques (incluant douze millions de segments) en fonction de trois jeux de modèles acoustiques :
CD : modèles dépendants du contexte : pour un phonème de la langue des modèles (allophones) différents sont estimés en fonction des contextes (phonèmes gauche et droit et suivant que le phonème est en position interne de mot ou en frontière de mot) ;
CI : modèles indépendants du contexte : l'ensemble des occurrences d'un même phonème du corpus d'apprentissage est moyenné dans un seul modèle ;
CI-long : modèles indépendants du contexte estimés en sélectionnant uniquement les segments d'une durée supérieure à 50 ms corpus d'apprentissage. Ici nous faisons l'hypothèse que ces segments longs sont plus représentatifs du phonème en question que des segments très courts.

En plus du dictionnaire de prononciation canonique n'incluant que les variantes majeures des mots les plus fréquents, nous avons utilisé un deuxième un dictionnaire de prononciation noté *var* avec plus de variantes, rendant notamment tous les schwas optionnels, en particulier pour les fins de mots se terminant par une consonne. La table 4.5 montre des résultats similaires pour les quatre configurations. Les durées moyennes des voyelles et des consonnes sont proches et varient peu en fonction de la configuration du système. On peut noter que les voyelles sont en moyenne légèrement plus longues que les consonnes et que leur écart-type est presque deux fois plus important que celui des consonnes. Ce résultat est attendu, dans la mesure où ce sont surtout les voyelles qui permettent de changer le débit de parole. La durée moyenne

	Configuration du système d'alignement							
	CD		*CI*		*CI-var*		*CI-long-var*	
	dur. moy.	(éc-type)	dur. moy.	(éc-type)	dur. moy.	(éc-type)	dur. moy.	(éc-type)
Voyelles	80,8	(2,7)	80,8	(2,9)	82,5	(2,8)	82,5	(2,8)
Consonnes	75,8	(1,5)	76,1	(1,4)	75,7	(1,5)	75,6	(1,5)
Pauses	153,2	(13,0)	180,2	(12,7)	175,6	(13,9)	173,9	(13,4)

Tableau 4.5 – Durées moyennes et écarts-types (en ms) calculés sur 270 heures de données journalistiques (12M segments) en fonction de différentes segmentations : *CD* : modèles contexte-dépendants ; *CI* : modèles contexte-indépendants ; *var* : variantes à schwa optionnel ; *long* : modèles acoustiques estimés à partir de segments supérieurs à 50 ms.

des consonnes est particulièrement stable à travers les configurations (l'écart maximum entre moyennes est de 0,5 ms), celle des voyelles augmente très légèrement (de 1,7 ms) si on propose plus de variantes de prononciation (configurations *CI-var CI-long-var*) pour lesquelles la modélisation acoustique des mots permet un schwa optionnel en fin de mot. Des schwas épenthétiques, alignés avec la consonne voisine comme un seul segment consonantique (cf. figure 4.8) dans la configuration initiale, sont maintenant segmentés de manière plus satisfaisante dans un segment de voyelle autonome.

Globalement, les résultats montrent que la configuration précise du système n'a que très peu d'influence sur des résultats comme la durée moyenne de segments et leur écart-type. Les *Pauses*, c'est-à-dire l'ensemble des segments incluant des segments de pauses, silences ou bruits peuvent être librement insérés entre les mots et leur nombre d'occurrences n'est pas imposé par la transcription. Souvent un grand nombre de silences courts alignés peut témoigner d'un décalage entre le signal à décoder et les modèles, l'alignement utilisant ce modèle de pause comme un joker.

Styles de parole

Nous comparons ensuite les durées des segments (voyelles et consonnes) pour différents styles de parole. Existe-t-il des différences importantes ? Deux types de corpus ont été alignés et comparés : parole journalistique et conversations téléphoniques. Nous avons également effectué ces comparaisons sur deux langues, français et anglais, afin de vérifier si le même genre de résultats peuvent être mesurés sur ces deux langues. De plus nous avons examiné si les résultats diffèrent suivant qu'il s'agisse de locuteurs hommes ou femmes. La table 4.6 montre les quantités de données utilisées pour cette comparaison.

La figure 4.6 montre les distributions des durées segmentales en français (à gauche) et en anglais (à droite), avec les données journalistiques (en haut) et les données

	Volume (h)	
Français	Hommes	Femmes
Parole préparée journalistique	270	90
Conversations téléphoniques	25	70
Anglais	Hommes	Femmes
Parole journalistique	420	300
Conversations téléphoniques	1000	1300

Tableau 4.6 – Volume horaire de parole journalistique et conversationnelle examiné en français et en anglais.

conversationnelles (en bas). On peut d'abord remarquer que sur l'ensemble des courbes les distributions ne sont pas gaussiennes. Toutes les courbes montrent à gauche une forte concentration de segments de durées relativement courtes (de moins de 6 cs) et, à droite, une évolution en demi-cloche plus ou moins étalée. La distribution pour l'anglais est un peu plus plate, ce qui pourrait être mis en lien avec le rythme accentuel de cette langue, avec des durées syllabiques plus variables que celles d'une langue syllabique comme le français. On peut aussi remarquer que les courbes sont très similaires pour nos deux populations hommes et femmes dans chaque configuration. Pour revenir à la question des différences entre les durées segmentales en fonction du style de parole, les courbes permettent de répondre affirmativement. En comparant les courbes du haut (journalistique) à celles du bas (conversations), on observe pour les deux langues le même effet : un étalement des durées (la proportion de segments au maximum des courbes de la parole journalistique diminue pour les courbes de parole conversationnelle) et une proportion accrue de segments très courts (3 cs, qui est la durée minimale d'un segment dans le système). Pour le français, cette proportion de segments très courts passe d'environ 8% pour la parole préparée à plus de 18% pour le spontané. La tendance est similaire, quoique plus atténuée en anglais. Ce pic de segments à durée minimale en parole conversationnelle, au-delà d'une réalisation de segments courts, semble pointer des différences de prononciation avec éventuellement moins de phonèmes produits par les locuteurs. En effet, le système d'alignement recherche en général un nombre de phonèmes correspondant à une prononciation canonique de la transcription orthographique. L'absence de certains phonèmes modélisés dans la réalisation produite conduit lors de l'alignement à des segments courts (correspondant aux phonèmes absents), et à des frontières mal placées. Des recherches sont en cours pour clarifier cette hypothèse sur les variantes de prononciation spécifiques à la parole conversationnelle et en particulier les métaplasmes.

Des distributions de durées ont également été calculées sur les données du corpus PFC [DUR 03]. La distribution pour les entretiens et conversations se superpose parfaitement à celle des données conversationnelles, ce qui montre qu'il existe des

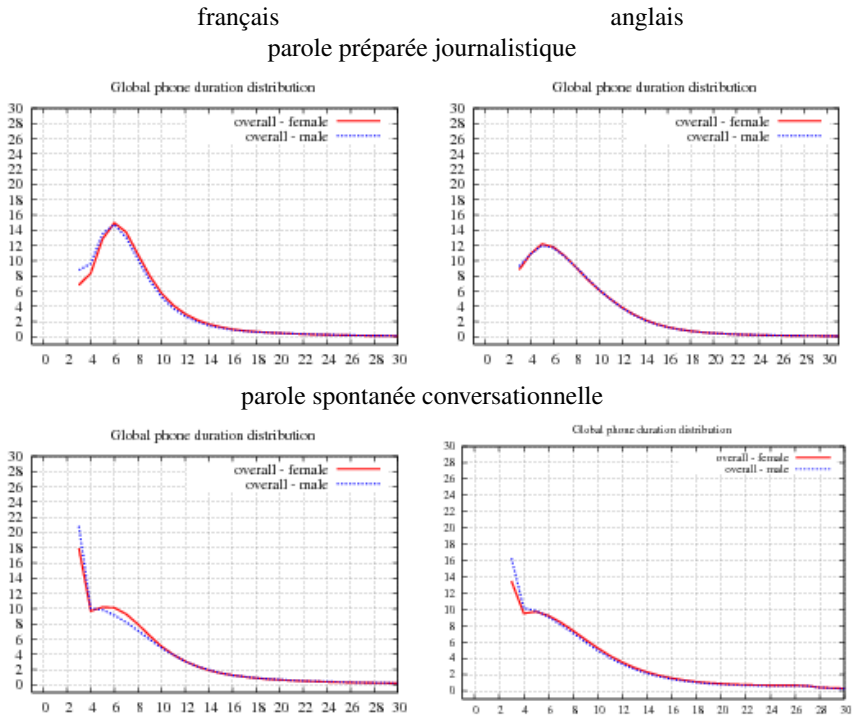

Figure 4.6 – Distributions de durées segmentales en français (gauche) et en anglais (droite) pour un style de parole journalistique (haut) et des conversations téléphoniques (bas). L'abscisse donne les durées en cs (de 3 à 30) et l'ordonnée le pourcentage de segments dans le corpus.

grandeurs caractérisant un certain type de parler qu'on peut mesurer de la même manière sur des corpus peu contrôlés (conversations téléphoniques) et contrôlés (PFC). La distribution correspondant aux textes lus est relativement proche de la distribution des données journalistiques, quoique un peu plus étalée : la lecture du texte du `Maire de Beaulieu` par des amateurs ne conduit pas à un débit aussi rapide et régulier, que celui obtenu par des journalistes professionnels, majoritaires dans le corpus de parole journalistique.

4.3.2. *Distribution de phonèmes du français*

Nous pouvons également étudier, grâce à l'alignement automatique, les fréquences de phonèmes sur les différents corpus mentionnés dans la section précédente.

La figure 4.7 donne les pourcentages d'occurrence des phonèmes dans les corpus de parole journalistique, de conversations téléphoniques et PFC (l'ensemble des lectures, entretiens et conversations) en français. On peut voir que les trois types de corpus suivent globalement la même évolution. Nous utilisons la courbe du corpus journalistique (en rouge, points carrés) comme référence, pour laquelle nous avons indiqué les pourcentages des phonèmes en ordonnée. Concernant la figure pour les voyelles (figure 4.7 haut), le schwa /ə/ et la voyelle centrale ouverte /œ/ sont comptés ensemble et notés par le symbole `x` ; la voyelle centrale fermée /ø/ est codée `eu`. Les deux /O/ ouvert et fermé sont comptés ensemble (o,c) et occupent ainsi 3,6% du corpus, chacun représentant environ 1,8%. Pour PFC nous donnons une courbe globale intégrant les divers types de parole de 12 points d'enquête. On peut observer que pour les conversations téléphoniques par rapport à la parole journalistique, on a pour les voyelles antérieures fermées, 1% (absolu) de /e/ en moins et pour les voyelles ouvertes, environ 2% de /ɛ/ et 2% de /a/ en plus, ainsi qu'un peu moins de 1% de /o,ɔ/ en moins. Pour les voyelles nasales on observe un petit déficit pour la voyelle /ã/ dans les conversations téléphoniques. La courbe PFC reste proche des deux autres courbes.

La figure 4.7 (bas) montre le pourcentage des consonnes du français observées dans les trois corpus. Globalement les courbes d'occurrence des consonnes suivent la même évolution sur les différents genres de corpus. Les consonnes les plus fréquentes du français sont /R/, /l/, /s/, /t/. On voit que le classement varie légèrement en fonction du genre du corpus examiné. Pour la parole spontanée on peut voir que la parole téléphonique génère des proportions de /m/ et de /w/ plus élevées. Ceci est largement dû, pour le /m/, à une proportion élevée de mots comme `mais, moi, me` et des interventions de « backchannel » `hum` en nombre particulièrement élevé ici. Le /w/ provient de nombreuses occurrences de `oui, ouais, moi, voilà, toi, vois, crois`. On peut remarquer que ces mêmes mots enrichissent les statistiques du /a/ et du /ɛ/, ce qui est cohérent avec ce que l'on peut observer sur la figure 4.7 (haut).

4.3.3. *Réalisations du schwa*

Dans les différents corpus, le schwa a une fréquence élevée : environ 5% des segments observés sont des schwas réalisés, et il se place régulièrement parmi les 5 voyelles les plus fréquentes du français (/a/, /e/, /ɛ/, /i/, /ə/). Le rang précis dépend de différents facteurs, dont le style et l'origine régionale du corpus analysé. Concernant les voyelles centrales du français, le schwa est la voyelle la plus observée dans les corpus, loin devant le /œ/ et le /ø/. Le corpus utilisé pour les analyses présentées ici comprend 100 heures de parole d'émissions journalistiques.

Pour le schwa en français nous retenons essentiellement deux propriétés pertinentes lors de sa modélisation dans un système de reconnaissance automatique de la parole :

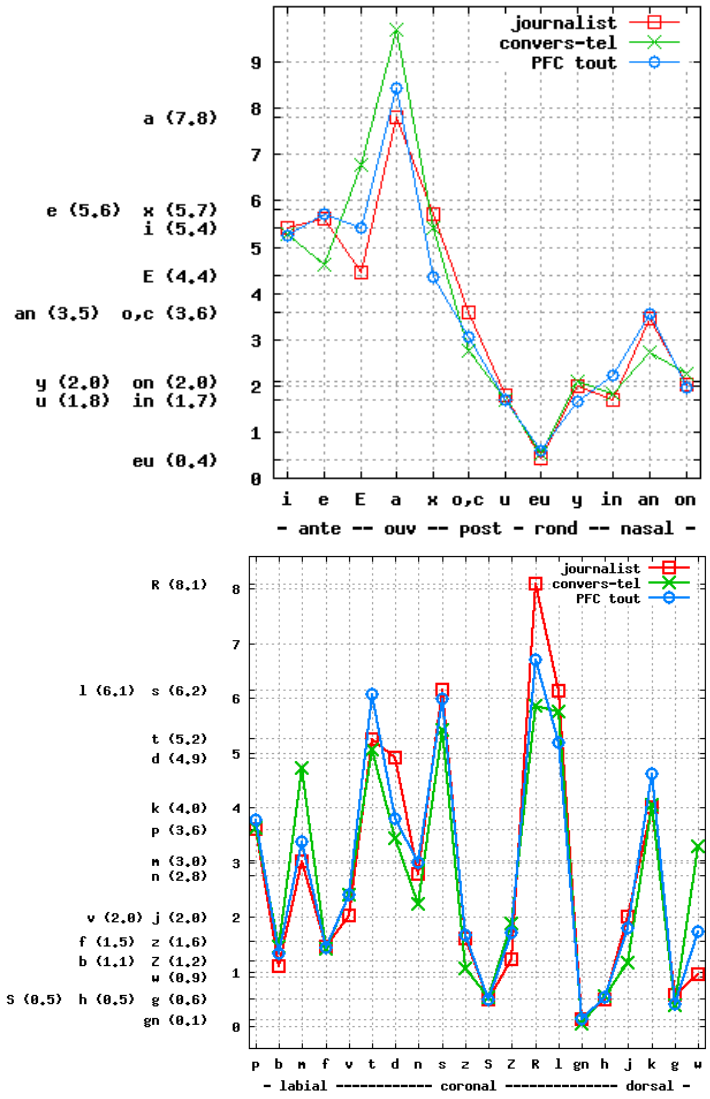

Figure 4.7 – Pourcentage d'occurrence des voyelles (en haut) et des consonnes (en bas) sur 3 types de corpus **journalistique, conversations téléphoniques, PFC**. À gauche en ordonnée, sont reportées les phonèmes avec leur pourcentage dans le corpus journalistique, établissant ainsi une échelle de comparaison.

(1) Le schwa est une **voyelle optionnelle** : sa réalisation est peu certaine et dépend fortement d'un contexte (phonotactique, prosodie, lexique, style de parole, accent. . .) plus large. Il y a potentiellement absence de voyelle (schwa désigne traditionnellement l'absence de voyelle entre deux consonnes en hébreu).

(2) Le schwa est la seule voyelle du système français qui apparaît presque exclusivement comme **noyau de syllabe inaccentuée**.

Ceci amène à supposer pour le schwa des propriétés particulières, notamment concernant son timbre (plutôt central), sa durée (probablement plus courte), sa fréquence fondamentale (éventuellement plus basse) que celle des autres voyelles. Nous avons vu précédemment que le schwa peut poser problème pour la reconnaissance automatique de la parole si un segment est prévu par le modèle mais non produit par le locuteur ou inversement. Dès lors il est intéressant de s'intéresser plus précisément aux contextes facilitant la chute ou le maintien du schwa. Pour cela, nous avons explicité de manière exhaustive les variantes concernant le schwa dans le dictionnaire de prononciation. L'alignement automatique ne permettra peut-être pas de les localiser et de les identifier toutes, cependant il permet de faire émerger des tendances, même avec des modèles acoustiques "bruités".

Afin de mieux comprendre les erreurs observées et d'améliorer la modélisation du schwa dans les systèmes, il est important d'examiner les données, à la fois « manuellement » en quantités limitées, et en adaptant les systèmes de transcription« automatique » pour pouvoir tirer bénéfice des très grands corpus. Différents types d'informations peuvent ainsi être extraits : à partir de descriptions prévoyant tous les sites potentiels de schwa on peut mesurer si la voyelle schwa est réalisée ou non et en déduire des régularités en fonction de paramètres contextuels. Ici nous rencontrons deux types de problèmes : 1) comment définir les sites potentiels de schwa en français ? 2) comment décider si un schwa est réalisé ou non ?

L'alignement automatique permet de décider de la présence ou de l'absence d'un schwa en fonction de l'adéquation entre le signal et les modèles. Cette décision objective est d'autant plus fiable que le modèle représente bien la voyelle centrale en question, et que la durée du segment aligné dépasse la durée minimale de 3 cs. Des mesures de formants des voyelles [GEN 05] ont montrées que le schwa ne présente pas d'écart-type plus important que les autres voyelles orales du français.

Concernant la question des **sites potentiels de schwa**, le plus simple est de les lier à la graphie (c'est-à-dire si l'entrée lexicale contient un e-muet graphémique). Cette approche néglige cependant un certain nombre d'autres voyelles épenthétiques réalisées notamment en frontière de mots (voir figure 4.8). Une telle voyelle est d'autant plus facilement insérée que l'enchaînement des deux mots réalise une séquence de consonnes difficile à articuler dans le système de la langue (cf. règle des trois

consonnes). Une voyelle épenthétique peut également apparaître en fin d'un mot prosodiquement appuyé (`donc` réalisé comme [dõkə] ; `le blocus des ports de la Méditerrannée` réalisé comme ... [blɔkysɔde] ...). Nous regroupons ici sous le terme de schwa l'ensemble de ces voyelles épenthétiques. Une autre méthode pour décider des sites potentiels de schwa, motivée par la phonologie, serait de prévoir un schwa potentiel entre toute séquence de deux consonnes ; de même aux frontières de mots sans contrainte sur la syllabe ouverte ou fermée. Ces voies restent à explorer.

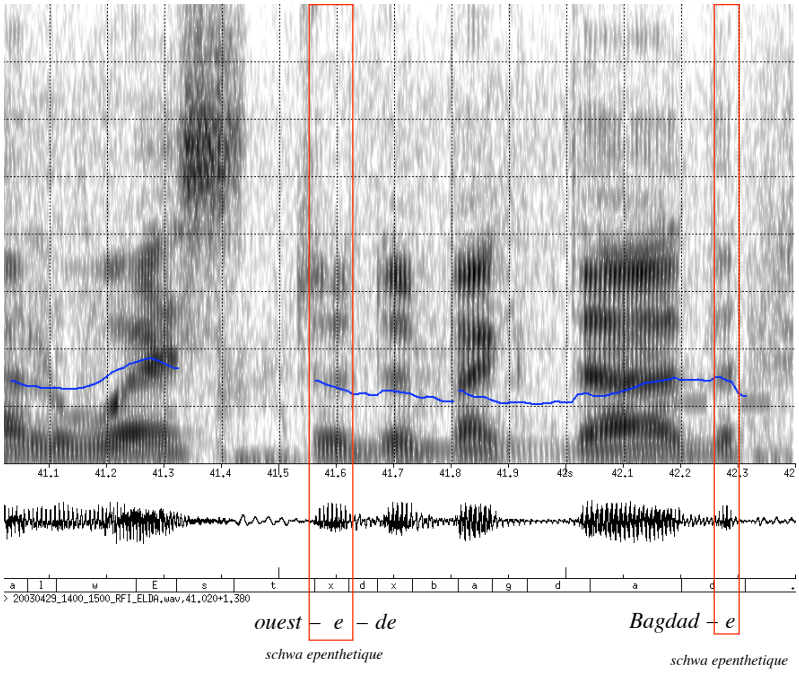

Figure 4.8 – Spectrogramme d'un extrait de radio, illustrant deux voyelles épenthétiques, la première (finale de `ouest` était prévue par le modèle, la deuxième (finale de `Bagdad`) est absorbée par le modèle du /d/. L'alignement montré ici a été fait avec le dictionnaire de prononciation "sans variantes".

Le tableau 4.7 illustre pour quelques mots-types les sites potentiels de schwa retenus pour l'alignement, ainsi que la prononciation maximale et les variantes dérivées. Le tableau 4.8 donne des résultats de réalisation et d'effacement du schwa pour le mot monosyllabique `le` dans différentes conditions : globalement le schwa est réalisé dans 82.6% des cas. Ce taux est intermédiaire entre des mesures effectuées précédemment [ADD 99a] sur de la parole lue (97%) et de la parole spontanée (65%). Ces

mot	prononciation maximale	variantes
site potentiel de schwa : e-muet graphémique intra-mot		
l<u>e</u>	lə	l
c<u>e</u>la	səla	sla
fera	fəʁa	fʁa
d<u>e</u>v<u>e</u>nu	dəvəny	dəvny, dvəny, dvny
site potentiel de schwa : + fin de mot en syllabe fermée		
r<u>e</u>vanch<u>e</u>	ʁəvãʃə	ʁəvãʃ ʁvãʃə ʁvãʃ...
d<u>e</u>v<u>e</u>nir#	dəvəniʁə	dəvənir dəvnirə...
cauch<u>e</u>mar#	kɔʃəmaʁə	kɔʃmaʁ kɔʃmaʁə...

Tableau 4.7 – Extrait du dictionnaire de prononciation utilisé lors de l'alignement illustrant les sites de schwas potentiels considérés, ainsi que les variantes générées.

chiffres varient ensuite de manière assez importante suivant les contextes gauche et droit. La réalisation du schwa est favorisée si le mot est précédé d'une respiration ou d'une consonne, en revanche une voyelle peut favoriser l'effacement. Le taux d'effacement le plus important a pu être mesuré pour un contexte gauche /u/ et un contexte droit /m/, provenant en large partie de la locution `tout le monde`. Ainsi, les résultats montrent qu'à partir de taux de réalisation moyens, de grandes différences peuvent apparaître dès lors qu'on analyse finement les données. Phonotactique, prosodie et fréquence d'occurrence y jouent un rôle important.

contexte gauche	%ə	%∅	#occ.
non-spécifié	82.6	17.4	23950
respiration	92.7	7.3	2730
consonne	89.6	10.4	6480
voyelle	76.7	23.3	12310
contexte gauche __ droit	%ə	%∅	#occ.
respiration __ /t/	98.8	1.2	170
/r/ __ /t/	94.7	5.3	360
/r/ __ /m/	83.6	16.4	540
/u/ __ /m/	14.3	85.7	230

Tableau 4.8 – Taux de réalisation (%ə) et d'effacement (%∅) du schwa pour le mot `le` et différentes conditions de contexte gauche (en haut) et des contextes gauche et droit (bas). #occ. donne le nombre d'occurrences par condition.

4.4. Conclusion

Les progrès accomplis en traitement automatique permettent d'aborder bon nombre de recherches sous un angle nouveau. La disponibilité de grands corpus et d'instruments pour accéder à leur contenu permet aux chercheurs de formuler simultanément de nombreuses questions et de connaître rapidement, si ce n'est des réponses définitives, au moins des tendances permettant d'affiner les questions. Nous vivons actuellement une révolution technologique qui permet d'enrichir le domaine de la linguistique de l'oral de nouveaux instruments et de méthodologies expérimentales exploitant de grands corpus.

De grands corpus sont indispensables pour la mise au point de systèmes de transcription automatique performants. En retour, ces systèmes peuvent produire des annotations de corpus au niveau lexical, phonémique ou syllabique avec des balises temporelles correspondantes. Si les segmentations automatiques diffèrent légèrement des segmentations manuelles, nos travaux montrent que de nombreuses informations peuvent être obtenues en exploitant ce type de données. En particulier, des mesures contrastives sur différents sous-ensembles permettent de minimiser le biais lié à l'instrument utilisé.

Nous avons cherché en premier lieu à déterminer l'influence de la configuration du système d'alignement sur des mesures globales telles que les durées moyennes des voyelles et des consonnes. Nous avons modifié le système d'alignement en utilisant différents jeux de modèles acoustiques dépendants et indépendants du contexte et des dictionnaires de prononciation avec plus ou moins de variantes. Les durées moyennes restent stables pour les différentes configurations de systèmes testées, avec en moyenne 75 ms pour les consonnes et 80 ms pour les voyelles. Comme attendu, l'écart-type des voyelles est supérieur à celui des consonnes.

Nous avons ensuite montré que les distributions de durées segmentales peuvent contribuer à caractériser les langues, les styles de parole et à mettre en évidence des différences de prononciation entre différents styles.

Les distributions de phonèmes du français calculées sur des corpus de français parlé très différents, comme les données journalistiques d' ESTER, des conversations téléphoniques et le corpus PFC restent très similaires avec de légères variations qui peuvent s'expliquer en général par l'apparition de quelques mots très fréquents caractéristiques de la situation de communication ou du sujet traité. Ainsi, en français parlé, le /w/ est globalement peu fréquent, mais apparaît significativement plus dans des conversations familières que dans les journaux radio- et télédiffusés. Un examen comparatif des corpus montre que ceci est dû à de nombreuses occurrences de mots comme `oui`, `ouais`, `moi`, `voilà`, `toi`, `vois`, `crois` en parole familière.

Les alignements permettant de rendre la réalisation du schwa optionnelle via des variantes de prononciation spécifiques, nous ont permis d'examiner le maintien ou

Bibliographie

[ABR 85a] ABRY C., AUTESSERRE D., BARRERA C., BENOÎT C., BOË L.-J., CAELEN J., CAELEN-HAUMONT G., ROSSI M., SOCK R., VIGOUROUX N., « Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 156–163, juin 1985.

[ABR 85b] ABRY C., BENOÎT C., BOË L.-J., SOCK R., « Un choix d'événéments pour l'organisation temporelle du signal de parole », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 133-137, juin 1985.

[ADA 03] ADANK P., Vowel normalization : a perceptual-acoustic study of Dutch vowels, PhD thesis, Radboud University Nijmegen, 2003.

[ADA 04] ADANK P., SMITS R., VAN HOUT R., « A comparison of vowel normalization procedures for language variation research », *Journal of Acoustical Society of America*, vol. 116, n°5, p. 3099–3107, 2004.

[ADD 99a] ADDA-DECKER M., BOULA DE MAREÜIL P., LAMEL L., « Pronunciation variants in French : schwa and liaison », *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Fransisco, p. 2239-2242, août 1999.

[ADD 99b] ADDA-DECKER M., LAMEL L., « Pronunciation variants across system configuration, language and speaking style », *Speech Communication*, vol. 29, p. 83-98, 1999.

[ADD 05] 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.

[ADD 06] ADDA-DECKER M., « De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux », *Actes des XXVI^{es} Journées d'Étude sur la Parole*, Dinard, p. 389–400, juin 2006.

[ADD 07] ADDA-DECKER M., « Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole », *Actes des 5^{e} Journées d'Études Linguistiques*, Nantes, p. 211-216, juin 2007.

[ADD 08] ADDA-DECKER M., GENDROT C., N. N., « Contributions du traitement automatique de la parole à l'étude des voyelles orales du français », *TAL*, vol. 49, n°3, p. 13-46, 2008.

[ADD 11] ADDA-DECKER M., SNOEREN N., « Quantifying temporal speech reduction in French using forced speech alignment », *Journal of Phonetics*, vol. 39, p. 261–270, 2011.

[ADD 12] ADDA-DECKER M., GENDROT C., NGUYEN N., « Apport du traitement automatique à l'étude des voyelles », NGUYEN N., ADDA-DECKER M., Eds., *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*, Hermès, Paris, 2012.

[AME 04] AMELOT A., Étude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français, PhD thesis, Université de Paris III, Paris, 2004.

[AME 06] AMELOT A., ROSSATO S., « Velar movements for the feature [±nasal] for two French speakers », *Proceedings of the 7th International Seminar on Speech Production*, Ubatuba, Brésil, p. 459–467, 13-15 décembre 2006.

[ARM 98] ARMSTRONG N., BOUGHTON Z., « Identification and evaluation responses to a French accent : some results and issues of methodology », *Revue Parole*, vol. 5–6, p. 27–60, 1998.

[ARM 99] ARMSTRONG N., UNSWORTH S., « Sociolinguistic variation in southern French schwa », *Linguistics*, vol. 37, p. 127–156, 1999.

[ARM 02] ARMSTRONG N., JAMIN M., « Le français des banlieues : Uniformity and discontinuity in the French of the Hexagon », SALHI K., Ed., *French In and Out of France*, p. 107–136, Peter Lang, Hamburg, 2002.

[ARM 08] ARMSTRONG N., LOW J., « C'est encœur plus jeuli, le Mareuc : some evidence for the spread of /ɔ/-fronting in French », *Transactions of the Philological Society*, vol. 106, p. 432–455, 2008.

[ASS 82] ASSMANN P. F., HOGAN J. T., NEAREY T. M., « Vowel identification : Orthographic, perceptual and acoustic factors », *Journal of Acoustical Society of America*, vol. 71, p. 975–989, 1982.

[AUD 10] AUDIBERT N., FOUGERON C., FREDOUILLE C., MEUNIER C., « Évaluation d'un alignement automatique sur la parole dysarthrique », *Actes des XXVIII^{es} Journées d'Études sur la Parole*, Mons, Belgique, p. 353–356, 2010.

[AUR 03] 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.

[AUT 85] AUTESSERRE D., ROSSI M., « Propositions pour la segmentation et l'étiquetage de la base de données acoustiques du G.R.E.C.O. Parole », *Actes des XIVèmes Journées d'Études sur la Parole*, Paris, p. 147–151, juin 1985.

[BAD 91] BADIN P., « Fricative consonants : acoustic and X-ray measurements », *Journal of Phonetics*, vol. 19, p. 397–408, 1991.

[BAH 89] BAHL L., BROWN P., DE SOUZA P., MERCER R., « A tree-based statistical language model for natural language speech recognition », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, n°7, p. 1001-1008, 1989.

[BAK 75] BAKER J. M., « The DRAGON system - an overview », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, n°1, p. 24-29, 1975.

[BAR 01] 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.

[BAS 01] BASSET P., AMELOT A., VAISSIÈRE J., ROUBEAU B., « Nasal flow in French spontaneous speech », *Journal of the International Phonetic Association*, vol. 31, p. 87–100, 2001.

[BAU 06] BAUDE O., *Corpus Oraux - Guide des Bonnes Pratiques*, CNRS édition, 2006.

[BEC 99] BECKMAN M., COHEN K., « Modeling the articulatory dynamics of two levels of stress contrast », HORNE M., Ed., *Prosody : Theory and Experiment*, p. 169–200, Kluwer Academic, 1999.

[BEL 57] BELLMAN R., *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[BEN 74] BENGUEREL A.-P., COWAN H., « Coarticulation of upper lip protrusion in French », *Phonetica*, vol. 30, p. 41–55, 1974.

[BER 08] BERTRAND R., BLACHE P., ESPESSER R., FERRÉ G., MEUNIER C., PRIEGO-VALVERDE B., RAUZY S., « Le CID – Corpus of Interactional Data – Annotation et exploitation multimodale de parole conversationnelle », *Traitement Automatique des Langues*, vol. 49, p. 105–134, 2008.

[BEZ 99] VAN BEZOOIJEN R., GOOSKENS C., « Identification of language varieties. Contribution of different linguistic levels », *Journal of Language and Social Psychology*, vol. 18, p. 31–48, 1999.

[BIN 03] BINISTI N., GASQUET-CYRUS M., « Les accents de Marseille », *Cahiers du français contemporain*, vol. 8, p. 107–129, 2003.

[BLA 81] BLADON A., LINDBLOM B., « Modeling the judgement of vowel quality differences », *Journal of Acoustical Society of America*, vol. 69, p. 1414–1422, 1981.

[BLA 82] BLADON A., « Arguments against formants in the auditory representation of speech », CARLSON R., GRANSTRÖM B., Eds., *The representation of speech in the peripheral auditory system*, p. 95–102, Elsevier Biomedical Press, Amsterdam, 1982.

[BLA 99] BLANCHE-BENVENISTE C., « Constitution et exploitation d'un grand corpus », *Revue Française de linguistique appliquée*, vol. IV, n°1, p. 65-74, 1999.

[BOE 01] BOERSMA P., « Praat, a system for doing phonetics by computer », *Glot International*, vol. 5, p. 341–345, 2001.

[BOI 00] BOITE R., BOURLARD H., DUTOIT T., HANCQ J., LEICH H., *Traitement de la parole*, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse, 2000.

[BOM 10] BOMBIEN L., MOOSHAMMER C., HOOLE P., KÜHNERT B., « Prosodic and segmental effects on EPG contact patterns of word-initial German clusters », *Journal of Phonetics*, vol. 38, n°3, p. 388–403, 2010.

[BOT 86] BOTHOREL A., SIMON P., WIOLAND F., ZERLING J.-P., *Cinéradiographie des voyelles et consonnes du français*, Travaux de l'Institut de Phonétique de Strasbourg, Strasbourg, 1986.

[BOU 82] BOURCIEZ E., BOURCIEZ J., *Précis historique de phonétique française*, Klincksieck, Paris, 1982.

[Bou 00] BOULA DE MAREÜIL P., FAGYAL Z., « Autour de l'harmonie vocalique en français », *Actes des XXIII^es Journées d'Étude sur la Parole*, Aussois, p. 85–88, juin 2000.

[Bou 02] BOULA DE MAREÜIL P., ADDA-DECKER M., « Studying pronunciation variants in French by using alignment techniques », *Proceedings of Interspeech*, Denver, p. 2273-2276, 16-20 septembre 2002.

[BRI 74] BRIDLE J., BROWN D., An Experimental Automatic Word-Recognition System, Rapport n°1003, Joint Speech Research Unit, Ruislip, England, 1974.

[BRO 92] BROWMAN C., GOLDSTEIN L., « Articulatory phonology : an overview », *Phonetica*, vol. 49, n°3-4, p. 155–180, 1992.

[BUR 98] BURGER S., C. D., « Identifying dialects of German from digit strings », *Proceedings of the First Language Resources and Evaluation Conference*, Grenade, Espagne, p. 1253–1357, 1998.

[BÜR 07a] BÜRKI A., FOUGERON C., GENDROT C., FRAUENFELDER U., « De l'ambiguïté de la chute du schwa en français », *Actes des 5èmes Journées d'Etudes Linguistiques*, Nantes, p. 83–88, 2007.

[BÜR 07b] BÜRKI A., FOUGERON C., GENDROT C., FRAUENFELDER U., « Chute du schwa en français : un processus sans ambiguïté ? », *Actes des 5^e Journées d'Études Linguistiques*, Nantes, p. 83-88, juin 2007.

[BÜR 07c] BÜRKI A., GENDROT C., « Reconnaissance automatique et analyse linguistique : l'exemple du schwa », *Actes des 7^e Rencontres Jeunes Chercheurs en Parole*, Paris, p. 40-43, juillet 2007.

[BÜR 08] BÜRKI A., GENDROT C., GRAVIER G., LINARÈS G., FOUGERON C., « Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa », *Traitement Automatique des Langues*, vol. 49, 2008.

[CAL 89] CALLIOPE, Ed., *La Parole et son traitement automatique*, Masson, Paris, 1989.

[CAP 05] CAPPEAU P., SEIJIDO M., Les corpus oraux en français. (inventaire 2005 v.1.0), Rapport, Rapport de la DGLFLF, 2005.

[CAP 07] CAPPEAU P., GADET F., « Où en sont les corpus sur les français parlés ? », *Revue Française de Linguistique Appliquée*, vol. 12, p. 129–133, 2007.

[CAR 74] CARTON F., *Introduction à la phonétique du français*, Bordas, Paris, 1974.

[CAR 83] CARTON F., ROSSI M., AUTESSERRE D., LÉON P., *Les Accents des Français*, Hachette, Paris, 1983.

[CAR 84] CARRÉ R., DESCOUT R., ESKÉNAZI M., MARIANI J., ROSSI M., « The French language database : defining, planning and recording a large database », *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Diego, California, USA, mars 1984.

[CAS 03] CASTELLOTTI V., DE ROBILLARD D., « Des français devant la variation : quelques hypothèses », *Cahiers de l'Institut de Linguistique de Louvain*, vol. 29, p. 223–240, 2003.

[CAU 02] CAUBET D., « Métissages linguistiques ici (en France) et là-bas (au Maghreb) », *Ville-École-Intégration Enjeux*, vol. 130, p. 117–132, 2002.

[CHI 41] CHIBA T., KAJIYAMA T., *The Vowel : Its Nature and Structure*, Tokyo-Kaiseikan, Tokyo, 1941.

[CHI 79] CHISTOVICH L. A., SHEIKIN R. L., LUBLINSKAYA V. V., « "Centres of gravity" and spectral peaks as the determinants of vowel quality », LINDBLOM B., ÖHMAN S., Eds., *Frontiers of Speech Communication Research*, p. 143–157, Academic Press, New York, 1979.

[CHI 85] CHISTOVICH L. A., « Central auditory processing of peripheral vowel spectra », *Journal of Acoustical Society of America*, vol. 77, n°3, p. 789–805, 1985.

[CLA 05] CLAIRET S., « Les voyelles nasales en français méridional et non méridional : comparaison à partir d'une étude des paramètres aérodynamiques », *Actes du colloque international "La Méditerranée et ses langues"*, Montpellier, p. 239–246, mars 2005.

[CLO 04] CLOPPER C., PISONI D., « Some acoustic cues for the perceptual categorization of American English regional dialects », *Journal of Phonetics*, vol. 32, p. 111–140, 2004.

[COH 93] COHN A., « Nasalisation in English : phonology or phonetics ? », *Phonology*, vol. 10, p. 42–81, 1993.

[COL 05] COLEMAN J., *Introducing Speech and Language Processing*, Cambridge University Press, Cambridge, UK, 2005.

[COR 97] CORREDOR-ARDOY C., GAUVAIN J.-L., ADDA-DECKER M., LAMEL L., « Language Identification with Language-Independent Acoustic Models », *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, septembre 1997.

[COS 00] COSERIU E., *L'Homme et son langage*, Peeters, Louvain et Paris, 2000.

[COV 01] COVENEY A., *The Sounds of Contemporary French : Articulation and Diversity*, Elm Bank Publications, Exeter, UK, 2001.

[DAV 52] DAVIS K., R. B., S. B., « Automatic Recognition of Spoken Digits », *Journal of the Acoustical Society of America*, vol. 24, n°6, 1952.

[DAV 89] DAVIS B., MERMELSTEIN P., « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n°4, p. 357–366, 1989.

[DEL 55a] DELATTRE P., LIBERMAN A., COOPER F., « Acoustic loci and transitional cues for consonants », *Journal of the Acoustical Society of America*, vol. 27, p. 769–773, 1955.

[DEL 55b] DELATTRE P., LIEBERMAN A., COOPER F., « Acoustic loci and transitional cues for consonants », *Journal of Acoustical Society of America*, vol. 27, n°4, p. 769–773, 1955.

[DEL 65] DELATTRE P., *Comparing the phonetic features of English, Spanish, German and French*, Julius Gross Verlag, Heidelberg, 1965.

[DEL 00] DELVAUX V., « Étude aérodynamique de la nasalité en français », *Actes des XXIII^es Journées d'Étude sur la Parole*, Aussois, p. 141–144, 19–23 juin 2000.

[DEL 02a] DELVAUX V., METENS T., SOQUET A., « French nasal vowels : articulatory and acoustic properties », *Proceedings of ICSLP 2002*, Denver, p. 53–56, septembre 2002.

[DEL 02b] DELVAUX V., METENS T., SOQUET A., « Propriétés acoustiques et articulatoires des voyelles nasales du français », *Actes des XXIV^es Journées d'Etudes sur la Parole*, p. 348–352, 2002.

[DEL 03a] DELAIS-ROUSSARIE E., « Quelques outils d'aide à la transcription et à l'annotation de données audio pour constituer des corpus oraux », DELAIS-ROUSSARIE E., DURAND J., Eds., *Corpus et variation en phonologie du français : Méthodes et analyses*, Presses Universitaires du Mirail, Toulouse, 2003.

[DEL 03b] DELAIS-ROUSSARIE E., MEQQORI A., TARRIER J.-M., « Annoter et segmenter des données de parole sous Praat », DELAIS-ROUSSARIE E., DURAND J., Eds., *Corpus et variation en phonologie du français : Méthodes et analyses*, p. 159–185, Presses Universitaires du Mirail, Toulouse, 2003.

[DEL 03c] DELVAUX V., Contrôle et connaissance phonétique : Les voyelles nasales du français, PhD thesis, Université Libre de Bruxelles, Bruxelles, 2003.

[DEL 04] DELIC, « Présentation du Corpus de référence du français parlé », *Recherches sur le Français Parlé*, vol. 18, p. 11–42, 2004.

[DES 02] DESROSIÈRES A., THÉVENOT L., *Les catégories socioprofessionnelles*, Éditions La Découverte, Paris, 2002.

[DIC 12] DICANIO C., NAM H., WHALEN D., BUNNELL H., AMITH J., GARCÍA R., « Assessing agreement level between forced alignment models with data from endangered language documentation corpora », *Proceedings of Interspeech*, Portland, page4 pages, septembre 2012.

[DIS 80] DISNER S. F., « Evaluation of vowel normalization procedures », *The Journal of the Acoustical Society of America*, vol. 67, n°1, p. 253–261, 1980.

[DOL 97] DOLMAZON J., BIMBOT F., ADDA G., EL BÈZE M., CAEROU J., ZEILIGER J., ADDA-DECKER M., « Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale », *Actes des JST97*, Avignon, avril 1997.

[DRE 50] DREYFUS-GRAF J., « Sonograph and sound mechanics », *Journal of the Acoustical Society of America*, vol. 22, n°6, page731 ?739, 1950.

[DRE 72] DREYFUS-GRAF J., « Parole codée (phonocode) : reconnaissance automatique de langages naturels et artificiels », *Revue d'Acoustique*, vol. 21, p. 3-12, 1972.

[DUB 05] DUBOIS S., « Un siècle de français cadien parlé en Louisiane : Persistance linguistique, hétérogénéité géographique et évolution », VALDMAN A., AUGER J., PISTON-HATLEN D., Eds., *Le Français en Amérique du Nord. Etat présent*, p. 287–305, Les Presses Universitaires de Laval, Saint-Nicolas, 2005.

[DUE 01] DUEZ D., « Manifestation phonétique de la réduction et de l'assimilation contextuelle des segments de la parole conversationnelle », *Revue PArole*, vol. 17-18-19, p. 89–111, 2001.

[DUE 03a] DUEZ D., « Acoustic properties of consonant sequences in conversational French speech », *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelone,

Espagne, p. 2965–2968, 2003.

[DUE 03b] DUEZ D., « Modelling aspects of reduction and assimilation in spontaneous French speech », *Proceedings of the IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, University of Tokyo, p. 120-124, avril 2003.

[DUR 87] DURAND J., SLATER C., WISE H., « Observations on schwa in southern French », *Linguistics*, vol. 25, p. 883–1004, 1987.

[DUR 88] DURAND J., « Les phénomènes de nasalité en français du midi : phonologie de dépendance et sous-spécification », *Recherches Linguistiques de Vincennes*, vol. 17, p. 29–54, 1988.

[DUR 90] DURAND J., *Generative and Non-Linear Phonology*, Longman, London, UK, 1990.

[DUR 02a] DURAND J., LAKS B., LYCHE C., « La phonologie du français contemporain : usages, variétés et structure », PUSCH C., RAIBLE W., Eds., *Romanistische Korpuslinguistik – Korpora und gesprochene Sprache*, p. 93–106, Gunter Narr Verlag, Tübingen, 2002.

[DUR 02b] DURAND J., LAKS B., LYCHE C., (PFC) Protocole d'enquête, Bulletin PFC n ° 1, CNRS UMR 5610 et Université Toulouse-Le Mirail, 2002.

[DUR 03] 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.

[DUR 04] DURAND J., LYCHE C., « Structure et variation dans quelques systèmes vocaliques du français : l'enquête Phonologie du Français Contemporain (PFC) », COVENEY A., SANDERS C., Eds., *Variation et francophonie*, p. 217–240, L'Harmattan, Paris, 2004.

[DUR 05a] DURAND J., « La phonétique classique : l'Association phonétique internationale et son alphabet », NGUYEN N., WAUQUIER-GRAVELINES S., DURAND J., Eds., *Phonologie et Phonétique : Forme et substance*, Hermès, Paris, 2005.

[DUR 05b] DURAND J., LAKS B., LYCHE C., « Un corpus numérisé pour la phonologie du français », WILLIAMS G., Ed., *La Linguistique de corpus*, p. 205–217, Presses Universitaires de Rennes, Rennes, 2005.

[DUR 06] DURAND J., TARRIER J.-M., « PFC, corpus et systèmes de transcription », *Cahiers de Grammaire*, vol. 30, p. 139–158, 2006.

[DUR 08] DURAND J., « Essai de panorama phonologique : les accents du Midi », BARONIAN L., MARTINEAU F., Eds., *Mélanges offerts à Yves-Charles Morin*, Presses de l'Université Laval, 2008.

[DUR 09] DURAND J., LAKS B., LYCHE C., « Le projet PFC : une source de données primaires structurées », DURAND J., LAKS B., LYCHE C., Eds., *Phonologie, variation et accents du français*, p. 19–61, Hermès, Paris, 2009.

[EYC 06] EYCHENNE J., Aspects de la phonologie du schwa dans le français contemporain. Optimalité, visibilité prosodique, gradience, PhD thesis, Université de Toulouse-Le Mirail, 2006.

[FAG 03] FAGYAL Z., NGUYEN N., BOULA DE MAREÜIL P., « From dilation to coarticulation : is there vowel harmony in French ? », *Studies in Linguistic Sciences*, vol. 32, p. 1–21, 2003.

[FAN 60] FANT G., *Acoustic Theory of Speech Production. With Calculations based on X-ray Studies of Russian Articulations*, Mouton, The Hague, 1960.

[FAN 68] FANT G., « Analysis and synthesis of speech processes », MALMBERG B., Ed., *Manual of phonetics*, p. 173–177, North-Holland, Amsterdam, 1968.

[FAN 73] FANT G., *Speech, Sounds and Features*, MIT Press, Cambridge, MA, 1973.

[FAN 75] FANT G., « Non-uniform vowel normalization », *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 2-3, p. 1–19, 1975.

[FAR 97] FARNETANI E., « Coarticulation and connected speech processes », HARDCASTLE W., LAVER J., Eds., *The Handbook of Phonetic Sciences*, p. 371–404, Blackwell, 1997.

[FOR 73] FORNEY G. D., « The Viterbi algorithm », *Proc. of the IEEE*, vol. 61, n°3, p. 268-278, 1973.

[FOU 99] FOUGERON C., SMITH C., « French », *Handbook of the International Phonetic Association*, p. 78–81, Cambridge University Press, Cambridge, UK, 1999.

[FOU 01] FOUGERON C., « Articulatory properties of initial segments in several prosodic constituents in French », *Journal of Phonetics*, vol. 29, p. 109–135, 2001.

[FOU 05] FOUGERON C., « La phonologie articulatoire : une introduction », NGUYEN N., WAUQUIER-GRAVELINES S., DURAND J., Eds., *Phonologie et phonétique : Forme et substance*, p. 265–290, Hermès, Paris, 2005.

[FOU 06] FOULKES P., DOCHERTY G., « The social life of phonetics and phonology », *Journal of Phonetics*, vol. 34, p. 409–438, 2006.

[FOU 07] FOUGERON C., GENDROT C., BÜRKI A., « Le schwa : une voyelle comme les autres ? », *Actes des 5^{e} Journées d'Études Linguistiques*, Nantes, p. 191-198, juin 2007.

[FOU 10] FOUGERON C., AUDIBERT N., FREDOUILLE C., MEUNIER C., GENDROT C., PANSERI O., « Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique », *Actes des XXVIII^{es} Journées d'Études sur la Parole*, Mons, Belgique, p. 365–368, 2010.

[FOW 86] FOWLER C., « An event approach to the study of speech perception from a direct-realist perspective », *Journal of Phonetics*, vol. 14, p. 3–28, 1986.

[FRY 62] FRY D. B., ABRAMSON A. S., EIMAS P. D., LIBERMAN A., « The identification and discrimination of synthetic vowels », *Language and Speech*, vol. 5, p. 171–189, 1962.

[GAD 92] GADET F., *Le français populaire*, Presses Universitaires de France, Paris, 1992.

[GAI 07] GAILLARD-CORVAGLIA A., LÉONARD J.-L., DARLU P., « Testing cladistics on dialect networks and phyla (Gallo-Romance vowels, Southern Italo-Romance diasystems and Mayan languages) », *9th Meeting of the ACL SIG in Computational Morphology and Phonology*, Prague, p. 23–30, 2007.

[GAL 95] GALES M., Model-based techniques for noise robust speech recognition, PhD dissertation, Cambridge University, Cambridge, 1995.

[GAL 05] 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.

[GAL 06] GALANTUCCI B., FOWLER C., TURVEY M., « The motor theory of speech perception reviewed », *Psychonomic Bulletin & Review*, vol. 13, p. 361–377, 2006.

[GAL 09] GALLIANO S., GRAVIER G., CHAUBARD L., « The ESTER 2 evaluation campaign for the rich transcription of French broadcasts », *Proceedings of Interspeech*, Brighton, p. 2583-2586, septembre 2009.

[GAR 93] GAROFOLO J., « TIMIT Acoustic-Phonetic Continuous Speech Corpus », *Linguistic Data Consortium*, Philadelphia, 1993.

[GAU 94] GAUVAIN J.-L., LEE C., « Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, n°2, p. 291-298, 1994.

[GAU 05a] 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 2005.

[GAU 05b] GAUVAIN J., ADDA G., LAMEL L., LEFÈVRE F., SCHWENK H., « Transcription de la parole conversationnelle », *TAL*, vol. 45, n°3, 2005.

[GEN 04] 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, novembre 2004.

[GEN 05] 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, septembre 2005.

[GEN 06] GENDROT C., ADDA-DECKER M., « Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique », *Actes des XXVI^{es} Journées d'Études sur la Parole*, Dinard, p. 407-410, juin 2006.

[GEN 07a] 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*, Saarbrücken, Allemagne, p. 1417–1420, 2007.

[GEN 07b] 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, août 2007.

[GEN 07c] 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, décembre 2007.

[GER 68] GERSTMAN L., « Classification of self-normalized vowels », *IEEE Transactions of Audio Electroacoustics*, vol. 16, p. 78–80, 1968.

[GHI 07] GHIO A., PINTO S., « Résonance sonore et cavités supralaryngées », AUZOU P., ROLLAND V., PINTO S., OZSANCAK C., Eds., *Les dysarthries*, p. 101–110, Solal, Marseille, 2007.

[GIL 10] GILLIÉRON J., EDMONT E., *Atlas linguistique de la France*, Champion, Paris, 1902-1910.

[GOE 02] GOEBL H., « Analyse dialectométrique des structures de profondeur de l'ALF », *Revue de Linguistique Romane*, vol. 66, p. 5–63, 2002.

[GRA 02] GRAFF D., « An overview of Broadcast News corpora », *Speech Communication*, vol. 37, p. 15–26, 2002.

[HAB 05] HABERT B., « Portrait de linguiste(s) à l'instrument », *Texto ! Textes et cultures*, vol. X, n°4, 2005.

[HAN 94] HANSEN A., « Étude du E caduc — stabilisation en cours et variations lexicales », *Journal of French Language Studies*, vol. 4, p. 25–54, 1994.

[HAN 01] HANSEN A., « Lexical diffusion as a factor of phonetic change : The case of Modern French nasal vowels », *Language Variation and Change*, vol. 13, p. 209–252, 2001.

[HAR 99] HARRINGTON J., CASSIDY S., *Techniques in Speech Acoustics*, Kluwer Academic Publishers, Foris, Dordrecht, 1999.

[HAR 04] HARRISON P., Variability of formant measurements, Master's thesis, University of York, 2004.

[HAW 03] HAWKINS S., « Roles and representations of systematic fine phonetic detail in speech understanding », *Journal of Phonetics*, vol. 31, p. 373–405, 2003.

[HEE 04] HEERINGA W., Measuring dialect pronunciation differences using Levenshtein distance, PhD thesis, Rijksuniversiteit, Groningen, 2004.

[HEL 85] HELMHOLTZ H., *On the Sensations of Tone (Translated by A.J. Ellis) (2nd edition)*, Longmans, New York, 1885.

[HIN 78] HINDLE D., « Approaches to vowel normalization in the study of natural speech », SANKOFF D., Ed., *Linguistic Variation, Models and Methods*, p. 161–171, Academic Press, New York, 1978.

[HIN 00] HINTZE M.-A., POOLEY T., JUDGE A., Eds., *French Accents : Phonological and Sociolinguistic Perspectives*, AFLS/CiLT, Londres, 2000.

[HOL 95] HOLST T., NOLAN F., « The influence of syntactic structure on [s] to [ʃ] assimilation », CONNELL B., ARVANITI A., Eds., *Papers in Laboratory Phonology IV : Phonology and Phonetic Evidence*, p. 315–333, 1995.

[HUE 10] HUET S., GRAVIER G., SÉBILLOT P., « Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition », *Computer Speech and Language*, vol. 24, p. 663–684, 2010.

[IKE 06] IKENO A., J.H.L. H., « The role of prosody in the perception of US native English Accents », *Proceedings of Interspeech 2006*, Pittsburgh, p. 437–440, 17-21 septembre 2006.

[JEL 76] JELINEK F., « Continuous speech recognition by statistical methods », *Proc. of the IEEE*, vol. 64, n°4, p. 532–536, 1976.

[JEL 91] JELINEK F., « Self-organized language modeling for speech recognition », WAIBEL A., LEE K., Eds., *Readings in Speech Recognition*, p. 450–506, Morgan Kaufmann, 1991.

[JEL 98] JELINEK F., *Statistical Methods for Speech Recognition*, MIT Press, 1998.

[JOH 90] JOHNSON K., « The role of perceived speaker identity in F0 normalization of vowels », *Journal of the Acoustical Society of America*, vol. 88, p. 642–654, 1990.

[JOH 96] JOHNSON K., STRAND E., « Gradient and visual speaker normalization in the perception of fricatives », GIBBON D., Ed., *Natural Language Processing and Speech Technology. Results of the 3rd KOVENS Conference*, p. 14–26, Mouton de Gruyter, Berlin, 1996.

[JOH 02] JOHNSON K., *Acoustic and Auditory Phonetics*, Blackwell, 2002.

[JOH 04] JOHNSON K., « Massive reduction in conversational American English », YONEYAMA K., MAEKAWA K., Eds., *Spontaneous Speech : Data and Analysis*, p. 29–54, The National International Institute for Japanese Language, 2004.

[JOH 05a] JOHNSON K., « Decisions and mechanisms in exemplar-based phonology », *UC Berkeley Phonology Lab Annual Report*, p. 289–311, 2005.

[JOH 05b] JOHNSON K., « Speaker normalization in speech perception », PISONI D., REMEZ R., Eds., *The Handbook of Speech Perception*, Blackwell, Malden, MA, 2005, 363–389.

[JOL 95] JOLY G., *Précis de phonétique historique du français*, Armand Colin, Paris, 1995.

[JOO 48] JOOS M., « Acoustic Phonetics », *Language Monograph*, vol. 23, 1948.

[KEA 03] KEATING P., FOUGERON C., HSU C., « Domain-initial articulatory strengthening in four languages », LOCAL J., OGDEN R., TEMPLE R., Eds., *Papers in Laboratory Phonology VI : Phonetic Interpretation*, p. 143–161, Cambridge University Press, Cambridge, UK, 2003.

[KOH 92] KOHLER K., « Gestural reorganization in connected speech : a functional viewpoint on 'articulatory phonology' », *Phonetica*, vol. 49, p. 205–211, 1992.

[KOH 95] KOHLER K., « The realization of plosives in nasal/lateral environments in spontaneous speech in German », *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 2, Stockholm, p. 210–213, 1995.

[KOH 98] KOHLER K., « The disappearance of words in connected speech », *ZAS Working Papers in Linguistics*, vol. 11, p. 21–34, 1998.

[KUI 05] KUIPER L., « Perception is reality : Parisian and Provençal perceptions of regional varieties of French », *Journal of Sociolinguistics*, vol. 9, p. 28–52, 2005.

[LAB 66] LABOV W., *The Social Stratification of English in New York City*, Cambridge University Press, New York, 1966.

[LAB 72] LABOV W., *Sociolinguistic Patterns*, University of Pennsylvania Press, Philadelphia, 1972.

[LAB 76] LABOV W., *Sociolinguistique*, Éditions de Minuit, Paris, 1976.

[LAB 94] LABOV W., *Principles of Linguistic Change. Vol. 1 : Internal Features*, Blackwell, Oxford, GB, 1994.

[LAB 06a] LABOV W., « A sociolinguistic perspective on sociophonetic research », *Journal of Phonetics*, vol. 34, p. 500–515, 2006.

[LAB 06b] LABOV W., ASH S., BOBERG C., *The Atlas of North American English : Phonology, Phonetics, and Sound Change. A Multimedia Reference Tool*, Mouton de Gruyter, Berlin, 2006.

[LAD 57] LADEFOGED P., BROADBENT D., « Information conveyed by vowels », *Journal of the Acoustical Society of America*, vol. 29, p. 98–104, 1957.

[LAD 82] LADEFOGED P., BLADON A., « Attempts by human speakers to reproduce Fant's nomograms », *Speech Communication*, vol. 1, p. 185–198, 1982.

[LAD 96a] LADEFOGED P., *Elements of Acoustic Phonetics*, The University of Chicago Press, Chicago, IL, USA, 2nd édition, 1996.

[LAD 96b] LADEFOGED P., MADDIESON I., *The Sounds of the World's Languages*, Blackwell, Oxford, UK, 1996.

[LAD 97] LADEFOGED P., « Instrumental techniques for linguistic phonetic fieldwork », HARDCASTLE W., LAVER J., Eds., *The Handbook of Phonetic Sciences*, p. 137–166, Blackwell, Oxford, UK, 1997.

[LAK 08] LAKS B., « Pour une phonologie de corpus », *Journal of French Language Studies*, vol. 18, p. 3–32, 2008.

[LAM 91] LAMEL L., GAUVAIN J., ESKENAZI M., « BREF, a Large Vocabulary Spoken Corpus for French », *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Gênes, Italie, septembre 1991.

[LAM 99] LAMEL L., ROSSET S., GAUVAIN J., BENNACEF S., « The LIMSI ARISE system for train travel information », *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, USA, mars 1999.

[LAN 75] LANDERCY A., RENARD R., *Éléments de phonétique*, Didier, Bruxelles, 1975.

[LAN 95] LANDICK M., « The mid-vowels in figures : Hard facts », *The French Review*, vol. 69, p. 88–102, 1995.

[LÉO 93] LÉON P., *Précis de phonostylistique*, Fernand Nathan, Paris, 1993.

[LÉO 97] LÉON P., LÉON M., *La prononciation du français*, Fernand Nathan, Paris, 1997.

[LIB 54] LIBERMAN A., DELATTRE P., COOPER F., GERSTMAN L., « The role of consonant-vowel transitions in the perception of the stop and nasal consonants », *Psychological Monographs*, vol. 68, p. 1–13, 1954.

[LIB 85] LIBERMAN A., MATTINGLY I., « The motor theory of speech perception revised », *Cognition*, vol. 21, p. 1–36, 1985.

[LIN 63] LINDBLOM B., « Spectrographic study of vowel reduction », *Journal of the Acoustical Society of America*, vol. 35, p. 1773-1781, 1963.

[LIN 86] LINDBLOM B., « Phonetic universals in vowel systems », OHALA J., JAEGER J., Eds., *Experimental Phonology*, p. 13–44, Academic Press, Orlando, Fl, 1986.

[LIN 90] LINDBLOM B., « Explaining phonetic variation : a sketch of the H&H theory », HARDCASTLE W., MARCHAL A., Eds., *Speech Production and Speech Modelling*, p. 403–439, Kluwer, Dordrecht, 1990.

[LOB 71] LOBANOV B. M., « Classification of Russian vowels spoken by different speakers », *Journal of the Acoustical Society of America*, vol. 49, p. 606–608, 1971.

[LON 88] LONCHAMP F., Études sur la production et la perception de la parole, les indices acoustiques de la nasalité vocalique, la modification du timbre et de la fréquence fondamentale, PhD thesis, Université de Nancy 2, Nancy, 1988.

[LUC 83] LUCCI V., *Étude phonétique du français contemporain à travers la variation situationnelle*, Publications de l'Université des Langues et Lettres de Grenoble, Grenoble, 1983.

[LUD 12] LUDUSAN B., « UNINA System for the EVALITA 2011 Forced Alignment Task », *Working Notes of EVALITA 2011*, Rome, Italie, janvier 2012.

[MAC 98] MACNEILAGE P., « The frame/content theory of evolution of speech production », *Behavioral and Brain Sciences*, vol. 21, p. 499–546, 1998.

[MAC 00] MACWHINNEY B., *The CHILDES project : Tools for analyzing talk. Third Edition*, Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[MAC 09] MACHAC P., SKARNITZL R., *Principles of Phonetic Segmentation*, Epocha Publishing House, Prague, 2009.

[MAD 07] MADDIESON I., « Phonology, naturalness and universals », *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken, Allemagne, p. 77–82, 6-10 août 2007.

[MAE 93] MAEDA S., « Acoustics of vowel nasalization and articulatory shifts in French nasal vowels », HUFFMAN M., KRAKOW R., Eds., *Phonetics and Phonology, vol. 5 : Nasals, Nasalization, and the velum*, p. 147–167, Academic Press, San Diego, 1993.

[MAL 76] MALÉCOT A., LINDSAY P., « The neutralization of /ɛ̃/-/œ̃/ in French », *Phonetica*, vol. 33, p. 45–61, 1976.

[MAL 94] MALDEREZ I., « Vers la perte de l'opposition du lieu d'articulation des voyelles moyennes arrondies dans la parole des jeunes gens d'Île-de-France », *Actes des XX^{es} Journées d'Étude sur la Parole*, Trégastel, p. 361–366, juin 1994.

[MAL 03] MALFRÈRE F., DEROO O., DUTOIT T., RIS C., « Phonetic alignment : speech-synthesis-based versus Viterbi-based », *Speech Communication*, vol. 40, p. 503–517, 2003.

[MAN 86] MANTAKAS M., SCHWARTZ J., ESCUDIER P., « Modèle de prédiction du deuxième formant effectif F'2 – application à l'étude de la labialité des voyelles avant du français », Société Française d'Acoustique, p. 157–161, 1986.

[MAR 45] MARTINET A., *La prononciation du français contemporain*, Droz, Paris, 1945.

[MAR 58] MARTINET A., « C'est jeuli le Mareuc ! », *Romance Philology*, vol. 11, p. 345–355, 1958.

[MAR 77] MARTINET A., WALTER H., *Dictionnaire de la prononciation française dans son usage réel*, France-Expansion-Egena, Paris, 1977.

[MAR 80] MARCHAL A., *Les sons et la parole*, Guérin, Montréal, 1980.

[MAR 88] MARCHAL A., « Coproduction : evidence from EPG data », *Speech Communication*, vol. 7, p. 287–295, 1988.

[MAR 99] MARIANI J., PAROUBEK P., « Human Language Technologies Evaluation in the European Framework », *Proceedings of the DARPA Broadcast News Workshop*, p. 237-242, Morgan Kaufman Publishers, Washington, février 1999.

[MAR 05] MARIANI J., « Developing Language Technologies with the Support of Language Resources and Evaluation Programs », *Language Resources and Evaluation*, vol. 39, n°1, p. 35-44, 2005.

[MAR 08] MARTIN P., *Phonétique acoustique*, Armand Colin, Paris, 2008.

[MAT 81] MATTINGLY I., « Phonetic representation and speech synthesis by rule », MYERS T., LAVER J., ANDERSON J., Eds., *The Cognitive Representation of Speech*, p. 415–420, North Holland, Amsterdam, 1981.

[MAT 09] MATROUF D., « Variabilités acoustiques nuisibles pour le traitement automatique de la parole », 2009, Habilitation à Diriger les Recherches, Université d'Avignon et des Pays de Vaucluse, Académie d'Aix-Marseille.

[MCE 01] MCENERY T., WILSON A., *Corpus Linguistics*, Edinburgh University Press, Edimbourg, 2001.

[MÉN 02] MÉNARD L., SCHWARTZ J., BOË L., KANDEL S., VALLÉE N., « Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood », *Journal of the Acoustical Society of America*, vol. 111, n°4, p. 1892–1905, 2002.

[MER 76] MERMELSTEIN P., « Distance measures for speech recognition, psychological and instrumental », CHEN C., Ed., *Pattern Recognition and Artificial Intelligence*, p. 374–388, Academic Press, New York, 1976.

[MEU 94] MEUNIER C., Les groupes de consonnes. Problématique de la segmentation et variabilité acoustique, PhD thesis, Université de Provence, 1994.

[MEU 05] MEUNIER C., « Invariants et variabilité en phonétique », NGUYEN N., WAUQUIER S., DURAND J., Eds., *Phonologie et phonétique : Forme et substance*, p. 349-374, Hermès, Paris, 2005.

[MEU 11] MEUNIER C., ESPESSER R., « Vowel reduction in conversational speech : the role of lexical factors », *Journal of Phonetics*, vol. 39, p. 271–278, 2011.

[MEU 12a] MEUNIER C., « Contexte et nature des réalisations phonétiques en parole conversationnelle », *Actes des 29èmes Journées d'Études sur la Parole*, p. 1–8, 2012.

[MEU 12b] MEUNIER C., NGUYEN N., « Le traitement et l'analyse du signal de parole », NGUYEN N., ADDA-DECKER M., Eds., *Méthodes et outils pour l'analyse phonétique des grands corpus oraux*, Hermès, Paris, 2012.

[MEY 01] MEYNADIER Y., « La syllabe phonétique et phonologique : une introduction », *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, vol. 20, p. 91–148, 2001.

[MEY 03] MEYNADIER Y., Interaction entre prosodie et (co)articulation en français, PhD thesis, Université de Provence, Aix-en-Provence, 2003.

[MIL 89] MILLER J. D., « Auditory-perceptual interpretation of the vowel », *Journal of the Acoustical Society of America*, vol. 85, p. 2114–2134, 1989.

[MON 04] MONTAGU J., « Les sons sous-jacents aux voyelles nasales en français parisien : indices perceptifs des changements », *Actes des XXV^{es} Journées d'Étude sur la parole*, Fès, Maroc, p. 385–388, avril 2004.

[MOO 83] MOORE B. C., GLASBERG B. R., « Suggested formulae for calculating auditory-filter bandwidths and excitation patterns », *Journal of the Acoustical Society of America*, vol. 74, p. 750–753, 1983.

[MOO 97] MOORE B., *An introduction to the psychology of hearing*, Academic Press, San Diego, CA, 1997.

[MOU 91] MOUGEON R., BENIAK E., *Linguistic Consequences of Language Contact and Restriction : The Case of French in Ontario, Canada*, Oxford University Press, Oxford, 1991.

[NEA 78] NEAREY T. M., Phonetic Feature Systems for Vowels, PhD thesis, Indiana University Linguistics Club, 1978.

[NEA 89] NEAREY T., « Static, dynamic, and relational properties in vowel perception », *Journal of the Acoustical Society of America*, vol. 85, p. 2088–2113, 1989.

[NEY 84] NEY H., « The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32-2, p. 263–271, 1984.

[NGU 01] NGUYEN N., « Rôle de la coarticulation dans la reconnaissance des mots », *L'Année Psychologique*, vol. 101, p. 125–154, 2001.

[NGU 04a] NGUYEN N., FAGYAL Z., COLE J., « Perceptual relevance of long-domain phonetic dependencies », *Proceedings of the IVth Linguistic Studies Workshop*, Nantes, p. 173–178, mai 2004, (Accessible en ligne à http ://www.lling.fr/actes/actes-jel2004.pdf).

[NGU 04b] NGUYEN N., WAUQUIER S., TULLER B., « Méthodes et outils pour l'analyse des systèmes vocaliques », *Bulletin Phonologie du français contemporain*, vol. 3, p. 77-85, 2004.

[NGU 08] NGUYEN N., FAGYAL Z., « Acoustic aspects of vowel harmony in French », *Journal of Phonetics*, vol. 36, p. 1–27, 2008.

[NGU 09] NGUYEN N., WAUQUIER S., TULLER B., « The dynamical approach to speech perception : From fine phonetic detail to abstract phonological categories. », PELLEGRINO F., MARSICO E., CHITORAN I., COUPÉ C., Eds., *Approaches to Phonological Complexity*, p. 191-218, Mouton de Gruyter, Berlin, 2009.

[NIE 08] NIEBHUR O., LANCIA L., MEUNIER C., « On place assimilation in French sibilant sequences », *Proceedings of the 8th International Seminar on Speech Production*, Strasbourg, p. 221–224, 2008.

[NOR 75] NORDSTRÖM P. E., LINDBLOM B., « A normalization procedure for vowel formant data », *Proceedings of the 8th International Congress of Phonetic Sciences*, Leeds, 1975.

[OHA 81] OHALA J., « The listener as a source of sound change », MASEK C., HENDRICK R., MILLER M., Eds., *Papers from the Parasession on Language and Behavior*, Chicago, IL, USA, Chicago Linguistic Society, p. 178–203, 1981.

[PAL 84] PALTRIDGE J., GILES H., « Attitudes towards speakers of regional accents of French : effects of regionality, age and sex of listeners », *Linguistische Berichte*, vol. 90, p. 71–85, 1984.

[PAL 03] PALLETT D. S., « A Look at NIST's Benchmark ASR Tests : Past, Present, and Future », *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Îles Vierges, Etats-Unis, p. 1-6, 2003.

[PET 52] PETERSON G. E., BARNEY H. L., « Control methods used in the study of the vowels », *Journal of the Acoustical Society of America*, vol. 24, p. 175–184, 1952.

[PET 61] PETERSON G. E., « Parameters of vowel quality », *Journal of Speech and Hearing Research*, vol. 4, p. 10–29, 1961.

[PIT 05] PITT M., JOHNSON K., HUME E., KIESLING S., RAYMOND W., « The Buckeye corpus of conversational speech : Labeling conventions and a test of transcriber reliability », *Speech Communication*, vol. 45, p. 89–95, 2005.

[POT 50] POTTER R. K., STEINBERG J., « Towards the specification of speech », *Journal of Acoustical Society of America*, vol. 22, p. 807–820, 1950.

[PRE 89] PRESTON D., *Perceptual Dialectology*, Foris Publications, Dordrecht, 1989.

[PRI 90] PRICE P., « Evaluation of spoken language systems : the ATIS domain », *Proceedings of the DARPA Speech and Natural Language Workshop*, St. Thomas, Îles Vierges, Etats-Unis, June 1990.

[RAB 89] RABINER L. R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Proc. of the IEEE*, vol. 77, n°2, p. 257-286, 1989.

[RAC 02] RACINE I., GROSJEAN F., « La production du E caduc facultatif est-elle prévisible ? Un début de réponse », *Journal of French Language Studies*, vol. 12, p. 307–326, 2002.

[REC 09] RECASENS D., ESPINOZA A., « An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan », *Journal of the Acoustical Society of America*, vol. 125, p. 2288–2298, 2009.

[RID 11] RIDOUANE R., MEYNADIER Y., FOUGERON C., « La syllabe : objet théorique et nature physique », *Faits de Langue*, vol. 37, p. 213–234, 2011.

[ROS 90] ROSSI M., « Segmentation automatique de la parole : pourquoi ? Quels segments ? », *Traitement du Signal*, vol. 7, p. 315–326, 1990.

[ROS 94] ROSNER B., PICKERING J. B., *Vowel perception and production*, Oxford University Press, Oxford, 1994.

[ROS 98] ROSSATO S., FENG G., LABOISSIÈRE R., « Recovering gestures from speech signals : a preliminary study for nasal vowels », *Proceedings of ICSLP 98*, vol. 3, Sydney, Australie, p. 1091–1094, novembre-décembre 1998.

[RUB 92] RUBIN D. L., « Non-language factors affecting undergraduates' judgements of non-native English-speaking teaching assistants », *Research in Higher Education*, vol. 33, page 4, 1992.

[SAL 86] SALTZMAN E., « Task dynamic coordination of the speech articulators : a preliminary model », *Experimental Brain Research Series*, vol. 15, p. 129–144, 1986.

[SCH 68] SCHWARTZ M. F., « Identification of speaker sex from isolated voiceless fricatives », *Journal of the Acoustical Society of America*, vol. 43, p. 1178–1179, 1968.

[SCH 04] SCHMID H., « Probabilistic part-of-speech tagging using decision trees », *International Conference on New Methods in Language Processing*, Manchester, p. 44–49, 2004.

[SCH 12] SCHWARTZ J.-L., BASIRAT A., MÉNARD L., SATO M., « The Perception-for-Action-Control Theory (PACT) : A perceptuo-motor theory of speech perception », *Journal of Neurolinguistics*, vol. 25, p. 336–354, 2012.

[SHA 90] SHADLE C., « Articulatory-acoustic relationships in fricative consonants », HARDCASTLE W., MARCHAL A., Eds., *Speech Production and Speech Modelling*, p. 187–209, Kluwer Academic, Dordrecht, 1990.

[SIE 00] SIEMUND R., HÖGE H., KUNZMANN S., MARASEK K., « SPEECON - speech data for consumer devices », *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athènes, Grèce, p. 883-886, 2000.

[SOB 06] SOBOTTA E., « Continuum ou variétés ? La classification des accents de migrants aveyronnais à Paris », KREFELD T., RADTKE E., Eds., *Modellando lo spazio in perspectiva linguistica*, p. 195–214, Peter Lang, Frankfurt am Main, 2006.

[STE 40] STEVENS S. S., VOLKMANN J., « The relation of pitch to frequency : A revised scale », *American Journal of Psychology*, vol. 53, p. 329–353, 1940.

[STE 78] STEVENS K., BLUMSTEIN S. E., « Invariant cues for place of articulation in stop consonants », *Journal of the Acoustical Society of America*, vol. 64, p. 1358–1368, 1978.

[STE 89] STEVENS K., « On the quantal nature of speech », *Journal of Phonetics*, vol. 17, p. 3–45, 1989.

[STE 98] STEVENS K., *Acoustic Phonetics*, MIT Press, Cambridge, Mass., 1998.

[STR 63] STRAKA G., « La division des sons du langage entre voyelles et consonnes peut-elle être justifiée ? », *Travaux de Linguistique et de Littérature de Strasbourg*, vol. I, p. 17–99, 1963.

[STR 64] STRAKA G., « L'évolution phonétique du latin au français sous l'effet de l'énergie et de la faiblesse articulatoires », *Travaux de Linguistique et de Littérature de Strasbourg*, vol. II, p. 17–98, 1964.

[STR 89] STRANGE W., « Evolving theories of vowel perception », *Journal of the Acoustical Society of America*, vol. 85, p. 2081–2087, 1989.

[SYR 86] SYRDAL A. K., GOPAL H. S., « A perceptual model of vowel recognition based on the auditory representation of American English vowels », *Journal of the Acoustical Society of America*, vol. 79, p. 1086–1100, 1986.

[TAR 03a] TARRIER J.-M., « L'enregistrement et la prise de son », DELAIS-ROUSSARIE E., DURAND J., Eds., *Corpus et variation en phonologie du français : Méthodes et analyses*, p. 187–212, Presses Universitaires du Mirail, Toulouse, 2003.

[TAR 03b] TARRIER J.-M., « L'enregistrement et la prise de son », DURAND J., DELAIS-ROUSSARIE E., Eds., *Corpus et variation en phonologie du français : Méthodes et analyses*, p. 187–212, Presses Universitaires du Mirail, Toulouse, 2003.

[TES 01] TESTON B., « L'enregistrement numérique de la voix et la parole : problèmes et méthodes », *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, vol. 20, p. 219–232, 2001.

[THO 91] THOMAS A., « Évolution de l'accent méridional en français niçois : les nasales », *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix en Provence, France, p. 194–197, août 1991.

[THO 07] THOMAS E. R., KENDALL T., « NORM : The vowel normalization and plotting suite. », [Online Resource : http ://ncslaap.lib.ncsu.edu/tools/norm/], 2007.

[TOD 09] TODA M., Étude articulatoire et acoustique des fricatives sibilantes, PhD thesis, Université de Paris III, Paris, 2009.

[TOR 10] TORREIRA F., ADDA-DECKER M., ERNESTUS M., « The Nijmegen corpus of casual French », *Speech Communication*, vol. 52, p. 201–212, 2010.

[TRA 87] TRANEL B., *The Sounds of French : An Introduction*, Cambridge Univ. Press, Cambridge, UK, 1987.

[TRO 39] TROUBETZKOY N., *Grundzüge der Phonologie / Principes de phonologie*, Vandenhoek & Rupprecht / Klincksieck, Göttingen / Paris, 1939.

[TUB 70] TUBACH J., Reconnaissance automatique de la parole : étude et réalisation fondées sur les niveaux acoustique, morphologique et syntaxique, Thèse, Université de Grenoble, 1970.

[VAI 07] VAISSIÈRE J., « Area Functions and Articulatory Modeling as a Tool for Investigating the Articulatory, Acoustic, and Perceptual Properties of Sounds Across Languages », SOLE M., BEDDOR P., OHALA M., Eds., *Experimental Approaches to Phonology*, p. 55–72, Oxford University Press, 2007.

[VAL 94] VALLÉE N., Systèmes vocaliques : de la typologie aux prédictions, PhD thesis, Université Stendhal, Grenoble, 1994.

[VER 76] VERBRUGGE R., STRANGE W., SHANKWEILER P., EDMAN T. R., « What information enables a listener to map a speaker's vowel space ? », *Journal of the Acoustical Society of America*, vol. 60, p. 198–212, 1976.

[VIN 68] VINTSYUK T., « Speech discrimination by dynamic programming », *Kibernetika*, vol. 4, p. 81–88, 1968.

[WAL 76] WALTER H., *La dynamique des phonèmes dans le lexique français contemporain*, France-Expansion, Paris, 1976.

[WAL 82] WALTER H., *Enquête phonologique et variétés régionales du français*, Presses Universitaires de France, Paris, 1982.

[WAN 11] WANG Y., GALES M., « Speaker and noise factorisation on the AURORA4 task », *Proceedings of ICASSP '11*, Prague, République tchèque, mai 2011.

[WES 96] WESENICK M. B., KIPP A., « Estimating the quality of phonetic transcriptions and segmentations of speech signals », *Proceedings of ICSLP 1996*, p. 129–132, 1996.

[WIL 99] WILLIAMS A., GARRETT P., COUPLAND N., « Dialect recognition », PRESTON D., Ed., *Handbook of Perceptual Dialectology*, p. 345–358, John Benjamins, Amsterdam, 1999.

[WIL 05] WILLIAMS G., Ed., *La Linguistique de corpus*, Presses Universitaires de Rennes, Rennes, 2005.

[WIR 52] WIREN-STUBBS, « Automatic Recognition of Spoken Digits », *Journal of the Acoustical Society of America*, vol. 24, n°6, 1952.

[WOE 07] WOEHRLING C., BOULA DE MAREÜIL P., « Comparing Praat and Snack formant measurements on two large corpora of northern and southern French », *Proceedings of Interspeech 2007*, Anvers, Belgique, p. 1006–1009, 27-31 août 2007.

[WOE 08] WOEHRLING C., BOULA DE MAREÜIL P., ADDA-DECKER M., « Aspects prosodiques du français parlé en Alsace, Belgique et Suisse », *Actes des XXVII^{es} Journées d'Études sur la parole*, Avignon, juin 2008.

[WOE 09] WOEHRLING C., Accents régionaux en français. Perception, analyse et modélisation à partir de grands corpus, PhD thesis, Université Paris-Sud, Orsay, 2009.

[WOO 86] WOOD S., « The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels », *Journal of the Acoustical Society of America*, vol. 80, p. 391–401, 1986.

[WRI 04] WRIGHT R., « A review of perceptual cues and cue robustness », HAYES B., KIRCHNER R., STERIADE D., Eds., *Phonetically-Based Phonology*, p. 34–57, Cambridge University Press, Cambridge, UK, 2004.

[YOU 97] YOUNG S., ADDA-DECKER M., AUBERT X., DUGAST C., GAUVAIN J., KERSHAW D., LAMEL L., LEEUWEN D., PYE D., ROBINSON A., STEENEKEN H., WOODLAND P., « Multilingual large vocabulary speech recognition : the European SQALE project », *Computer Speech and Language*, vol. 11, n°1, 1997.

[ZER 84] ZERLING J., « Phénomènes de nasalité et de nasalisation vocaliques : Etude ciné-radiographique pour deux locuteurs », *Travaux de l'Institut de Phonétique de Strasbourg*, vol. 16, p. 241–266, 1984.

[ZUE 90] ZUE V., SENEFF S., GLASS J., « Speech database development at MIT : TIMIT and beyond », *Speech Communication*, vol. 9, n°4, p. 351–356, 1990.

[ZWI 86] ZWICKER E., « Subdivision of the audible frequency range into critical bands as a function of frequency », *Journal of the Acoustical Society of America*, vol. 33, p. 248–249,