Université Paris–Sud — Faculté des sciences d'Orsay
École Doctorale d'Informatique de Paris-Sud
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

---

LARGE-SCALE ACOUSTIC AND PROSODIC INVESTIGATIONS OF FRENCH

Rena NEMOTO

Thèse pour le diplôme de Docteur en Sciences, spécialité Informatique
soutenue publiquement le mercredi 16 novembre 2011 à Orsay
devant le jury composé de

Rapporteurs	Yannick ESTÈVE
	François PELLEGRINO
Directrice	Martine ADDA–DECKER
Co-encadrante	Ioana VASILESCU
Examinateurs	Anne VILNAT
	Olivier FERRET

Remerciements

Quand je me suis inscrite en thèse, je n'avais en tête qu'il y avait une soutenance pour moi. Mais quelques années plus tard, cela m'est arrivé.

Je tiens tous d'abord à remercier mes directrices de thèse, Martine Adda-Decker et Ioana Vasilescu pour m'avoir offert une occasion de faire une thèse avec elles et aussi pour tous leurs conseils, disponibilité, gentillesse, patience et soutient, et ce dès mon premier stage au LIMSI. Grâce à elles, ma curiosité et mon intérêt pour la parole n'ont cessés de croître.

Je remercie également les membres de mon jury de thèse, Yannick Estève, François Pellegrino, Anne Vilnat et Olivier Ferret, pour leur participation à ma soutenance, pour leur lecture attentive et pour leurs suggestions intéressantes.

Je tiens aussi à remercier Jacques Durand avec qui j'ai pu faire évoluer mes études sur la prosodie.

Je remercie encore Jean-Luc Gauvain, chef du groupe Traitement du Langage Parlé, et Lori Lamel pour m'avoir accueillie au sein de groupe et pour m'avoir offert la possibilité de terminer ma thèse dans de bonnes conditions.

Je tiens à remercier tous les membres du groupe TLP et les membres du LIMSI pour les années agréables. Je pense en particulier à Bianca et Cécile qui m'ont toujours soutenue même après leurs thèses ; Nadi, Souhir, Penny, Hai-Son et Thiago qui sont encore petits doctorants ainsi que futurs docteurs ; Marc, Laurence, Tanel, Gary, Dong avec qui j'ai partagé les premières années de ma thèse ; Viet Anh, Jáchym, Thomas, Ilya, Adrian, Artem, Marco, Hervé, Qing-Qing, Giulia avec qui j'ai partagé les dernières années de ma thèse ; Natalie, Nadège, Viet Bac, Josep, Philippe, Eric, Claude, Gilles, Faouzi et tous ceux que je ne cite pas ici.

Je remercie aussi et surtout ma famille et mes amis pour leur soutien de loin comme de près. Merci à Julien, Yoko, Aurélie, Silvana, Michael, ... avec qui j'ai pu partager ou je partage toujours d'excellents et drôles moments à la Cité U, au bord de la Seine, sur la pelouse, ... enfin partout dans le monde.

Je remercie enfin les futurs lecteurs qui trouveront cette thèse et les futurs collègues avec qui je travaillerai avec enthousiasme.

This thesis was partially financed by RTRA-DIGITEO, Région Île-de-France, under the projet AMADEO (**A***pprentissage à partir de grandes* **ma***sses de* **d***onn***é***es* **o***rales*, 2007-01D), and by the OSEO *Quaero* program.

Contents

List of Figures

List of Tables

List of Acronyms

AFCP *Association Francophone de la Communication Parlée*, French-speaking Speech Communication Association

AM Acoustic Model

ARPA Advanced Research Projects Agency

ASR Automatic Speech Recognition

BN Broadcast News

C Consonant

CI Context Independent

CTS Conversational Telephone Speech

DARPA Defense Advanced Research Projects Agency

DGA *Délégation Générale de l'Armement*, General Delegation for Ordnance

DP Dynamic Programming

ELDA Evaluations and Language resources Distribution Agency

ESTER *Évaluation des Systèmes de Transcription d'Émissions Radiophoniques*, Evaluation of Radio Broadcast Rich Transcription Systems

ESTER1 *Évaluation des Systèmes de Transcription d'Émissions Radiophoniques pour la 1ère campagne*, Evaluation of Radio Broadcast Rich Transcription Systems for the 1st campaign

ESTER2 *Évaluation des Systèmes de Transcription d'Émissions Radiophoniques pour la 2nd campagne*, Evaluation of Radio Broadcast Rich Transcription Systems for the 2nd campaign

f_0 Fundamental Frequency

F1 First Formant

F2 Second Formant

F3 Third Formant

IWER Individual Word Error Rate

HMM Hidden Markov Model

HSR Human Speech Recognition

INTSINT INternational Transcription System for INTonation

IPA International Phonetic Alphabet

LIMSI *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur*, Computer Sciences Laboratory for Mechanics and Engineering Sciences

LM Language Model

LMT Logistic Model Trees

LPC Linear Predictive Coding

LVCSR Large Vocabulary Continuous Speech Recognition

MFCC Mel-Frequency Cepstral Coefficient

NIST National Institute of Standards and Technology

OOV Out-Of-Vocabulary

PFC *Phonologie du Français Contemporain*, Phonology of Contemporary French

POS Part-Of-Speech

RT Response Time, or Reaction Time

RT Rich Transcription

S Semivowel

SVM Support Vector Machines

ToBI Tone and Break Indices

V Vowel

WER Word Error Rate

Résumé

Introduction

Cette thèse est consacrée aux analyses acoustiques et prosodiques du français à partir de grandes masses de données orales illustrant deux styles de parole différents. D'une part, il s'agit de données de parole préparée consistant en des enregistrements de journaux radiodiffusés francophones ayant été utilisés lors de la campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques (ESTER) [Galliano *et al.*, 2005; Galliano *et al.*, 2009]. D'autre part, nous avons utilisé un sous-ensemble d'entretiens face à face issus du corpus PFC (Phonologie du Français Contemporain) [Delais-Roussarie and Durand, 2003; Durand *et al.*, 2005] qui contient des enregistrements de variétés de français des régions.

L'objectif à long terme des études proposées dans cette thèse concerne l'amélioration des systèmes de reconnaissance automatique de la parole (RAP) à travers l'amélioration des modèles de prononciation. Dans une perspective à plus court terme, notre but est d'augmenter nos connaissances de la variabilité dans la prononciation propre à l'oral telle qu'illustrée dans différents styles de parole. Nos travaux se basent sur l'analyse des paramètres acoustiques et prosodiques des mots et des syntagmes du français. Ces paramètres sont utilisés pour discriminer des mots homophones et des séquences de mots. En effet, une raison majeure d'erreurs observées dans la transcription automatique de la parole concerne les mots et/ou multi-mots homophones ayant ainsi une prononciation identique ou très peu distinctive pour lesquels la solution automatique repose entièrement sur les modèles de langage.

Le français inclut une grande proportion de mots et multi-mots homophones. Un phonème peut ainsi correspondre à un mot écrit et nombre de mots pouvant être transcrits phonétiquement par un même phonème ont des graphies très différentes (par exemple, /a/ : *a*, *as*, *à* ; /o/ : *au*, *aux*, *eau*, *eaux*, *haut*, *hauts*, *oh* ; /s/ : *s'*, *c'* ; ...). Cette observation concerne aussi les séquences de phonèmes : un mot incluant deux phonèmes (/ma/ *ma*) peut ainsi être décomposé en une séquence de mots homophones plus courts (/m#a/ : *m'a*, *m'as*). La question des corrélations entre les paramètres acoustico-prosodiques et les frontières de mot devient appropriée dans de telles situations.

La proportion de mots homophones dans un corpus de parole est ainsi en lien avec les spécificités de la langue. Au-delà de ces caractéristiques, le style de parole pourrait lui aussi contribuer à augmenter la proportion de séquences homophones en raison des prononciations réduites et des effets de la parole hypo-articulée [Lindblom, 1990].

Nous nous sommes surtout intéressées à des particularités liées aux aspects segmentaux (l'articulation des phonèmes) et prosodiques (l'accentuation et l'intonation) qui pourraient caractériser la prononciation. Nos analyses acoustiques et prosodiques reposent sur des catégories grammaticales ou bien sur la position d'une syllabe à l'intérieur d'un mot ou d'un syntagme nominal. Le terme de prosodie est largement utilisé pour indiquer accent, ton, stress, etc. au

niveau lexical et intonation au niveau post-lexical ou non-lexical [Hirst and Di Cristo, 1998; Lacheret-Dujour, 2000]. Les paramètres mesurables relevant du niveau prosodique sont principalement la fréquence fondamentale (f_0) liée à la hauteur de voix, la durée en lien avec le rythme et le débit de parole, et l'intensité pour exprimer la force ou la puissance de la voix.

Méthode

Nous avons fait appel à des techniques automatiques afin de caractériser l'ensemble des données. Le système d'alignement en phones du LIMSI [Gauvain *et al.*, 2005] a été utilisé pour mesurer des paramètres acoustiques tels que les formants, la fréquence fondamentale (f_0), la durée et l'intensité en nous appuyant sur les frontières temporelles des réalisations des phones données par un système d'alignement standard, ce qui nous a également permis de dégager des patrons prosodiques spécifiques. La durée minimale d'un segment est de 30 *ms*. Le logiciel PRAAT [Boersma and Weenink, 2008] a été utilisé afin d'extraire un certain nombre de paramètres acoustiques. Nous avons ainsi extrait les premiers trois formants (F1, F2, F3), la f_0 et l'intensité toutes les 5 *ms*. Pour chaque segment, un taux de voisement peut être calculé, correspondant au rapport donné par le nombre de points de mesure avec $f_0 > 0$ sur le nombre total de points de mesure. Un taux de voisement peut être calculé ainsi : une trame est considérée comme voisée dès lors que la f_0 y est définie ($Pv = \frac{nombre\ de\ trames\ voisées}{nombre\ de\ trames}$). Nos analyses se sont basées sur des catégories grammaticales ou bien sur la position des segments à l'intérieur d'un mot ou une phrase prosodique.

Classification automatique de mots fréquents homophones

En français, de nombreuses erreurs de transcription automatique de la parole sont souvent causées par des mots homophones fréquents, par exemple un verbe au participe passé prononcé de la même manière à l'infinitif (*toussé*, *tousser*), engendre des confusions lors de la transcription automatique. Nous nous sommes interrogées si les mots homophones étaient discriminables avec des informations acoustiques et prosodiques, en particulier s'il s'agit d'homophones issus de classes syntaxiques différentes ou ayant des positions différentes à l'intérieur des mots/syntagmes prosodiques. Pour cette étude, deux paires d'homophones *et*/*est* et *à*/*a* ont été choisies pour la classification automatique (cf. Tableau 1).

TAB. 1: Nombres d'occurrences de mots.

	ESTER (préparée, 66 heures)		PFC (spontanée, 11 heures)	
mot	#occ.	phonème	#occ.	phonème
à	20,4k	/a/	3,6k	/a/
a	11,3k	/a/	3,4k	/a/
et	19,1k	/e/	5,0k	/e/
est	14,5k	[ɛ]5,0k, [e]9,5k	6,2k	[ɛ]1,9k, [e]4,3k

Analyses acoustico-prosodiques

Nous avons effectué des analyses prosodiques concernant la durée, le taux de voisement, la f_0 et les co-occurrences de pauses. Pour la paire *et*/*est*, les distributions en termes de durée et taux de voisement montrent des différences nettes permettant de distinguer ces deux mots homophones. Cependant les mesures observées pour la paire *à*/*a* sont moins claires. Dans l'ensemble, la distribution de durée pour la conjonction *et* s'avère plus plate que celle du verbe *est*, tandis que la comparaison des mots de la paire *à*/*a* ne montre pas de différence significative. Nous avons noté également que les mots grammaticaux sont plus souvent précédés de pauses. Les mots grammaticaux (*et*/*à*) ont un taux de voisement plus faible que les verbes (*est*/*a*). Ces mesures suggèrent que les homophones, réalisés a priori avec les mêmes phonèmes, peuvent présenter des différences dans leur réalisation prosodique (f_0, durée, pause, etc.).

Choix d'attributs

Les mesures acoustiques ont montré des différences entre les homophones sélectionnés. Nous avons par la suite effectué des tests de classification automatique pour vérifier si ces différences acoustiques mesurées pouvaient être utilisées pour discriminer automatiquement ces paires homophones. Combinant f_0, formants, durée, taux de voisement, et co-occurrence de pauses, 62 attributs acoustico-prosodiques sont définis pour la classification automatique en utilisant le logiciel WEKA [Witten and Frank, 2005]. Les 62 attributs acoustiques et prosodiques ont été choisis pour modéliser à la fois le mot cible (**attributs intra-phonémiques**) et sa relation au contexte (**attributs inter-phonémiques**). Ces attributs sont :

Attributs intra-phonémiques (40) : durée, f_0 (moyenne par segment, début, milieu, fin), taux de voisement, trois formants (F1, F2, F3), intensité. Nous avons également calculé les différences (notées Δ) début-milieu, milieu-fin et début-fin pour la f_0, les trois formants et l'intensité.
Attributs inter-phonémiques (22) : durée, f_0, trois premiers formants, intensité, pauses. Le paramètre durée est mesuré comme suit : la différence entre la durée au centre du segment correspondant au mot cible et le centre de la voyelle précédente/suivante, même s'il y a des consonnes ou des pauses entre ces phonèmes. Pour la f_0, les trois formants et l'intensité, Δ a été calculée comme différence entre la valeur moyenne du phonème du mot cible et celle de la voyelle précédente et suivante, et entre ces deux voyelles précédant et suivant le mot cible. Les paramètres pause à gauche et pause à droite ont été également rajoutés.

Résultats de classification automatique

Pour classifier automatiquement les mots homophones à partir de ces attributs, nous avons testé 25 algorithmes implémentés dans le logiciel Weka (classification bayésienne, arbres, règles et fonction etc.). Les expériences de classification sont effectuées à l'aide de la méthode de validation croisée. Le tableau 2 montre l'algorithme ayant permis la meilleure discrimination de chaque paire, la moyenne des 10 meilleurs algorithmes par paire de mots et la moyenne des 25 algorithmes par paire de mots.
Nous avons obtenu des taux moyens d'identification entre 60 et 77% (cf. Tableau 2). Les résultats de la classification automatique utilisant soit l'ensemble des attributs, soit des sous-ensembles limités au niveau linguistique ou au 15 attributs sélectionnés caractérisant le segment ou son environnement, montrent que les catégories d'attributs prosodiques et d'attributs inter-segmentaux

ainsi que les 15 attributs sélectionnés sont aussi performantes que les résultats donnés par tous les 62 attributs. Parmi les 15 attributs sélectionnés, les attributs les plus importants sont pause gauche, Δintensité, F2, Δf_0, taux de voisement, et durée. Par ailleurs, la paire *et*/*est* a été mieux discriminée que la paire *à*/*a* pour les deux styles de parole. Cela va dans le sens des analyses acoustico-prosodiques où l'on observait que la paire *et*/*est* se distinguait mieux que la paire *à*/*a*. Cela s'explique également en partie par le fait qu'un tiers environ des occurrences du verbe *est* ne sont pas de vrais homophones (prononciation /ɛ/ pour *est*) de la conjonction *et*, ce qui engendre des attributs plus discriminants. Les résultats pour la paire *et*/*est* sont particulièrement intéressants pour le corpus PFC puisque la parole spontanée présente en général plus d'erreurs lors de la transcription automatique.

TAB. 2: Comparaison des % de classification des mots homophones en fonction des types d'attributs. Dans le tableau le meilleur %, la moyenne sur 10 meilleurs algorithmes et la moyenne sur 25 algorithmes sont montrés. Le nombre d'attributs pour chaque catégorie est marqué entre parenthèses. **En haut** : *et*/*est*, ESTER (gauche), PFC (droite). **En bas** : *à*/*a*, ESTER (gauche), PFC (droite).

Mots	*et* vs *est*					
Corpus	ESTER			PFC		
	meill.	10 meill.	moy.	meill.	10 maill.	moy.
tous (62)	79,8	77,8	**71,3**	83,1	81,1	**76,3**
formants (30)	67,5	65,9	62,3	66,6	65,3	62,7
prosodie (32)	79,5	77,7	**70,9**	82,4	81,0	**77,3**
intra- (40)	73,2	71,3	65,7	71,7	70,4	67,0
inter- (22)	75,7	74,4	**69,2**	81,2	80,5	**77,0**
15 meill. att. (15)	77,6	76,4	**70,5**	81,4	80,5	**76,9**
15 tous meill. att. (15)	76,1	75,0	**69,5**	80,4	80,3	**76,7**

Mots	*à* vs *a*					
Corpus	ESTER			PFC		
	meill.	10 meill.	moy.	meill.	10 meill.	moy.
tous (62)	72,9	71,4	66,3	69,4	66,4	61,6
formants (30)	69,0	67,7	64,3	62,7	61,2	58,5
prosodie (32)	72,3	70,6	65,6	67,7	65,9	60,7
intra- (40)	68,9	68,0	64,0	60,0	59,3	57,0
inter- (22)	71,0	70,1	65,5	65,9	65,1	60,1
15 meill. att. (15)	70,9	69,7	65,5	67,5	65,4	61,2
15 tous meill. att. (15)	68,9	67,8	64,2	62,1	60,9	58,4

Tests perceptifs de mots fréquents homophones

Enfin, des tests perceptifs ont été également menés pour estimer la capacité des humains à effectuer la même tâche de discrimination ainsi que les stratégies perceptives aboutissant à une

meilleure différenciation des homophones. Pour vérifier si des humains comptent sur des paramètres acoustico-prosodiques pour discriminer des mots homophones ou bien s'ils ont tendance à utiliser l'information contextuelle similaire à *n*-gram de modèles de langage (ML) pour des systèmes de RAP, deux types de tests perceptifs ont été menés : simulation perceptive d'un décodage grâce aux modèles acoustiques + de langage (MA + ML : décodage du mot cible en l'écoutant en contexte) et d'un modèle de langage (ML : décodage du mot cible grâce au contexte droit et gauche mais sans audio). Ces tests ont été menés en utilisant les données ESTER. Les n-grams sélectionnés pour le test (ici des 7-grams, c'est-à-dire 3 mots à gauche et à droite d'un mot cible correspondant au contexte maximal d'un modèle de langage) comportaient à la fois des séquences correctement transcrites par le système automatique et des séquences présentant des erreurs de transcription. La motivation de cette sélection était d'observer une différence éventuelle entre ces extraits et de l'associer à une ambigüité éventuellement plus forte des séquences ayant généré des erreurs de transcription automatique. Les résultats des tests perceptifs ont été mesurés en tant que taux d'erreur des mots cibles et ont été comparés avec la transcription de référence du corpus et avec la solution du système automatique.
Nous avons noté que les humains ont fait très peu d'erreurs sur les stimuli correspondant à des extraits correctement transcrits par le système automatique. Une augmentation importante dans le taux d'erreurs humain a été observée sur l'ensemble de stimuli concernant les extraits ayant généré des erreurs automatiques et surtout ceux se prêtant à une confusion réciproque *et*/*est*.

TAB. 3: Taux d'erreur de mtos sur l'ensemble de 4 stimuli par les conditions de la transcription automatique/humaine : RAP (critères de sélection) ; ML (test écrit sur l'ambigüité locale) ; MA+ML (test audio).

	TEM (taux d'erreur de mots)		
Stimuli	**RAP**	**Humains**	
Condition	**MA+ML**	**MA+ML**	**ML**
5 distracteurs	0	0	-
10 corrects	0	1,4	8,2
20 *et/est confusions symétriques*	100	25,5	27,6
48 autre erreurs d'*et/est* (6 sets/4 types/2 mots cibles)	100	16,0	-

Analyses prosodiques à grande échelle d'unités lexicales et syntaxiques

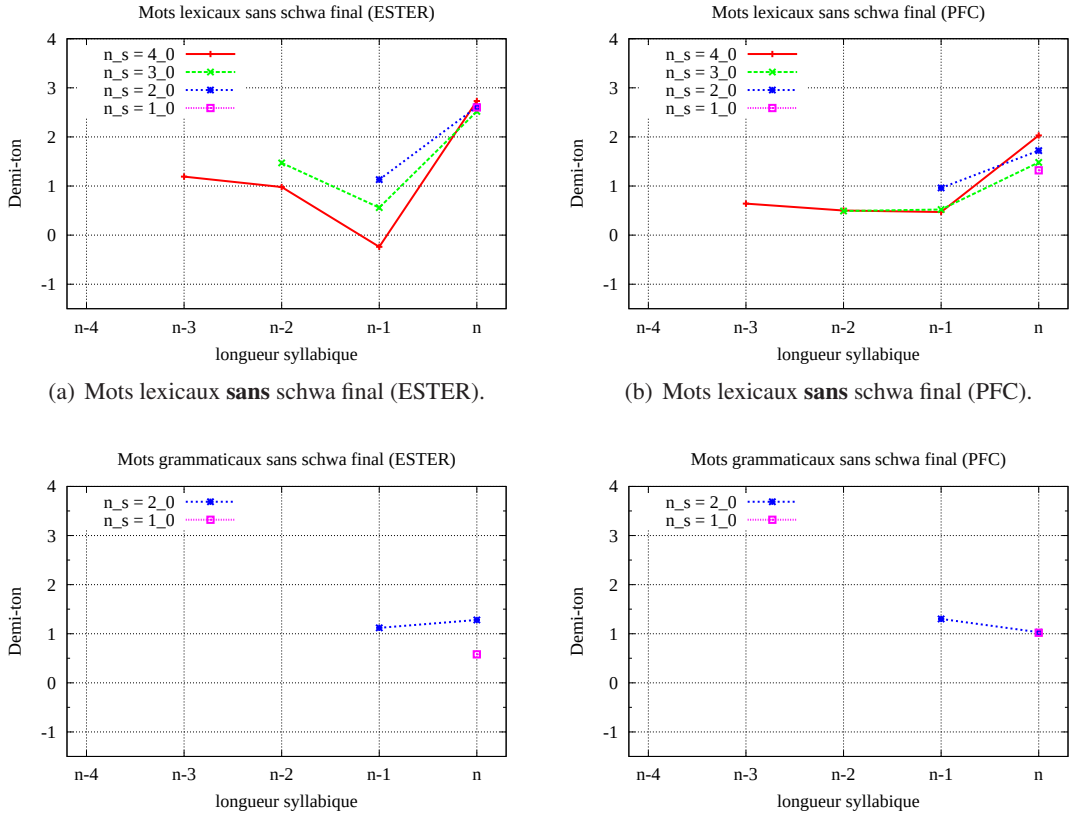
Lors d'une seconde étape de notre travail, nous avons entrepris des analyses de profils prosodiques moyens (f_0, durée, intensité) sur des classes de mots et de syntagmes nominaux en tenant compte du paramètre longueur syllabique afin de mettre en évidence des régularités. Cette étude est basée sur l'hypothèse que les variantes de prononciation sont les résultats de différentes contraintes prosodiques. De plus, l'étude proposée soulève la question du lien entre de nombreuses variantes de prononciation et les différences prosodiques mesurables. Les objectifs de cette étude sont : (1) d'établir une méthodologie automatisée pour étudier des propriétés prosodiques des mots français dans l'ensemble ; (2) de comparer des propriétés prosodiques à travers les différents styles de parole (préparée et spontanée).

Les variantes de prononciation sont souvent responsables de versions plus courtes ou plus longues des mots, des segments et/ou des syllabes ajouté(e)s ou éliminé(e)s, et introduisent des réalisations prosodiques spécifiques. Pour cette étude, nous avons utilisé des enregistrements de locuteurs masculins correspondant à 13 heures de parole manuellement transcrites du corpus TECHNOLANGUE-ESTER et à 6 heures de parole du corpus PFC. Les réalisations prosodiques des variantes de prononciation ont été étudiées à travers des profils prosodique en tenant compte de la distinction entre mots lexicaux et mots grammaticaux et de celle des noms par rapport aux syntagmes nominaux. Le tableau 4 montre une description quantitative des deux corpus en fonction des mots mono- et polysyllabiques. La méthodologie utilisée exploite les transcriptions automatiques grâce à l'alignement en phonèmes et mots, ainsi que l'étiquetage prosodique et morpho-syntaxique afin de comparer des profils prosodiques moyens en fonction des classes de mots de différentes longueurs syllabiques, de la présence ou non d'un schwa final, de la durée et de l'appartenance ou non du mot à un syntagme.

TAB. 4: Description quantitative des corpus ESTER et PFC en termes de mots (tokens) de longueur syllabique *n*, pour *n* = 0-4 . Les comptes sont séparés en fonction du schwa final réalisé (s=1, bas) ou non (s=0, haut). Concernant *classe syll.* : n_s indique n : longueur syllabique du mot ; s : présence(1)/absence(0) d'un schwa final.

n	*classe syll.*	*#mots*		*exemples*
	n_s	ESTER	PFC	
0	0_0	12578	13921	l' ; d' ; de
1	1_0	72249	65521	vingt ; reste
2	2_0	36027	20346	beaucoup
3	3_0	15994	4959	notamment
4	4_0	6053	1408	présidentielle
n	*classe syll.*	*#mots* + /ə/		*exemples*
0	0_1	12295	5056	de ; le ; que
1	1_1	3918	1642	reste ; test
2	2_1	2087	716	ministre
3	3_1	698	208	véritable
4	4_1	174	49	nationalistes

Les paramètres prosodiques (f_0, durée, intensité) sont mesurés à travers des profils moyens de longueurs syllabiques n du mot. En utilisant cette méthode, l'impact de la longueur syllabique du mot est considérée avec l'idée que les syllabes internes du mot sont plus susceptibles d'avoir un phénomène de réduction temporelle et d'engendrer des variantes de prononciation. Les profils étudiés peuvent ainsi fournir des contours synthétisés de f_0, durée et intensité selon les positions différentes des syllabes à l'intérieur des mots. Les données sont divisées en plusieurs sous-ensembles : mots lexicaux, mots grammaticaux, présence ou absence du schwa final, nom, syntagme nominal et différents styles de parole.



(a) Mots lexicaux **sans** schwa final (ESTER).

(b) Mots lexicaux **sans** schwa final (PFC).

(c) Mots grammaticaux **sans** schwa final (ESTER).

(d) Mots grammaticaux **sans** schwa final (PFC).

FIG. 1: Profils de f_0 pour les mots lexicaux (haut) et grammaticaux (bas) sans schwa final : ESTER (gauche) et PFC (droite).

Mots grammaticaux vs lexicaux

Profils de f_0 moyenne

Tout d'abord, nous présentons les profils de f_0. Les valeurs de f_0 correspondant aux mots lexicaux montrent des valeurs de f_0 plus élevée pour la syllabe finale n dans les deux corpus (cf. Figure 1). La présence de schwa final est en lien avec une augmentation globale de la f_0. A la parole spontanée du corpus PFC correspondent des profils plus plats de f_0 par rapport à la parole préparée du corpus ESTER. Les profils des mots grammaticaux montrent des valeurs de f_0 globalement plus basses que celles des mots lexicaux.

Profils de durée moyenne

Dans un deuxième temps nous nous sommes intéressées aux profils de durée. Des durées plus longues de la syllabe finale sont observées pour les mots lexicaux dans les deux corpus. La variation du paramètre durée de la syllabe finale montre des valeurs plus étendues lorsque le schwa final est absent. La parole spontanée (corpus PFC) montre une variation plus large par rapport à la syllabe finale en fonction de la longueur syllabique par rapport à la parole préparée. Concernant

les mots grammaticaux, des durées plus longues de la syllabe finale ne sont pas notées et cette observation est indépendante du paramètre style de parole.

Profils d'intensité moyenne

Troisièmement nous nous sommes intéressées aux paramètre intensité. Contrairement aux deux paramètres précédents (f_0 et durée), ce paramètre ne génère pas de profils particuliers de la syllabe finale. Pour les mots lexicaux, dans la plupart des cas, les valeurs d'intensité de la syllabe finale sont aussi hautes que celles de la première syllabe. Des valeurs plus basses sont toutefois à noter lorsque le schwa final est présent. Pour les mots grammaticaux, les valeurs d'intensité de la syllabe finale sont presque les mêmes que celles des mots lexicaux. Ces résultats montrent que l'intensité est un paramètre moins distinctif que la f_0 et la durée.

Impact de la durée sur la f_0

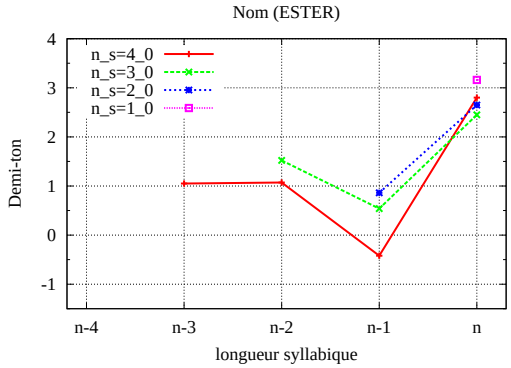
Les données ci-dessus soutiennent l'observation suivante : l'accentuation finale présente une corrélation avec la f_0 et la durée. Afin d'approfondir cet aspect, l'étude de l'impact de la durée sur la f_0 est menée pour les mots lexicaux sans schwa final. Les données sont divisées en deux catégories en fonction du débit de la parole (lent et rapide). Les items de la catégorie "lent" montrent des valeurs de f_0 plus élevées pour les deux styles de paroles. Moins de variation de f_0 dans la catégorie "rapide" est observée pour la parole spontanée en comparaison avec la catégorie "lent". Ces résultats confirment que la variation de durée de mots pourrait introduire de la variation de profils prosodiques, qui pourrait finalement être en lien avec la variation de la prononciation.

Nom vs syntagme nominal

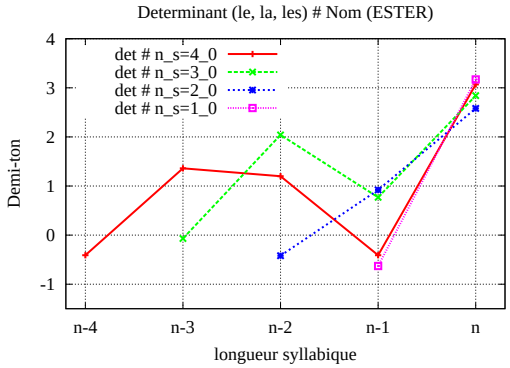
Après les études comparant des mots grammaticaux et lexicaux, nous étendons la mesure des profils prosodiques aux unités plus larges que sont les syntagmes ou mots/syntagmes prosodiques. Nous limitons pour l'instant nos analyses aux bigrammes de type `déterminant - nom` en comparaison avec les profils correspondant aux noms seuls. Les profils sont analysés pour répondre à la question des profils prosodiques en lien avec les frontières de mots. Les mesures sont destinées à répondre à la question suivante : le profil moyen d'un syntagme `déterminant - nom` de longueur n peut-il se distinguer de celui d'un nom de longueur n ? Les déterminants concernés sont *le*, *la* et *les*. La comparaison entre nom et syntagme nominal pour les deux styles de la parole considérés dans ce travail est effectuée en fonction des paramètres f_0, durée et intensité. Nous présentons ci-dessous les bigrammes sans schwa final.

Profils de f_0 moyenne

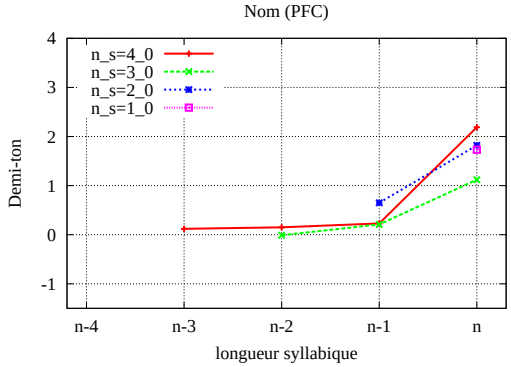
Les résultats (cf. Figure 2) montrent que les indices prosodiques pour les frontières de mots sont moins distinctifs pour la parole spontanée telle qu'illustrée par le corpus PFC d'entretiens face à face. Le style de parole pourrait être à l'origine des différences en termes de profils de f_0 entre déterminant et nom. Nous avons ainsi noté que les valeurs de f_0 sont plus basses pour la parole spontanée (PFC) que pour la parole préparée (ESTER). Cependant, pour les deux styles de parole, il est à noter que les valeurs de f_0 dans les syntagmes nominaux commencent par des valeurs plutôt basses et montent vers la première syllabe du nom suivant. Cette information concernant l'augmentation de la f_0 lors du passage du déterminant à la première syllabe du nom pourrait être



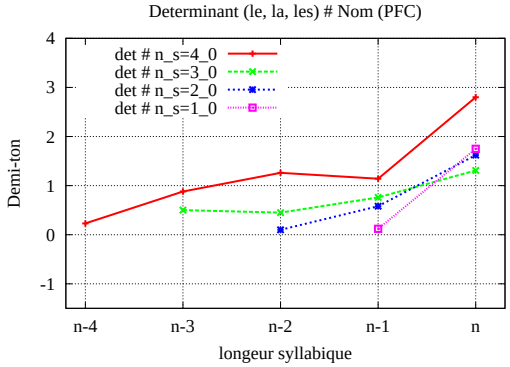
(a) Profils de f_0 moyenne de noms pour ESTER.



(b) Profils de f_0 moyenne de syntagme pour ESTER.



(c) Profils de f_0 moyenne de noms pour PFC.



(d) Profils de f_0 moyenne de syntagme pour PFC.

FIG. 2: Profils de f_0 moyenne pour longueur syllabique *n*. **Gauche :** Noms (sans schwa final), **Droite :** syntagmes nominaux (`déterminant–nom`). **Haut :** ESTER, **Bas :** PFC.

utilisée comme indice pour localiser des frontières de mots et pour désambigüiser des homophones comme *déblocage* et *des blocages*.

Profils de durée moyenne

Pour les profils de durée moyenne, des résultats similaires sont observés pour les deux corpus. Les profils de durée ne mettent pas en avant des indices pour distinguer des mots homophones comme "*lézard*" et "*les arts*" dans les deux styles de parole.

Profils d'intensité moyenne

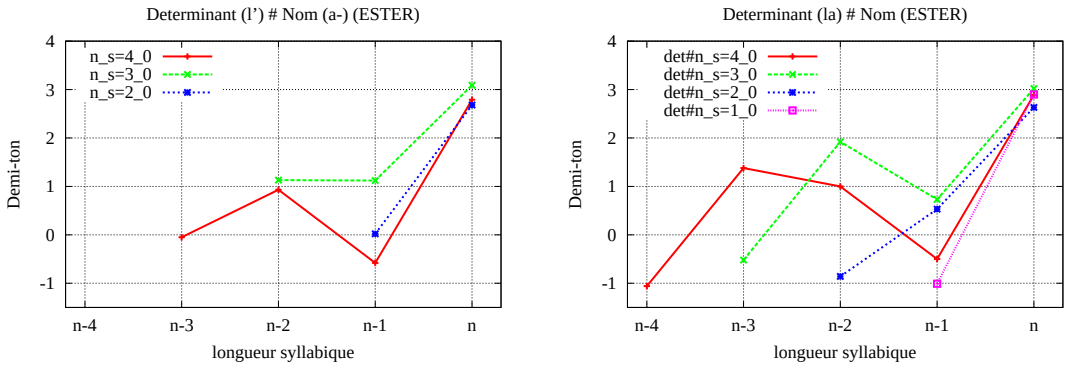
Les profils d'intensité montrent que les premières syllabes et les syllabes finales d'un nom ont presque les mêmes valeurs. A l'intérieur des syntagmes nominaux, les valeurs de la syllabe du déterminant, et de la première et la dernière syllabe du nom sont très proches. Il est intéressant de noter que les valeurs d'intensité sont légèrement plus basses sur le déterminant que sur la première syllabe du nom. Ceci pourrait également être un indice de segmentation des mots.

Analyses intervocaliques

Pour finir, des analyses de f_0 intervocaliques montrent que la plupart des déterminants présentent une chute de valeurs de f_0 en comparaison avec la voyelle précédente. Les premières et dernières syllabes des noms ont des valeurs de f_0 montantes en comparaison avec les voyelles précédentes. Les profils de durée intervocaliques montrent des durée longues entre la voyelle du déterminant et celle du mot précédent en soulignant une frontière de syntagme. Ces résultats suggèrent la présence d'indices mesurables contribuant à localiser les frontières de mots dans les grands corpus audio.

Syntagme nominal homophone

Une étude limitée aux syntagmes nominaux ambigus avec des homophones a été menée pour estimer la capacité des mesures prosodiques à distinguer entre des syntagmes comme *la fiche* vs *l'affiche*. Pour cette étude préliminaire nous avons utilisé uniquement les locuteurs masculins du corpus ESTER. Cette étude nous a permis d'aborder la question de la pertinance des "détails prosodiques fins" (*fine prosodic details*) pour discriminer des syntagmes nominaux homophones. Les profils liés à la f_0 soulignent une différence nette en valeurs moyennes concernant la première syllabe. En regardant la Figure3(b), il est à noter que les noms précédés par le déterminant *la* comme dans le syntagme *la fiche*, montrent des valeurs de f_0 plus basses. En effet, si la première voyelle de la séquence /la/ appartient au déterminant (*la*), la valeur de f_0 peut être basse. Par contre, dans le cas de *l'affiche* (Figure 3(a)) où le déterminant (*l'*) est suivi par une voyelle *a* qui appartient au nom, des valeurs intermédiaires par rapport aux résultats obtenus pour le syntagme *la fiche* (entre le déterminant *la* avec la f_0 basse et la première voyelle des mots lexicaux avec la f_0 haute) sont observées pour la première voyelle du syntagme de *l'affiche*. Ce résultat basé sur des données extraites de grands corpus s'avère en lien avec des résultats issus d'études psycholinguistiques [Spinelli *et al.*, 2007; Spinelli *et al.*, 2010] sur une petite quantité de matériel contrôlé.



(a) Profil de f_0 moyenne de syntagme nominal : *l'* # *a-* nom. (b) Profil de f_0 moyenne de syntagme nominal : *la* # nom).

FIG. 3: Profils de f_0 moyenne pour des longueur syllabique n en comparaison avec syntagme nominal ambigu pour ESTER. **Gauche :** *la* # nom, **Droite :** *l'* # *a-* nom

Perspectives

Dans cette thèse nous avons mené des études de profils prosodiques de mots et de syntagmes afin de déduire des spécificités liées aux variantes de prononciation. Nous nous sommes intéressées à la fois à des cas très ambigus comme les mots homophones et à des parties de discours en général (ici les noms dont les profils prosodiques ont été évalués selon les critères mots lexicaux/grammaticaux et inclusion ou non dans un syntagme nominal). Les propriétés prosodiques des mots et syntagmes telles que révélées par ces études pourraient être considérées comme des éléments préliminaires pour l'élaboration de règles de réalisation acoustico-prosodiques spécifiques aux classes de mots. Notre but à long terme est d'améliorer les modèles acoustiques des systèmes de RAP pour un traitement plus efficace des variantes de prononciations en utilisant des paramètres acoustico-prosodiques afin de réduire les taux d'erreurs de transcription automatique. D'un point de vue "recherche fondamentale", ces études ont également eu comme objectif de mettre en évidence le rôle proéminent des caractéristiques prosodiques en français. Enfin, le travail proposé dans cette thèse montre l'efficacité des études de grands corpus audio en utilisant des outils automatiques issus de la RAP pour l'extraction et la description des caractéristiques acoustiques et prosodiques d'une langue.
A plus long terme, ce type d'approches pourrait être exploité de manière plus large en lien avec la RAP à travers des applications telles que la détection d'entités-nommées, la recherche d'information, la compréhension de la parole, la détection d'événements, la traduction automatique, etc. Pour l'heure, le travail de localiser des focus et/ou des entités-nommées intégrant des caractéristiques acoustiques et prosodiques est en cours en utilisant un classifieur discriminant tel que les Champs Conditionnels Aléatoires (Conditional Random Fields - CRF) en collaboration avec des collègues du LIMSI.
À plus grande échelle, des études futures pourraient inclure des séquences d'étiquetage morpho-syntaxiques plus vastes, et des analyses plus détaillées de profils de f_0. Une extension de la méthodologie pour d'autres styles de paroles et langues pourrait être considérée afin de poursuivre des études acoustiques et prosodiques comparatives.

Introduction

This thesis focuses on acoustic and prosodic analyses of French from large-scale audio corpora portraying different speaking styles including prepared and spontaneous speech. We are especially interested in particularities of segmental phonetics and prosody that may characterize pronunciation in terms of grammatical categories and position within a word. A long-term objective of the proposed investigations concern improved automatic speech recognition (ASR) systems by improving pronunciation modeling. On a more short-term perspective, our goal is to increase our knowledge concerning pronunciation variation across speaking styles focusing on acoustic-prosodic features. In particular, these features attract our attention with the objective of discriminating between homophone words and word sequences. As a matter of fact, a major reason to ASR errors is homophones and near homophones, which arise more or less in different languages. French is known to admit a larger proportion of homophones than English for example. So, the proportion of homophones is first related to the characteristics of the studied language. Beyond these language specificities, speaking style may contribute to increase the proportion of homophonic sequences due to reduced pronunciations and hypo-articulated speech.

Nowadays ASR systems achieve high performances in transcribing speech and in particular prepared or semi-prepared data that are broadcast news-like recordings, although human speech recognition (HSR) still remains up to five times more accurate. Among the various reasons of the human-machine gap, we may cite our lack of knowledge concerning pronunciation variants represents a serious bottleneck to further improvements of ASR systems across conditions, and in particular across speaking styles. As a consequence, large collections of style-specific training data are required to implicitly capture pronunciation variation within the context-dependent acoustic phone models. ASR experience shows that the systematic introduction of a large number of pronunciation variants into the pronunciation dictionary does not tend to decrease word error rates, as more pronunciation variants tend to increase homophone pronunciations between different word types.

Words' pronunciation may vary according to communicative contexts (speaking styles, accents, type of interaction, etc.), this variation being observed at the acoustic, phonetic, and prosodic levels (hypo/hyper-articulation, variation of speech rate, of intensity, and of fundamental frequency, etc.). As already mentioned, acoustic modeling of ASR systems implicitly takes into account such sources of variation by relying on selections of specific training data. However, current acoustic models are not able to precisely model all levels of information as for instance fine-grained prosodic features in charge with the disambiguation of the syntactic structure of an utterance or with the semantic or pragmatic information.

Back to the ASR current challenges, it is relatively straightforward to introduce pronunciation

variants using phonological rules (e.g. schwa insertion or deletion rules, liaison rules, consonant cluster reduction rules, voicing assimilation rules...). Applying these rules to the full system vocabulary results in high pronunciation variant rates and, as mentioned above, in increased homophone rates. For frequently observed words in the acoustic training corpora, it is possible to select the most relevant variants from the observed tokens, and even estimate probabilities for all the different variants. However, the occurrence of lexical entries in the language follows a Zipf law (the frequency of any word type is inversely proportional to its frequency rank), which entails a small number of word types with a large number of observed tokens, and a large number of types with very few tokens in the training data. This means that reliable variant probabilities cannot be estimated for a large number of words of the vocabulary. To tackle this problem, we have to move from words to word classes, where each class comprises a large number of tokens.

The work proposed here focuses on experimenting a new methodology able to capture well-known prosodic properties of French such as word final lengthening, f_0 rise... in large-scale audio data. In particular, with respect to multiword homophones, we are interested in prosodic cues to word boundary location. Consequently, a first step of our work will focus on prosodic cues for automatic classification of homophone words. More specifically, we will focus on short homophonic function words which frequently occur in the French language, and which contribute to a significant amount of ASR transcription errors, either due to substitutions, deletions, or insertions. Furthermore, perceptual evaluations are conducted via an original ASR-related protocol to investigate why humans remain superior to discriminate between such ambiguous homophones.

As a second step, the study is extended to the whole French vocabulary (as observed in the speech corpora) by introducing prosody and syntax-related classes to answer the following question: how do prosodic parameters vary across grammatical categories, and as a function of the syllabic length of words and syntagms. To do so, we propose to look into fine-grained phonetic/prosodic details as observed in two selected speech corpora. The acoustic and prosodic investigations are conducted on large-scale transcribed audio corpora of French. The selected corpora provide a reliable background to analyze pronunciation variation and general acoustic and prosodic tendencies. ASR tools facilitate processing large oral data with different speaking styles and a significant number of speakers. The automatic alignment system segments audio streaming to provide phone/phonemic transcription that allows us to study acoustic and prosodic features in segments. We also make use of automatic morpho-syntactic labeling to carry out contrastive measurements involving acoustic and prosodic features at phonemic and lexical levels.

Finally, a third focus concerns the speaking style issue. What differences can be measured between different speaking styles? How can these differences be interpreted in light of ASR results? We want to recall that spontaneous speech entails much higher word error rates than prepared (journalistic) broadcast speech. In spontaneous face-to-face speech, involved speakers share more context information and as a consequence less information needs to be conveyed by the acoustic channel. This may at least partially explain the higher word error rates for spontaneous speech. The proposed study aims at clarifying how speaking style differences between prepared and spontaneous French are reflected on a prosodic level.

Outline

This manuscript is organized as follows. A first part gives a background overview (Chapters 1 and 2) on speech recognition and prosody and the second part develops the realized works (Chapters 3, 4 and 5) on automatic and human classification of homophones and large-scale prosodic parameter analyses of different style French corpora.

Chapter 1 gives an overview of automatic speech recognition (ASR) system as compared to human speech recognition (HSR). The variability of articulation as a consequence of factors such as speech tempo, communication context, etc. in speech introduces pronunciation variations. This pronunciation variation may cause recognition errors for both humans and machines. We describe how ASR systems process the variation of pronunciations. Then errors encountered by speech recognition are evaluated for both ASR and HSR.

Chapter 2 details prosodic studies existing in the literature. After introducing a brief general definition of prosody, French prosodic specificities are discussed. We also summarize prosodic components (fundamental frequency, intensity, duration, formants, and pause) as exploited in further studies. Then prosody modeling in the framework of speech technology is introduced. The role of the prosody for human speech recognition is also outlined.

Chapter 3 describes the two audio corpora of different speaking styles (prepared and spontaneous speech) investigated here. These two corpora were automatically segmented by the automatic alignment system.

Chapter 4 covers automatic and human classification of (near-)homophones. First, we give a brief introduction of ASR transcription errors, and in particular frequent errors due to homophone function words. Next, acoustic analyses concerning prosody of the investigated homophone pairs are presented. Then, acoustic and prosodic parameters were defined and measured for automatic discrimination of homophone words. Automatic feature selection was implemented to identify the most relevant parameters to discriminate between our homophone candidates. In the final section, an original perceptual test protocol was proposed and perceptual transcription tests were run to shed light on humans' perceptual strategies as compared to automatic classification results.

Chapter 5 is dedicated to prosodic parameter analyses using fundamental frequency, duration, and intensity. The originality of this study consists in relying on the combination of large corpora and automatic speech alignment together with a variety of lexical classes to give an average overview of French prosodic regularities as well as changes across speaking styles. We explore prosodic regularities of French words via average prosodic profiles. Some influential factors are taken into consideration: word syllable length, word-final schwa, duration, and part-of-speech. Comparisons are made of grammatical and lexical words, of n-syllabic lexical words and n-syllabic noun phrase. A final study focuses more particularly on acoustic-prosodic detail by exploring average differences between two types of homophonic determiner-noun phrases: for the first noun phrase type, the first vowel belongs to the determiner (*la* # noun); for the second type, the first vowel of the noun phrase corresponds to first vowel of the noun (*l'* # a-noun).

Part I

Background

Chapter 1

Automatic and human speech recognition

In this chapter, an overview of the current standard automatic speech recognition (ASR) systems is presented. An ASR system is aimed to convert audio signals into word text outputs. Despite of a good performance, there is still place for improvement of the current ASR systems in particular to process speech variations encountered in spontaneous speech (e.g. conversational). Spontaneous speech data are indeed responsible of high word errors. Some of the errors are penalizing for the automatic output: in the following a particular focus will be dedicated to the type of automatic errors currently encountered in the automatically transcribed spoken data, to their sources and experimentations conducted to understand and avoid them, in particular in comparison with human transcription (cf. section 1.3).

First of all, we present an overview of the standard state-of-the-art ASR systems and their general architecture. In order to understand the differences between the automatic and human speech transcription, a short presentation of voice mechanism is also provided in section 1.1.

Current ASR systems have around 10% of word error rate for read fluent speech like broadcast new shows. The performance degrades considerably for spontaneous speech with up to 30% WER (word error rate) [Adda-Decker and Lamel, 2005]. This is because spontaneous speech exhibits more variability in terms of pronunciation, articulation, speech tempo, etc. In the section 1.2, the attempt to adapt pronunciation variation for the ASR system is particularly presented.

1.1 Automatic speech recognition system

This section focuses on the ASR systems history and architecture.

1.1.1 Voice mechanism

The air expelled from the lungs vibrates the vocal folds (vocal cords) which place at the level of the laryngeal prominence known as Adam's Apple where is center of the neck. You can feel their vibration when you pronounce vowels. The vocal folds are opened when we do not produce the

sound and they are closed when we produced the sound which engender vibration of the vocal folds (cf. Figure 1.1). Generally men have longer and thicker vocal fold size than women, so that is why men have lower voice. Then the vibrate air through vocal folds goes through vocal tract (cavity from vocal folds to mouth or to nose where voice or sound can be produced for humans). The difference cavity forms using tongue position, degree of mouth open, produce different sounds. Voice which is uttered from human mouth, vibrates air. This air vibration reaches at the tympanic membrane (called also eardrum) in ears of a listener. Humans recognize this eardrum vibration as sound.

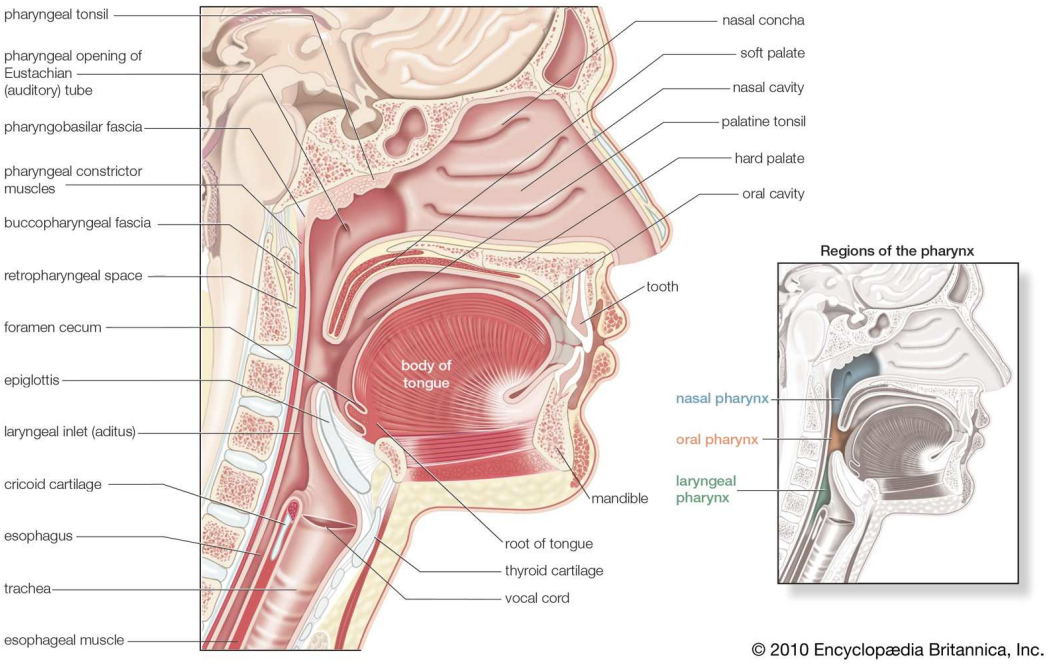


Figure 1.1: Vocal tract from Encyclopædia Britannica, retrieved from http://www.britannica.com/EBchecked/media/68641/Sagittal-section-of-the-pharynx.

1.1.2 Brief history of ASR

In order to adapt this human voice mechanism for recognizing sound to machine, humans have tried to develop the ASR system from its birth in 1950's. The ASR system has developed to recognize: phonemes, syllables, isolated words, connected words, read speech, conversational and spontaneous speech, conversational telephone speech (CTS). Further developments included more challenging tasks in terms of speakers and message complexity: from monologues (single speaker) to multiparty conversations (overlapping speech, specific items to capture audience attention, turn-taking, message co-elaboration, etc.) Some authors reviewed in detail the ASR developing history from its birth [Furui, 2005; Juang and Rabiner, 2005].

The first ASR system appeared in 1950's at Bell Laboratories [Davis *et al.*, 1952]. The first ASR system aimed at recognizing isolated ten digit words (from 1 to 9 and 0 'OH') for a single speaker

using first and second formants. This system got about 98% of word correct identification. Then fast improvements occurred: ten monosyllabic words for a single talker [Olson and Belar, 1956], a binary selection of phoneme classification [Wiren and Stubbs, 1956], 10-vowel (/b/ - vowel - /t/) recognition [Forgie and Forgie, 1956], a phoneme recognizer to identify four vowels and nine consonants [Fry, 1959; Denes, 1959], etc.

From the middle of 1960's, the ASR systems has progressed using algorithms. Vintsyuk [1968] applied dynamic programming (DP) method for connected word recognition. But his work had been unknown till 1980's in other countries. So this method was widespread in 1970's [Sakoe and Chiba, 1978]. Reddy [1966] also innovated continuous speech recognition research by dynamic tracking of phonemes through speech wave. Oppenheim [1968] adopted cepstral processing for speech.

In the 1970's, the use of fundamental pattern recognition technology to speech recognition was proposed: Linear Predictive Coding (LPC) [Itakura and Saito, 1970; Atal and Hanauer, 1971; Rabiner *et al.*, 1979]. Jelinek [1976] proposed continuous speech recognition using statistical methods. In 1971, the speech understanding project sponsored by the Advanced Research Projects Agency (ARPA)[1] [Klatt, 1977] of the U.S. Department of Defense started in the term of five-year program. The program aimed at building speech understanding systems with a small number of speakers and about a thousand words that could reach less than 10% of semantic error rate.

Most of current ASR systems are based on statistical modeling of speech in the 1980's: HMMs [Ferguson, 1980; Juang, 1985; Rabiner, 1989], n-gram model (language model) [Jelinek, 1985], Δ and $\Delta\Delta$ceptsrum [Furui, 1986], neural networks to speech recognition [Lippmann, 1987; Waibel *et al.*, 1989], etc. At the middle of 1980's, the DARPA threw a new program for speech and natural language. 1000-word database consists of read-speech sentences appropriate to a naval resource management (RM) task built around existing interactive database and graphics programs [Price *et al.*, 1988]. Unlike the former program with speaker-dependent tasks, this program needed to treat speaker-independent, speaker-adaptive and speaker-dependent speech recognition. The ASR system was applied for a dictation system as *Tangora* developed at IBM [Averbuch *et al.*, 1987; Das and Picheny, 1996].

In the 1990's, significant improvements on statistical models or pattern recognition have made the progress of the ASR systems suitable from read to spontaneous speech, speaker-dependent to speaker-independent. Some applications for the automotive navigation system, the human-machine dialogue system etc. have been developed. Along with the DARPA program in 1990's, large corpora, to transcribe read sentence speech of Broadcast News (BN) [Pallett *et al.*, b; Pallett *et al.*, a; Graff, 2002] and conversational speech [Godfrey *et al.*, 1992], have been collected. These large vocabulary speech corpora helped to improve the ASR systems. The benchmark tests of NIST (National Institute of Standards and Technology) evaluations supported by DARPA have spread to other domains: language recognition, speaker recognition, spoken document retrieval, machine translation, etc.[2]

In 1997 in France, a first evaluation campaign for French [Dolmazon *et al.*, 1997] was organized by ARC-B1 project "Dictée vocale" of the AUF (*Agence Universitaire de la Francophone*, Association of Universities of the Francophonie), based on journal read speech recognition. This evaluation contributed to enhance the ASR systems for French language [Adda *et al.*, 1997].

[1] Called now the DARPA (Defense Advanced Research Projects Agency)

[2] See in http://www.itl.nist.gov/iad/mig//tests/index.html.

In 2000's, applications of the ASR systems to mobile phone [Varga *et al.*, 2002], to subtitle for T.V. [Imai *et al.*, 2000], to speech-to-speech translation [Waibel *et al.*, 2003; Gao *et al.*, 2006], etc. have developed. Other evaluations for French language [Gravier *et al.*, 2004a; Galliano *et al.*, 2009] were also held. In order to increase recognition performance for spontaneous speech in Japanese, large vocabulary corpus including about 7 millions of words was collected [Maekawa, 2003]. The speaking types for the ASR systems have been shifted from read speech to spontaneous speech. The DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) program has reinforced robust speech recognition technology in order to address a range of languages and speaking styles.

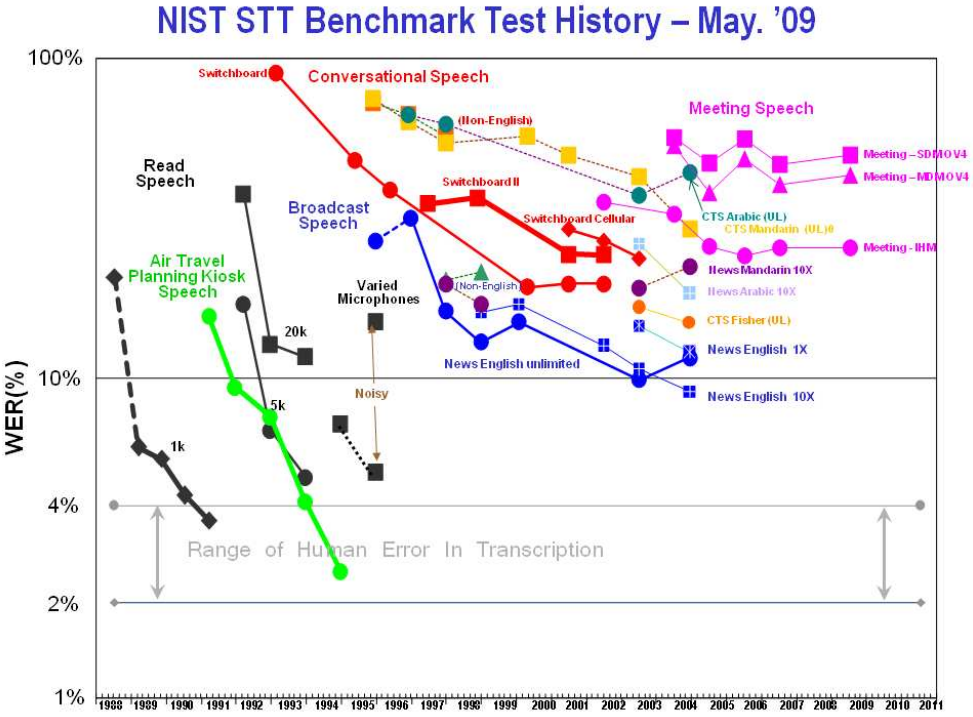


Figure 1.2: ASR Evaluations at DARPA/NIST from 1988–2009.

The progress of the ASR systems can be seen in Figure 1.2, which demonstrates word error rate (WER) of DARPA/NIST campaigns from its starting of 1988[3]. These results reveal that the ASR systems have good performance for read speech and prepared speech such as BN with less or around 10% of WER. However, in spite of improvement in the ASR systems, they keep quite high WER for conversational or spontaneous speech. So the challenge for amelioration of the current speech recognition is to enhance the ASR systems in order to become suitable for dealing with spontaneous speech and its numerous sources of variation.

[3] See in http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html.

1.1.3 ASR architecture

The purpose of ASR systems is to convert speech signals (human voice or spoken words) into text forms. To do that, statistical methods are widely used in the state-of-the-art ASR systems. The state-of-the-art ASR system are able to process different speakers (speaker-independent), and a significant amount of words which are connected and/or co-articulated (continuous speech), while at their beginning ASR systems mostly processed discrete words with silence before and after each word (isolated word recognition). The ASR systems able to use large vocabulary (about 20,000 to 60,000 words [Jurafsky and Martin, 2008a]) are called large vocabulary continuous speech recognition (LVCSR).

From 1970's, the ASR systems using statistical methods [Jelinek, 1998] have been applied. The purpose of the ASR systems is to link the speech (acoustic observation) input to the word sequence output. This can be represented as probabilistic formula of $P(W|O)$.

O denotes a sequence of acoustic observations or symbols. Each individual observation or symbol is represented by small letter o. In continuous speech, sounds are connected. Sounds are segmented by a certain time, for example every 30 milliseconds (ms), in order to extract acoustic information (frequencies, energy, etc.) from its segment that will be an observation or a symbol:

$$O = o_1, o_2, o_3, ..., o_m \tag{1.1}$$

W denotes a sequence of words including individual word w:

$$W = w_1, w_2, w_3, ..., w_n \tag{1.2}$$

The formula $P(W|O)$ represents the probability that a sequence of words W is uttered, under the condition that a sequence of acoustic observation O is determined[4]. The probability with such under the condition (the condition here is that 'acoustic observation sequence was determined') is called 'conditional probability'.

The ASR systems need to find the most probable word sequence from observed acoustic parameters. This can be represented as follows:

$$\hat{W} = \arg\max_{W} P(W|O) \tag{1.3}$$

In mathematics, '$\arg\max$ ' means the argument of the maximum. That is to say, the function $\arg\max_{x} f_{(x)}$ seeks the x value which maximizes the function $f_{(x)}$. Thus Equation 1.3 is to find the most likely word sequence ($\hat{W}$) obtained by $\arg\max_{W} P(W|O)$, in which the ASR systems try to search the word sequence W that maximize the probability of the word sequence W, given the acoustic observations O. Any conditional probability formula can be rewritten in Bayes' theorem. Thus, the equation $P(W|O)$ can be represented as follows:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{1.4}$$

[4]Conditional probability is often written in this way: the probability of W, given O.

Recall that the probability formula $P(W|O)$ is the probability of event W (a sequence of words), occurred under the condition of event O (a sequence of acoustic observations). So the formula $P(W|O)$ is called conditional probability, and also called the posterior probability in the Bayes' theorem. The posterior probability is given by the prior probability $P(W)$ multiplied the likelihood function $P(O|W)$. $P(W)$ addresses the probability of the event W, and called the prior probability. The prior probability does not need any condition or information of the event O. $P(O|W)$ is the conditional probability that when a sequence of words W is uttered, a sequence of acoustic observations O will be yielded. For the posterior probability $P(W|O)$, the result of event W is observed under the certain condition O. Inversely, likelihood function infers likely conditions O from the result W. The denominator of the acoustic observations $P(O)$ does not depend on the word sequences W. So the maximization for the word sequence W does not need introducing $P(O)$. Therefore Equation 1.4 can be simplified as:

$$P(W|O) = P(O|W)P(W) \tag{1.5}$$

Thus Equation 1.3 can be written as follows:

$$\hat{W} = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} = \arg\max_{W} P(O|W)P(W) \tag{1.6}$$

Recall that the ASR aims at finding the most probable word sequence $\hat{W}$ given the acoustic observations O. This can be computed by the two probability formulas $P(O|W)$ and $P(W)$. $P(O|W)$ is the acoustic observation likelihood given word sequence W, and this model is called the acoustic model. $P(W)$ includes the prior knowledge of word sequence and known as the language model. Figure 1.3 presents a diagram of a general ASR system.

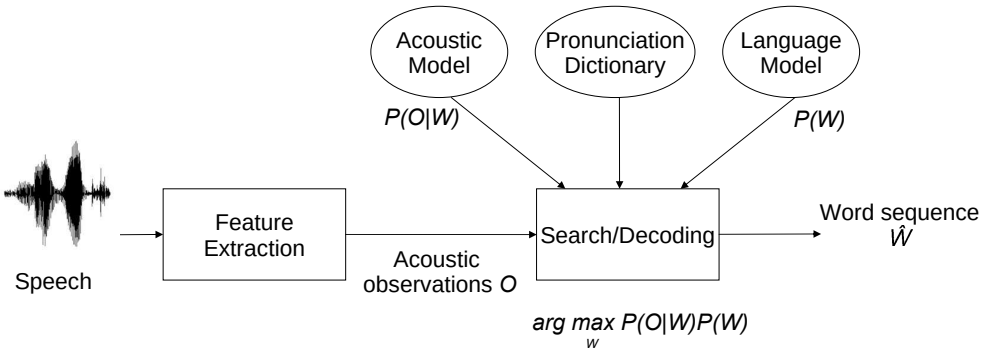


Figure 1.3: ASR system.

Acoustic model (AM) is used for modeling human voice or other sound like music, noise, breath etc., so as to find a certain phone/phoneme corresponding with a certain sound waveform. Such a system is called phone-based system. Here phones correspond to phonemes and also to allophones (several articulatory realizations for a phoneme) [Lamel and Gauvain, 2003]. A three-state left-to-

right Hidden Markov Model (HMM) [Rabiner, 1989; Young, 1996] is used for most of the ASR systems. Phone models are able to consider left and right context (neighbor phones), and this is called triphone. Or phone models may just investigate a single phone context (right- or left-context), or even without taking account of contexts (context-independent model). The variability of pronunciation is adjusted thanks to pronunciation dictionary to transform phoneme chains into words.

Pronunciation dictionary is used to model word pronunciation including variants. That is to say, pronunciation dictionary has a role to associate acoustic-level representations (phones/phonemes) yielded by AM to word representations. Table 1.1 gives some examples of pronunciation variations in French. The word "*ils*" ("they") has two possible pronunciation representations: its canonical pronunciation /il/ and its variant with liaison[5] /ilz/.

Pronunciation dictionary also admits an optional schwa /ə/ or also called mute *e* realization: the words *ministre* ("minister") and *seize* ("sixteen") in French can have an optional schwa at the end of the word finishing by a consonant [Adda-Decker and Lamel, 1998; Adda-Decker and Lamel, 1999; Adda-Decker *et al.*, 1999b]. Schwa deletion also can be seen inside the word "*venir*". Please see in section 1.2 for more detail on pronunciation variation.

Table 1.1: Examples of pronunciation including variants in the lexicon of the LIMSI ASR system.

word	**pronunciation**			
	canonical	variant		
`ils` ("they")	il	ilz		
`seize` ("sixteen")	sɛz	sɛzə		
`venir` ("come")	vəniʁ	vniʁ		
`est` ("is")	ɛ	ɛt	e	et
`ministre` ("minister")	ministʁ	ministʁə	minist	mnist

Language model (LM) is used to estimate the probability of word sequences. On one hand acoustic model comes from sound information; on the other hand language model is based on linguistic information, as knowledge of word sequences. For LM, statistical n-gram models are popular where n-grams mean sequence of n words. If certain n-gram words are frequent in texts, these word sequences have high probabilities. If rare sequences appear, less weight is assigned to these word sequences.

1.2 Pronunciation variations

Along with the ASR development from read speech to spontaneous speech and from monologue to dialogue, various factors which influence ASR results such as speaker characteristics (accent, gender, age, etc.), speaking styles (read, prepared, spontaneous, conversational), contexts (formal,

[5] In French, most of cases, a word-final consonant is not pronounced. However there are some cases that a word-final consonant is pronounced with the combination of a following vowel of the next word.

casual) are considered. Among such factors, pronunciation variations can be found within the spoken productions of a same speaker (intraspeaker) and different speakers (interspeaker) [Strik and Cucchiarini, 1999; Wester, 2002; Kessens, 2002].

Intraspeaker pronunciation variations refer to variations due to the differences in words pronunciations of a same speaker as a consequence of diverse factors. For instance, speaking style is one of such factors. Reading written texts is different from spontaneous utterances. While articulation is clear and speech tempo is quite constant and not fast when reading isolated speech, spontaneous speech is more variable in terms of fluency and in speech rate. That means that read speech can have more canonical pronunciations, whereas spontaneous speech can have more variants. Moreover, disfluencies (e.g. repetitions, repairs, filled pauses, false starts, lengthened schwa/vowel) highly occur in spontaneous speech compared with read and prepared speaking styles. Non lexical spoken events such as disfluencies represent one of the challenges of the current ASR systems. Contexts, which are circumstances of the spoken message delivery, are also one of the intraspeaker sources of variation. In the framework spontaneous speech data, it has been noticed that if the context change (conversation between unknown or less known persons (formal) and friends (casual)), articulations and speech rate can change as well. Labov [1972] stated that stylistic variations were a result of variations in the degree of formality of speech. Lindblom [1990] described the change of different articulation according to contexts as H&H (hyper- and hypo) speech theory. Carré and Hombert [2002] explained that a speaker produce clear speech (hyper-articulation) so that his/her listener(s) can understand a new message. As for hypo-articulation, a speaker and his/her listener(s) have shared information so that the speaker needs less constraints of articulation, which results in reduced articulation.

Pronunciation variations between different speakers may also occur as a consequence of the anatomical differences due to the gender or the age of the speaker. The geographic background engenders variations related to accents and dialects. Finally, the socioeconomic background and education level can also influence pronunciation variations.

Thus adapting the various speaking styles to the ASR systems is a crucial task, especially for spontaneous speech, since the ASR systems for spontaneous speech have higher Word Error Rate (WER) values [Adda-Decker *et al.*, 2003]. So considering canonical pronunciation, variants, and coarticulations is required for the ASR systems.

1.2.1 Pronunciation variation modeling for ASR

Two workshops concerning pronunciation variation took place in 1998 and 2002. One is the ESCA Tutorial and Research Workshop "Modeling pronunciation variation for automatic speech recognition" in 1998[6]; the other is ISCA Tutorial and Research Workshop named as "Modeling pronunciation variation for automatic speech recognition (PMLA)" in 2002 [7]. Various propositions of pronunciation variation modeling for ASR were presented at the two workshops.

[6] In Rolduc, the Netherlands from 4 to 6 May 1998.

[7] In Aspen Lodge, Estes Park, Colorado, USA, from 14 to 15 September 2002.

1.2.1.1 Workshop of "Modeling pronunciation variation for ASR"

Strik and Cucchiarini reviewed the workshop in [1998] and also gave an overview of the literature on modeling pronunciation variation for ASR in [Strik and Cucchiarini, 1999]. The review provides a special focus on the sources of information concerning variations in speech. The major approaches to derive information on pronunciation variation are the data-driven and the knowledge-based methods. Both methods need manually (e.g. [Riley *et al.*, 1999; Saraçlar and Khudanpur, 2004]) or automatically (e.g. [Adda-Decker and Lamel, 1999; Wester and Fosler-Lussier, 2000]) transcribed data from the acoustic signals in order to obtain information. The data-derived approaches are that pronunciation variation information can be directly obtained from acoustic signal data. In contrast, the knowledge-based methods need to get pronunciation variation information from sources that already exist in the linguistic literature with phonological or phonetic knowledge.

The second considered aspect is the types of pronunciation variation: within-word and cross-word variation. As indicated in section 1.1.3, the ASR system uses a pronunciation dictionary, called also lexicon, pronunciation variation modeling is mostly generated in the lexicon. At the word level, one may notice that a word can have several pronunciation candidates from canonical pronunciation to its variant(s). Variants can occur by substitutions, insertions and deletions of phones or phonemes related to canonical pronunciation. This type of variation is within-word variation. As for cross-word variation, multiwords (sequences of words) are treated as one entity in the lexicon [Sloboda and Waibel, 1996; Riley *et al.*, 1999]. The study of pronunciation variation about multiwords is emphasized in [Binnenpoorte *et al.*, 2005] to improve automatic speech recognition and automatic phonetic transcription.

A particular attention is dedicated to the representation of the information concerning the pronunciation variants. That is to say, the pronunciation variation information can be formalized or not [Strik and Cucchiarini, 1999]. In a data-driven method, the formalizations are done by e.g. rewrite rules, decision trees, artificial neural network, or phone confusion matrix. In a knowledge-based method, the formalizations of pronunciation variation information are extracted from linguistic studies. The obtained formalizations are generally added in the lexicon as optional phonological rules such as substitutions, insertions and deletions of phones or phonemes. The alternative choice is not using formalizations. It means that all possible variants are listed in the pronunciation dictionary without being generated by some rules.

Finally, a significant aspect, the level of pronunciation modeling represents a crucial aspect: the pronunciation may be modeled at the pronunciation dictionary level, the acoustic model (AM) level, and the language model (LM) level. In addition, these three levels are linked with each other. At the pronunciation dictionary level, pronunciation variation is generally generated manually or automatically by adding pronunciation variants to the pronunciation dictionary. Most of the time it is automatically generated and different methods are proposed as follows: rules, artificial neural networks, grapheme-to-phoneme converters, phone(me) recognizers, optimization with maximum likelihood criterion, and decision trees. The problem of adding pronunciation variants to the pronunciation dictionary is that added variant representations can confuse other entries by the same phone or phoneme representations with different entries. This problem can lead the increase of errors to correctly recognize words. In order to avoid the increase of word confusability, finding the counterbalance between errors and adding variants is crucial. To determine it, several studies were carried out: frequency of occurrence of the variants, maximum likelihood criterion,

confidence measures, degree of confusability between the variants, use of multiword entries.

By optimizing acoustic models, pronunciation variation can be represented at the acoustic model level. The optimization can be made by using forced alignment or also referred to as forced recognition in the training phase. During forced alignment, the recognizer is given the orthographic transcription which is to be recognized. The role of forced alignment is to find the most likely string of phones/phonemes that matches the provided words from the acoustic signal. In this way, this new phonetic/phonemic transcriptions can be obtained. These phones/phonemes are time-aligned.

At the language model level, as mentioned above, the most popular way to deal with the pronunciation variants is to add variants in the pronunciation dictionary. In this case, there is no change in LM. The second proposition is to integrate the variants to compute the n-grams. The third solution is to adopt the intermediate level in the general ASR system. This solution aims at finding the most likely sequence of words from a corresponding string of variants and a sequence of acoustic observations.

1.2.1.2 Workshop of PMLA

In 2002, four years later from the workshop "Modeling pronunciation variation for automatic speech recognition", another workshop concerning pronunciation modeling named "Pronunciation modeling and lexicon adaptation for spoken language technology (PMLA)" was held. In this workshop, new approaches in terms of holding pronunciation variation processing have been presented.

Bates and Ostendorf [2002] used prosodic features (fundamental frequency, duration, and energy) and word cues (part-of-speech label and content/function word tags, location of the word in the utterance) in their pronunciation model using a decision tree rules. The used data were phonetically hand-transcribed. Bates and Ostendorf got a slight improvement in phone error rate over the baseline model.

Bell *et al.* [2002] investigated the role of predictability on content word duration. Higher frequency words are likely to have shorter durations since higher frequency words are predictable from neighboring words. Bell *et al.* [2002] studied with regard to word frequency, conditional and joint probabilities. Bell *et al.* revealed that a word's duration is influenced by two predictability variables: word frequency and the conditional probability of a word given the following word.

Syllable unit was also adapted to [Adda-Decker *et al.*, 2002; Sethy *et al.*, 2002]. Adda-Decker *et al.* [2002] studied syllabic structure and its variation for French in which liaison and word final-schwa can be perplexities for pronunciation variation modeling. Restructuring syllables due to omitted vowels or syllables are also investigated. Sethy, Narayanan, and Parthasarthy [2002] used a syllable-based approach for spoken name recognition. Pronunciation variation modeling with syllables as the acoustic unit was proposed and this syllable-based system could improve error rate.

These workshops proposed different approaches for pronunciation variation modeling to improve the ASR system. In the following, we address the question of the pronunciation variation on the ASR systems performance in French.

1.2.2 Pronunciation variation modeling for French

Fouché [1969] described pronunciation variants in French as a consequence of several factors, like speaking style, speaking rate, individual speaker habits and dialectal region. The most common pronunciation variation in French is liaison and optional schwa /ə/ [Adda-Decker *et al.*, 1999b]. These two phenomena occur at the word boundaries, and they can sometimes lead word errors.

1.2.2.1 Liaison

Liaison is that a mute or latent word-final consonant is pronounced due to the context of a following word starting with a vowel, glide, or mute *h*.

For the examples of liaisons, the word "*ils*" ("they") can be pronounced as /il/ without considering the final consonant. The two following connected words "*ils sont*" ('they are') are uttered as /ilsɔ̃/. Since a first phone/phoneme of a second word is /s/, the word "*ils*" conserves its pronunciation without liaison. But the two following words "*ils ont*" ("they have"), given the influence of the second word starting with a vowel, are pronounced /ilzɔ̃/. In the framework of the liaison phenomenon, it is very common to insert the phoneme /z/ in the words ending with an *-s* or an *-x* which precede a word starting with a vowel [Adda-Decker and Lamel, 1999]. In the case of the word "*ils*", two different pronunciations are added in the pronunciation dictionary: /il/ and /ilz/. In fact, the number of consonants used for liaison is limited [Adda-Decker *et al.*, 1999b] as follows: /z/, /t/, /n/, /r/, /p/ phonemes which correspond to the written forms {*-s*, *-z*, *-x*} for /z/, {*-d*, *-t*}, *-n*, *-r*, *-p*, respectively.

Liaison is described as: obligatory, optional, or interdictory liaison. Boula de Mareüil, Adda-Decker, and Gendner [2003] examined liaison realizations in the 100 hours of read newspaper speech on the base of the 20 liaison rules described in the literature. The result showed about half of liaison contexts (45%) are realized among 90k liaison contexts in the corpus.

The followings are the examples of these three types of liaison categorized in [Boula de Mareüil *et al.*, 2003]:

Obligatory liaison: After determiner, between adjective and noun, after monosyllabic adverb other than "*pas*" ("not"), between verb and pronoun, after clitic pronoun, after auxiliary verb of 3rd person, after monosyllabic preposition, after the word "*quand*" ("when").

Optional cases: between plural noun and plural adjective, after the word "*pas*" ("not"), after participle, after the word "*mais*" ("but").

Interdictory cases: after non clitic pronoun, after main verb, after singular common noun, after polysyllabic adverb/conjunction/preposition, after the word "*et*" ("and"), between adjective and non noun word.

To conclude, it is worth noticing that the liaison in French shows heterogeneous realizations which make difficult the pronunciation variants processing within the ASR framework. As a consequence, building the pronunciation dictionary represents a major task.

1.2.2.2 Schwa

Insertion and deletion of schwa are a major phenomenon of pronunciation variation in French. Insertion and deletion of schwa may change syllabic structure. Some schwas are obligatory. Schwa is thus pronounced within a word (non-final) position, when a schwa is preceded by two or more consonants (e.g. "*brebis*" ("sheep") /brəbi/), or followed by a liquid+/j/ (e.g. "*chancelier*" ("chancellor") /ʃãsəlje/) [Adda-Decker *et al.*, 1999b]. In the other cases, schwas can be optional.

Insertion of schwa may occur at the non-final position of a word (epenthesis) or at the end of a word. When is the schwa inserted? According to [Adda-Decker *et al.*, 1999b], schwa is optional when the words are finishing orthographically by *e* and also *-es* or *-ent*. For the non-final cases, the schwa is optional in compound words formed with *garde-* and *porte-* if the second element is more than disyllabic (e.g. "*garde-manger*" ("pantry") /gard(ə)mãʒe/ , "*porte-bonheur*" ("charm") /pɔrt(ə)bɔnœr/). There are some cases for inserting final-schwas. Final-schwas are realized in words finishing by consonants. As described in [Adda-Decker, 2007], the optional word-final schwa is found at the end of the word as in the following example. In the reference transcription made by human experts one may find *en fait* "in fact", but these words are transcribed by the ASR systems with the addition of the word "*de*" like "*en fait **de***" ("in fact of"). The reason is that the speaker inserted a schwa at the end of the word "*fait*" Such cases are very confusing for the ASR system and represent a source of transcription errors. If a schwa is realized in the acoustic observation and this schwa is absent in the pronunciation dictionary, the decoder tries to find another solution by choosing or inserting another word for the schwa, most of the time a monosyllabic function word (e.g. article, conjunction, pronoun, etc.) [Adda-Decker *et al.*, 1999b].

Inversely, errors may also occur if a realized schwa in the pronunciation dictionary and in the acoustic model is absent in the effective speech. [Bürki *et al.*, 2007] investigated schwa elision from 4185 optional schwas in a broadcast news corpus. The authors found that 29% of schwas were deleted. The schwa deletion can occur at the intraword level but also between two words, especially at the end of a grammatical word [Fougeron and Steriade, 1999]. For example, the word "*devenir*" "become" having the canonical pronunciation /dəvənir/, can occur with the middle schwa deleted, forming then /dɛvnir/. Concerning the schwa deletion between two words: "*tout le monde*" "every one" with the canonical transcription /tuləmɔ̃d/, the schwa in the determinant "*le*" can be often omitted engendering the pronunciation /tulmɔ̃d/. The schwa deletion is responsible for word omission or confusion with a preceding or following word, via the phenomena described as assimilation or preceding/following word change. Here are some examples of schwa deletion from [Adda-Decker, 2007]: "*tout le temps*" in the reference transcription made by humans, is decoded as "*tout temps*" without the determiner "*le*" by the ASR system. Another example is "*quai de Seine*" in the reference, is transformed as "*quête seine*" by the machine. The word "*de*" is assimilated in the preceding word.

1.2.2.3 Variations for the spontaneous speech

The study described in [Adda-Decker *et al.*, 1999b] revealed that schwa is more often realized for both of final schwa and non-final schwa in read speech than spontaneous speech. Such observation suggests that spontaneous speech request a specific processing of the phenomenon for schwa deletion.

Dufour [2010, p.131] investigated pronunciation variation in the spontaneous speech in French in the framework of the EPAC project [Estève *et al.*, 2010]. The author proposed pronunciation variations not only for schwa elisions, but also for the deletion of [l], such as in the word "*il*" ("he") [il], pronounced [i] in spontaneous speech. Deletions of nasal vowel, the consonant [ʁ], and the semivowel [ɥ] are also investigated. As an example, frequent word "*parce-que*" ("because") [paʁs(ə)kə] is concerned with the deletion of /ʁ/ generating the variant [pas(ə)kə].

Finally, in [Dufour, 2010] it is underlined that the pronunciation can change from voiced consonants to voiceless consonants which locates in front of voiceless consonants. For the example in the phrase "*je pense*" ("I think") with the canonical pronunciation [ʒ(ə)pɑ̃s(ə)], the pronoun word "*je*" ("I") [ʒ(ə)] deletes the optional final-schwa and this voiced consonant [ʒ] will change the voiceless consonant [ʃ] in front of the first voiceless phoneme of the following word "*pense*", like in "*j' pense*" [ʃpɑ̃s(ə)]. Such observations are considered by the author when building an ASR system for spontaneous speech in French.

1.3 Errors

The preceding section (cf. section 1.2), revealed that pronunciation variations cause the speech recognition errors. However they are not the only problems. Where come from other speech recognition errors? As mentioned above, spontaneous, and especially conversational speech have difficulties for the ASR systems owing to numerous disfluencies, huge pronunciation variants, and acoustic and prosodic variability. In this section, it is presented which kinds of errors are seen in the ASR system compared to the human recognition system (HRS).

1.3.1 Errors by ASR

WER (word error rate) is the metric widely employed to evaluate the performance of an ASR system (cf. Equation 1.7). WER is computed as the total number of insertions, deletions, and substitutions made by the recognizer in comparison with the total number of reference words (normally transcribed by humans). Table1.2 shows examples of errors in French [Adda-Decker, 2007]. Among the errors, substitutions represent 40% (2 substitutions/5 reference words $\times$ 100), whereas deletions represent 25% of WER and 50% for insertions.

$$WER = \frac{\text{Substitutions + Insertions + Deletions}}{\text{Reference word count}} \times 100 \tag{1.7}$$

Explaining the source of ASR errors represents a recurrent topic among the studies dedicated to ASR performances. Recently [Goldwater *et al.*, 2010] reviewed the literature with the purpose to define the conditions responsible for increasing WER:

- infrequent words [Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001];
- speech rate (fast [Siegler and Stern, 1995; Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001] or very slow [Siegler and Stern, 1995; Shinozaki and Furui, 2001]);
- short words [Shinozaki and Furui, 2001].

Table 1.2: Examples of WER between the reference transcription (REF) and ASR output (HYP), with regard to substitutions (S), deletions (D), and insertions (I) from [Adda-Decker, 2007].

```
(S)  REF:  c' était le même marasme
     HYP:  c' est elle même marasme

(D)  REF:  confiance apréciable le tandem
     HYP:  confiance apréciable ** tandem

(I)  REF:  en fait **
     HYP:  en fait de
```

Misrecognized turns during human-computer dialogue systems were found with higher maximum pitch and energy [Hirschberg *et al.*, 2004]. Goldwater *et al.* also cited the study of [Adda-Decker and Lamel, 2005] which points out that gender could be a factor of errors since more errors were found in spoken regions produced by male speakers as a consequence to increasing disfluency and reduction rates. Word error rates also vary among speakers [Doddington and Schalk, 1981; Nusbaum and Pisoni, 1987].

Other factors likely to produce errors are out-of-vocabulary (OOV) words [Kawahara *et al.*, 2003], disfluencies [Adda-Decker *et al.*, 2003; Kawahara *et al.*, 2003], noise/overlapping speech, phonetically similar words (e.g. homophones) [Béchet *et al.*, 1999; Gauvain *et al.*, 2005], n-gram words [Adda-Decker *et al.*, 2011], quality of microphone [Lippmann, 1997], etc.

At the language model (LM) level, one problem responsible for increasing WER is that LM is based on texts whereas spontaneous speech shows specific verbal events such as disfluencies. The gap between the LM and the spoken data specificities may be a source of errors. At the acoustic model (AM) level, speech reduction is a source of errors. For instance, fast speech can yield more errors. Noisy data and overlapping speech represent also sources of errors. In [Adda-Decker and Lamel, 2005], the comparison of WER for broadcast news (BN) and conversational telephone speech (CTS) in English and French underlined that BN speech, that is prepared speech, has lower errors than CTS, the latter showing spontaneous speech-like phenomena, e.g. reduction, overlapping speech, etc. Finally, at the pronunciation dictionary level, if the word does not exist in the pronunciation dictionary (i.e. OOV word), the recognizer can not find the proper word.

In the study cited above, [Goldwater *et al.*, 2010] aimed at investigating what features of a reference word increase the probability of an error. Thus a measure differing WER is needed that computes the error attributable to individual words, hence the acronym IWER (individual word error rate). WER is calculated over the full utterances or corpora. The following equation 1.8 is used to compute the new measure IWER:

$$IWER(w_i) = del_i + sub_i + \alpha \cdot ins_i \tag{1.8}$$

where del_i and sub_i are the number of deletion and substitution in which w_i is deleted or substituted. And ins_i is counted if a word or words adjacent to w_i is inserted. α is fixed to guarantee the sum of ins_i over all w_i is equal to the total number of insertions. If there is an insertion, this insertion can be between two words. In this case, both two reference words count this insertion for one insertion. To prevent extra count of insertion, constant number of α is needed. Owing to the α, IWER over the whole corpus can be equal to WER.

In order to investigate which features can be problematic for the ASR system, Goldwater *et al.* [2010] compared two state-of-the-art ASR systems (SRI [Stolcke *et al.*, 2006] and Cambridge [Evermann *et al.*, 2005]) using telephone conversational speech data from the National Institute of Standards and Technology (NIST) 2003 Rich Transcription exercise (RT-03)[8]. Many features were annotated in the data: disfluency features (before/after filled pause, before/after fragment, before/after repetition, and position in repeated sequence), syntactic class (open class, closed class and discourse marker[9]), first word of turn, speaker gender, n-gram log probability (unigram, trigram), word length, number of pronunciations, number of homophones, number of neighbors, frequency-weighted homophones/neighbors, prosodic features (pitch, intensity, average speech rate in phones/second, word duration, log jitter).

Similar results have been obtained by the two systems for each feature of IWER. IWER results from the two systems have been computed using Monte Carlo permutation test[10] [Good, 2004] and standard logistic regression[11] implemented in `R` software [R Development Core Team, 2008]. The authors observed that the following features were most likely generate errors: turn-initial words, closed class words which are slightly worse error rates than open-class words, disfluencies (especially, fragments, non-final repetitions, and words preceding fragments), extreme prosodic values, speaker differences, and *doubly confusable pairs* which have similar-sounding and similar-context words. Results revealed as well that male speakers have higher error rate than female speakers. Results obtained by [Goldwater *et al.*, 2010] reinforced previous findings about the spoken events responsible of increasing WER.

1.3.1.1 Errors by ASR for French

As for the French language, [Adda-Decker *et al.*, 1999a] analyzed ASR errors in French and found that errors were often in relation with incorrect gender, number and tense agreement, and other homophone substitutions. [Adda-Decker and Lamel, 2005] found that male speakers have more WER in both English and French for prepared speech. The trend is confirmed by [Goldwater *et al.*, 2010]. As for disfluencies, ASR systems for French language, similarly to English, are also penalized by such non lexical events [Adda-Decker *et al.*, 2003].

[Gauvain *et al.*, 2005] found some differences between French and English. French language has much more ambiguous contexts than English with 20% of ambiguous spoken regions in the English journalistic texts while 75% of ambiguity was found in the French ones. Here is an

[8] http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/index.html.

[9] Open class (content word, lexical word): nouns, verbs, etc. Closed class (function word, grammatical word): prepositions, articles, etc. Discourse mark: well, okay, etc.

[10] The Monte Carlo permutation test is a standard nonparametric test that does not rely on the assumption that the data are drawn from a given probability distribution.

[11] Logistic regression is used to estimate the effects of each numeric features and to determine whether they have significant predictive value for the error rates.

example: the ambiguity of the sequence of phonemes /la/ which may be transcribed as `la`, `là`, `l'a`, `l'as`, or `las` [Adda-Decker, 2006]. There are thus five possibilities of correct transcription of the sequence, among them two are composed of two words (`l' a` and `l' as`). The choice of words can be solved thanks to the language model (LM). For the case of the sequence /la/, examples presented in Table 1.3 underline the difficulty to select between "`il`" and "`dit`". In the case of such ambiguous sequences, the (larger) context may help in the local disambiguation. However, an n-gram model normally employs the 3- or 4- grams, so the help of the context remains limited.

Table 1.3: Examples of candidate words for /la/ phoneme sequence with left & right contexts.

left context	**/la/**	**right context**		w **solutions**	
tu	--	vois	:	la	"You see her."
tu	--	vu	:	l'a	"You saw him/it."
il	--	dit	:	l'a, la	"You said it. You say it."
#	--	d'écriture	:	las	"tired of writing"

"*Doubly confusable*" (semantically and phonetically similar) pairs in French are also found in number (singular/plural) and gender (masculine/feminine) inflections. Here are some examples:

- For the masculine noun "*stylo*" "pen"
  - /stilo/: "*stylo*" (sing.), "*stylos*" (plu.)
- For the adjective "*beau*" (masc. sing.) "beautiful"
  - /bɛl/: "*bel*" (masc. sing. before vowels and mute *h*), "*belle*" (fem. sing.), "*belles*" (fem. plu.);
- For the verb "*arriver*" /aʁive/ "arrive":
  - present form /aʁiv/: "*arrive*" (1st&3rd, sing.), "*arrives*" (2nd, sing.), "*arrivent*" (3rd, plu.)
  - past form /aʁive/: "*arrivé*" (masc., sing.), "*arrivée*" (fem., sing.), "*arrivés*" (masc., plu.), "*arrivées*" (fem., plu.).

[Béchet *et al.*, 1999] found 72% of singular/plural homophonic inflection in the word dictionary developed from journalistic texts. Béchet *et al.* evaluated such homophonic words with the aim of disambiguating them via a comparison of four types of LM. In spite of non classical LM propositions such as phrase-based model and cache-based model, the result of the classical 3-gram LM and the 3-class LM on the Part-Of-Speech (POS) were better. However, the results with four model combination showed the best performance.

As explained in section 1.2.2.2, deletion and insertion of schwa due to pronunciation variation also trouble the speech recognizer. Finally, pronunciation variations in general are responsible for the erroneous transcription of spontaneous speech.

1.3.2 Errors by humans

The comparison between humans and machines on speech transcription tasks revealed that humans significantly outperformed machines in various speaking styles from read isolated words to spontaneous telephone speech and in different atmospheres (quiet and noisy) [Deshmukh *et al.*, 1996; Lippmann, 1997; Pols, 1999; Shinozaki and Furui, 2003; Shen *et al.*, 2008].

Interesting studies were made by [Lippmann, 1997] who compared the results of humans and machines in terms of word error rates using 6 different corpora (2 read isolated words (digits and alphabet letters), 2 read sentences, and 2 spontaneous telephone conversations). In the read digit task, the result showed that both machines and humans were good in transcribing proposed data (0.72% of error rates for machines and 0.105% with vocoded speech and 0.009% with wide-band speech for humans). Even though the machines demonstrated a good performance, the human performed at least 7 times better. Automatic transcription of another read alphabet letters corpus underlined that machines performed three times lower (5% for machines and 1.6% for humans). As for the read sentence task, worse results were obtained for both machines and humans in comparison with the two read isolated word studies. The result obtained from read sentence corpus using low to high quality 4 microphones speech revealed that humans did not suffer greatly from microphone quality sound (0.3–0.8% of word error rate), while machines were highly influenced (6.6%–23.9%). The word error rates for spontaneous speech transcription showed degraded results compared to read isolated words and read sentence words for both humans (4% of word error rate) and machines (43%). The difference of word error rates between machines and humans was wider for spontaneous speech than for read speech. These all results from 6 corpora comparing machines and humans from various speaking styles and conditions proved that humans outperformed the ASR system in all conditions. Lippmann [1997] suggested that "the performance gap between humans and machines can be reduced by basic research on improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech".

The difference between HSR and ASR is that HSR is focusing on fundamental understanding of human language processing, while ASR is focusing on the automatic decoding of the speech signal in order to minimize WER [Scharenborg, 2007]. In order to bridge a gap between humans speech recognition (HSR) and automatic speech recognition (ASR), some authors recently proposed methods which aim at adopting HSR findings to the ASR system architecture [Metze, 2007; Hogden *et al.*, 2007; Coy and Barker, 2007; Barker and Cooke, 2007; Moore, 2007], presented in "Special Issue" of "bridging the gap between human and automatic speech recognition" of the journal of Speech Communication, Volume 49, Number 5 in May 2007. In section 1.3.1, we proposed the error sources of state-of-the-art ASR systems. Errors have been linked to different factors such as: infrequent words, speech rate, short words, OOV, disfluency, noise/overlapping sounds, phonetically and contextually similar words, etc.

As for the human errors, psycholinguistic studies on spoken word recognition look often into response time or reaction time (RT) patterns and accuracy than error rates. It has been underlined, that RT increases for phonetically similar neighborhoods [Luce and Pisoni, 1998; Vitevitch and Luce, 1998; Marslen-Wilson and Zwitserlood, 1989]. Similar to the behavior of an ASR system, human listeners found difficult to differentiate phonetically related words. [Dahan *et al.*, 2001] highlighted the effect of word frequency: RT for high-frequency target words was longer than RT for infrequent words. Combining two factors (number of neighbors and frequency of neighbors)

into a single measure, called *frequency-weighted neighborhood density*, revealed the correlation between these two factors for RT [Luce and Pisoni, 1998; Vitevitch and Luce, 1998]. These results clarify that sounds/speech which make word recognition difficult for an ASR system is also hard for human listeners.

Vasilescu *et al.* [2009] compared the word error rate of the ASR system and the HSR system for the same 7-gram language model by doing perceptual tests involving frequent near-homophones in French and in American English. Human subjects were asked to transcribe 7-gram word chunks after listening to broadcast news (BN) corpus excerpts. Since the ASR system [Gauvain *et al.*, 2005] uses 4-gram language model (LM), it was suitable to choose 7-gram word chunks (3 words left and right with a central target word). The central words of the stimuli were concerned with ASR errors. The evaluation was made by the fourth (central) target word. The results from the perceptual tests underlined that humans produce 12% of errors for American English and 15% for French. Authors observed an increase of the amount of errors for the chunks missed by the automatic system (16% for American English and 18% for French) supporting the hypothesis of a local ambiguity due to the homophone targets. These results showed that "doubly confusable" pairs are confusable for both humans and machines. However humans outperformed about 5 to 6 times than the ASR system on the central word of 7-gram chunks, where the ASR system gave 100% word error rates.

In spite of the increasing progress of technology on the ASR system, it may remain yet something to "bridge a gap between humans speech recognition".

1.4 Conclusion

This chapter presented a general description of a standard automatic speech recognition (ASR) system after a brief introduction of the ASR history. The standard statistical ASR system consists in three components: acoustic model, n-gram language model, and pronunciation dictionary. Pronunciation can be varied according to speaking styles. Articulation is clearer for the read speech than the spontaneous speech which has more variations in pronunciation, rate of speech, syntax with disfluencies. And heterogeneous pronunciations can cause automatic errors. Thus spontaneous speech has more word error rates than read speech.

A particular focus has been put on the transcription errors for both current ASR systems and humans. The literature highlighted that humans are still higher-performance than machines. The literature underlines the effects in terms of increasing WER of "doubly confusable pairs" (phonetically and contextually similar pairs), even though humans outperform machines on such items. Bridging human and machine language processing remains an important objective in order to improve the ASR system as suggested in [Goldwater *et al.*, 2010].

Chapter 2

Prosody

For many years, researchers interested in the fields of human and automatic speech processing have tried to solve the question as to how listeners can detect word boundaries in the continuous speech stream of spoken words. Indeed, spoken words do not have clear acoustic markers that indicate the beginnings and ends of a word, such as the spaces that indicate lexical boundaries in the case of written words. The lexical segmentation problem may lead to comprehension problems when it results in two (near-)homophonic interpretations, as the following examples in French, from one the Jacques Durand's presentation in 2009), nicely illustrate:

1) /lezar/: *les arts*; *lézard* ('the arts'; 'lizard')
2) /ɔ̃sãdegut/: *on s'en dégoûte*; *on sent des gouttes* ('we are disgusted'; 'we feel drops')

How do automatic speech recognition (ASR) systems deal with lexical segmentation? What features do humans, better performing speech recognition than machines, use in order to recognize the audio stream? A review of the research in human speech recognition shows that a prominent role has been attributed to prosody, a vital component of human communication.

Hence, we will give an overview of prosody investigations from the raised question in this chapter. Firstly, we will introduce the general definition of prosody in section 2.1 and its application in speech technology. As we investigated French language, general prosody of French is also presented. Secondly, each prosodic parameter will be described in section 2.2. Next, from prosodic parameters, section 2.3 explains how speech is structured from a prosodic viewpoint. Then, the role of the prosody in human speech recognition will be outlined in section 2.4 before providing brief conclusions (cf. section 2.5).

2.1 General definition of prosody

What does mean "prosody"? According to [Crystal, 1997], prosody covers the study in suprasegmental phonetics and phonology of variations in pitch, loudness, tempo and rhythm. Sometimes it is used loosely as a synonym for "suprasegmental", but in a narrower sense it refers only to the above variables. Most of the authors indicate that the main components of the prosody are fundamental frequency (f_0), duration, and intensity. They are measurable physical terms. Prosodic information is thus conveyed by not just one parameter but pluri-parameters.

Prosody does not characterize written languages. Instead, written languages use punctuation marks[1] to structure phrases and sentences, and also employ question mark (?) and exclamation mark (!) to express paralinguistic information (nonverbal elements of communication) such as emotion, affirmation, and question or doubt, etc. An ellipsis of three dots (...) indicates that something is omitted or there is an intentional silence.

For [Delattre, 1966b], intonation is one of the prosodic phenomena which corresponds to suprasegmental characteristics. Effectively, prosodic phenomena are not individually concerned with segments like vowels and consonants, but are concerned with words and sense-groups. The prosodic phenomena include accent (final accent, insistence accent), rhythm, syllabication, and pause from a subjective viewpoint and intensity, duration, and fundamental frequency from an objective viewpoint as quantifiable via acoustic characteristics. Fundamental frequency plays a most important role in intonation perception: one can effectively distinguish a sentence as question or assertion, or recognize emotional attitudes such as surprise or joy thanks to fundamental frequency variations.

Hirst and Di Cristo [1998] explained that the term of prosody have been used interexchangeably in the literature with the term of intonation. Differences in the respective meanings of the two terms may be noticed according to different authors. Hirst and Di Cristo underlined that the term of prosody is used most broadly. The authors divided prosody in two levels: *lexical* level and *non-lexical* level. At the lexical level, lexical identity of words can be defined by tone, stress, and quantity. The non-lexical level, called also supralexical, and postlexical, includes larger range than the lexical level to express pitch patterns, declination, boundary phenomena, etc. (cf. Figure 2.1).

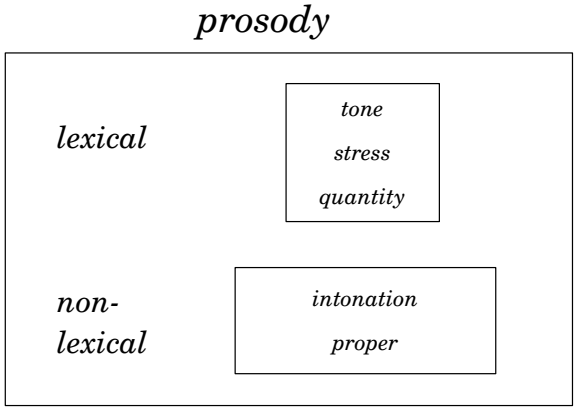


Figure 2.1: Prosodic functions from [Hirst and Di Cristo, 1998].

Lacheret-Dujour [2000] illustrates prosody as in Figure 2.2. The author divided prosody in two mechanisms: accentuation and intonation. Accentuation plays a role in composing an accentual or a stress group. Intonation is associated with intonation group. The intonation characterizing an intonation group is more remarkable than prosody associated to accentual groups. Besides, an intonation group can be followed by a pause. In lexical prosody, only accentuation is taken into

[1]Representative punctuation marks are: comma (,), period (.), colon (:), semicolon (;), quotation marks (‘’), apostrophe (’), etc.

consideration, whereas both accentuation and intonation are processed within the framework of the postlexical prosody.

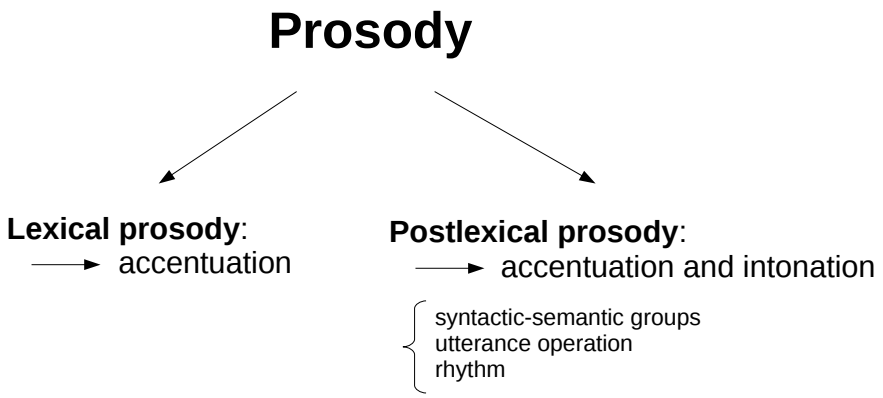


Figure 2.2: Prosodic functions from [Lacheret-Dujour, 2000].

[Vaissière and Michaud, 2006] defined prosody as consisting of accentuation, intonation and performance factors. Accentuation contains "all nonphonemic lexically distinctive properties". Intonation is often expressed by the fundamental frequency. Two levels of analysis are concerned with intonation features: syntactic and pragmatic. Syntactic intonation represents syntax in a large sense, although the author stated that syntactic intonation is not exactly linked with syntactic units each other. Pragmatic intonation is concerned with the information structure. Intonation can also correspond to attitudinal and emotional factors: speakers may express their attitudes and emotions via specific intonation patterns.

The common points of the authors above are that prosodic features are not linked to segmental level (i.e. phonemes levels), but with the lexical level, or more largely, the postlexical level. The term of accentuation is used at the lexical level and intonation at the non-lexical or postlexical level.

The voice height (pitch) can produce an intonation, a tone, and a melody as mentioned above. Languages which use tonal accent, pitch accent, or stress accent, distinguish words using prosody at the lexical level. Tone languages use pitch contour patterns in order to distinguish a lexical or grammatical meaning. Some tone languages such as Chinese, Thai, Vietnamese, etc, a tone change within a syllable. For example Mandarin Chinese have four tones: high level, mid rising, low dipping, and high falling. These tone patterns can be represented by Chao's iconic tone letters [Chao, 1930] as follows: [55, ˥ma 'mother'], [35,˧˥ma 'to be numb'], [214, ˨˩˦ma 'horse'], and [51,˥˩ma 'to curse'][2] respectively from 1 (low) to 5 (high).

While a pitch contour of tone languages changes within a syllable, Japanese uses a pitch contour within a word [Warner and Arai, 2001]. For example, a phoneme sequence of /hasi/ can have three different meanings with different pitch contours: bridge (**L**ow-**H**igh), chopsticks (H-L), and edge (flat or L-H) with around Tokyo city accent.

[2]Examples extracted from [Kratochvil, 1998, p.420]: ma, má, mă, mà

In languages as English, one word has several grammatical categories which can be distinguished by stress position. For example, the word “present”, stress put on first syllable “ˈpresent” means noun and adjective and on last syllable “preˈsent” for verb. Strictly, pronunciations for noun and verb are not the same pronunciation because unstressed vowels can become schwa-like vowels with weak syllables containing reduced vowels [Cutler, 1991, p.161–162]. Cutler and Carter [1987] revealed most of the lexical words in English have strong initial syllables and grammatical words have a weak syllable.

At the postlexical or non-lexical level, movement of f_0 helps us to recognize which kinds of information are carried. For example, at the sentence level, final rising f_0 patterns correspond to interrogative modality, whereas a falling f_0 portrays the assertive one. This tendency is observed in many languages (French, English, Japanese, Danish, etc.) [Vaissière, sous press].

Pike [1945] associated accentuation with rhythm. Accentuation is appeared in regular timed intervals. The author proposed a typological classification for languages rhythm, known as the “isochrony” theory (see also [Abercrombie, 1967]). Languages are categorized according to one of the three types of isochrony: stress-timed, syllable-timed, and mora-timed languages. A stress-timed language has the equal temporal duration between two stressed syllables. English, Arabic, Russian, etc. are considered as stress-timed languages [Pike, 1945; Abercrombie, 1967]. A stress syllable is related to higher f_0, stronger intensity, and longer duration. In such stress-timed languages, stressed syllables appear at a constant timing followed by zero or more unstressed syllables [Beckman and Edwards, 1990].

As for a syllable-timed language, the duration of each syllable is supposed to be equal. French, Spanish, Telugu, Yoruba, etc. are known as syllable-timed languages [Pike, 1945; Abercrombie, 1967].

A mora-timed language is considered as having the same duration for each mora. Japanese is one of mora-timed languages (see [Bloch, 1950; Warner and Arai, 2001]). In Japanese, V (vowel), CV (consonant + vowel) or CSV (consonant + semivowel[3] + vowel) syllable is considered as one timing unit. In addition, three other cases are also composed of mora. First, geminate obstruents[4] (transcribed as /Q/) are considered as one mora because its length is distinctive. Geminate obstruents do not occur before vowels or nasal consonants. Second, moraic nasal (transcribed as /N/) is also a part of morae. Third, a long vowel owns two morae that can distinguish the meaning. For example, the word /obasaN/ with one /a/ vowel between /b/ and /s/ means an aunt and the word /obaasaN/ with two /a/ consecutive vowel is grandmother. This is different from English which accounts one syllable even though there are short and long vowels like ‘sit’ /sɪt/ and ‘seat’ /siːt/. Nonetheless, experimental measurements do not fully support the isochronous hypothesis [Lehiste, 1977].

2.1.1 Prosody of French

In the section above general prosodic notions have been introduced. Our work focuses on French specific prosodic patterns as borne by large oral data: in the following some particular prosodic

[3] Also called semi-consonant or glide.

[4] A double same consonant blocks airflow for the Japanese case. For example /kita/ with two morae (/ki/ and /ta/) and /kitta/ becoming /kiQta/ with three morae (/ki/, /Q/, and /ta/).

features characterizing French are then introduced.

French is usually considered as a fixed stress language [Di Cristo, 1998, p.196] that is different from English with free word stress, tonal language, and pitch accent language, etc. French stress is a single rhythmic stress allocated to the final full syllable, excluding a final-schwa, of the last lexical word of a stress group. Hence, Vassière [1991] claimed that French is more a "boundary language" than a stress language. Lengthening and rise fundamental frequency are remarked at the final full syllable [Delattre, 1966a; Vaissière, 1991; Di Cristo, 1998], and many other authors. The final syllable accentuation of a stress group is called final (primary) stress which is obligatory and an optional non-final (secondary) stress can also be seen at the first syllable of a content word [Di Cristo, 1998, p.196-197].

2.1.2 Prosody for speech technology

The great interests of taking into account the prosody for speech technology is related to speech synthesis in which written texts are converted to acoustics (numeric signal). The process of synthesizing speech is as follows: written text, phonetic transcription, prosodic generation, and acoustic synthesis. For the speech synthesis, it is important to generate rules [Vaissière, 1980; Bailly, 1983; Aubergé and Bailly, 1995; Boula de Mareüil *et al.*, 2001; Mertens, 2002] in order to link syntax and rhythm and to produce a numeric voice as natural as possible.

For instance, in 1970's, Vassière [1971; 1980] investigated speech synthesis for French. The methodology adopted was based on prosodic contours for lexical or prosodic words. Prosodic contours are then associated to lexical words and generation rules specify a sequence of these contours which allow deriving intonation structure of a sentence. The contour movements were denoted as rise (R), fall (F), step (S) and lowering (L). Optional position markers can be added to contour movements: initial (i), final (f), and continuous (c). Thus the sentence "*la confédération générale a organisé des manifestations importantes* (the general confederation organized important demonstrations)" can be associated to specific prosodic contours as in Table 2.1. These sequences of contours for each prosodic word are gathered in order to make a final sentence contour.

Table 2.1: The pitch contours corresponding to the sentence "*la confédération générale a organisé des manifestations importantes*" in [Vaissière, 1980].

Prosodic words	**Pitch contours**
confédération	Ri + F
générale	Ri + S + <Rc + Lf >
organisé	Ri + S + <Rc + Lf>
manifestations	Ri + S + Lf
importantes	Ri + F + Rc

As for the ASR systems, the prosody is not exploited for speech recognition, state-of-the-art systems modeling vocal tract via MFCC instead of prosodic features [Jurafsky and Martin, 2008b]. More recently, some significant progress has been made in the area of automatic sentence segmentation by combining lexical information from a word recognizer that uses both spectral and

prosodic cues (see [Ostendorf *et al.*, 2003]). To predict the absence or presence of a sentence boundary between words, various ASR modeling approaches rely on lexical, prosodic and structural features. Lexical features typically consist of word n-grams and parts-of-speech (henceforth POS) n-grams. These features are very useful for identifying short utterances in spontaneous speech, and hold different representations according to the chosen modeling approaches (e.g., an n-gram LM in the HMM framework, or word n-tuple indicators in discriminative classifier approaches).

The role of prosodic cues has increasingly become the focus of research in the ASR domain, particularly in studies that examine sentence boundaries and disfluency locations in speech transcribed by automatic recognizers (see also [Stolcke *et al.*, 1998; Kolář *et al.*, 2010]). For instance, Vicsi and Szaszàk [2010] examined the contribution of prosody in Hungarian speech segmentation through the elaboration of a classical speech recognition system with two additional syntactic and semantic modules. At the syntactic level, the prosodic segmentation of the input speech allowed word boundary recovery and *N*-best lattice rescoring based on f_0 and energy values. These prosodic cues are known to play an important role in the fixed-stress language of Hungarian. Duration-like features were found to be less reliable in Hungarian and were therefore discarded (unlike for instance English, where duration information is more robust and stress is at least partly unpredictable, see e.g., Campbell, 1993). The rationale of the use of the stress-related features was the following. If the stressed units are carefully labeled and identified so that their boundaries coincide with actual word boundaries, then the alignment of such units should be conducive to enhance word segmentation. The *N*-best rescoring based on syntactic level word-stress unit alignment was shown to augment the number of correctly recognized words.

In light of the need for more realistic dialogue systems, Hirose, Sato, Asano, and Minematsu [2005] have developed a corpus-based method that involve generating f_0 contours from text in Japanese. Predictions of the f_0 model commands were conducted for each prosodic word (accent phrase) using binary decision trees with one tree for each model parameter. With the text input, the method generated f_0 contours through prediction of phrase commands, prosodic word boundaries, decision of accent types and accent commands. The method enabled the authors to develop a speech synthesis system for three different types of emotional speech (anger, joy, and sadness). Perceptual experiments further confirmed that the designated emotions could be successfully conveyed with the f_0 contours generated by the corpus-based method.

Prosodic cues may also be used for accent/dialect identification (cf. Arabic dialect [Rouas, 2007], French dialect [Woehrling, 2009]), language identification [Pellegrino, 2009], speaker identification [Leung *et al.*, 2008], and emotion [Liscombe, 2007; Vidrascu and Devillers, 2007].

2.2 Acoustic correlation of prosody

Analysis of prosodic features of a language is concerned with intonation (low/high of voice height), stress (weak/strong of voice strength) and rhythm (slow/rapid of voice length) of speech. These units have suprasegmental characteristics[5]. Prosody consists of mainly fundamental fre-

[5]In phonetics and phonology, a segment is a discrete unit like a phone or a phoneme which can be identified in the stream of speech. Prosody can occur over several segments, thus prosody is considered having suprasegmental characteristics.

quency (f_0), duration, and intensity at physical acoustic levels. Pause and formants can also be added to prosody's composition. The terms of physical acoustic levels for prosody differ from those of perceptual levels. In perception, the main three acoustic factors correspond to pitch, length, and loudness, respectively. While acoustic values are represented in physical values and thus considered as objective view, perceptual prosodic evaluation is subjective and this depends on individual feelings.

In [Hirst and Di Cristo, 1998, p. 6] and [Lacheret-Dujour and Beaugendre, 1999, p.233], the authors stated the dichotomy between linguistic and physical levels of analysis. We drew up a list in comparison of these terms of linguistic and physical levels (cf. Table 2.2). We note that linguistic level terms are also compatible with perception and physical level terms with acoustic terms.

Table 2.2: Comparison of terms between acoustic and perceptual levels.

Acoustics objective	**Perception** subjective	**Notes**	**units**
fundamental frequency	pitch	voice height, register (local) melody (global)	Hz,
duration	length	lengthening (local) rhythm, tempo (global)	second (s) millisecond (ms)
intensity	loudness	physical strength (amplitude)	dB, phon
formants	timbre	open/close front/back	Hz

As noted above, the prosody is defined by fundamental frequency, duration, and intensity with physical parameters and pitch, duration (length/rhythm/tempo), and loudness with perceptual terms (cf. Table 2.2). In addition pause also plays a prosodic role. Prosodic parameters and their role in characterizing speech are described in the following.

2.2.1 Fundamental frequency (f_0)/Pitch

Fundamental frequency (f_0) describes the rate at which the vocal folds vibrate at the level of the laryngeal prominence (around center of the neck) and determines voice height. f_0 corresponds to pitch in the perceptual term. f_0 is measured in Hz (hertz) and its value is calculated by vibration frequency per second. Semitone (1/2 tone) is often used as another measurement for perceptual scales, because the sensation of sound height is represented in logarithmic[6]. In [Ghio, 2007], the author gives the clear example using musical notes: the difference of 130 Hz between C3 (262 Hz) and G3 (392 Hz) is perceived as the difference of 261 Hz between C4 (523 Hz) and G4 (784 Hz). Thus the octave can be manifested in 110 Hz, 220 Hz, 440 Hz, 880 Hz from one to another octave. Semitone is the smallest musical interval between two adjacent notes like C and C♯ or D♭ used in the occidental music. One octave has 12 semitones which are equally spaced.

[6] $x = a^p$ where x is composed of the base a and the exponent p. And the logarithm is to calculate the exponent p which makes the base a the number x. If $x = a^p$, then p is the logarithm of x to base a, and thus it is written $p = log_a(x)$. So $log_{10}(100) = 2$ and $log_{10}(1000) = 3$.

The semitone calculation from f_0 can be made using a logarithmic frequency scale. Traunmüller [2005] use a standard musical octave to extract semitone as follows: $12 \times \log_2(f/127.09)$ where f is frequency in Hz. Mertens [2004] presents the conversion from Hz to semitone as: $12 \times \log_2(f/f_{ref})$. Another formula is presented by [Ghio, 2007]: $40 \times \log_{10}(f/f_0)$.

If the frequency is higher, voice height is also higher. The frequency can be changed with anatomic view as mentioned in section 1.1.1. Men generally have longer and thicker vocal folds that lead lower voice production with less frequency than women. The same phenomenon can be seen between adults and children. Children yield much higher voice than adults. This can be linked to the stringed instruments. You can imagine the cord length of the violin that is much shorter than those of contrabass which give lower sounds than the violin ones.

Each phoneme can influence the height of f_0. Adda-Decker [2007] revealed that each vowel has a different average f_0. The author investigated to compute the average f_0 of each vowel in French. The difference between the maximum vowel (/ø/) and the minimum vowel (/ə/) is about 20 Hz for the all durations and about 40 Hz for the duration more than 100 ms.

2.2.2 Intensity/Loudness

Intensity refers to the amplitude of sound waveforms to describe sound strength and correlates with loudness in perceptual term. The amplitude is determined by the vibration of sound pressure. As mentioned above in section 1.1.1 the mechanism of how humans produce voice and hear sound, sound pressure is transmitted in the air, of course, we can not neglect that sound pressure can also propagate in the water or other situations, we will explain sound pressure in the air with microscopic view. We recall that sound is propagated by air vibration which can be represented by sound waves. Air is composed of molecules. When a sound occurs in a certain direction, this sound moves air molecules around of it. This is because "when vocal folds are open, air is pushing up through the lungs, creating a region of high pressure and when the folds are closed, there is no pressure from the lungs" [Jurafsky and Martin, 2008b, p. 267]. Air molecules receive pressure and these pressured air molecules push other near molecules (compression). Then these pressured molecules moves away from each other (rarefaction) attenuating pressure. Pushed molecules push other molecules (compression). These alternate movements of compression and rarefaction repeat. These air pressure reached human's eardrums to vibrate them. The repetition of air molecular pressure can be represented as sound wave with its compression part as more than 0 and rarefaction part as less than 0.

Sound intensity is often represented in decibel (dB). In the experiments presented in following, intensity is considered along with pitch and duration as a potential cue in characterizing words boundaries.

2.2.3 Duration/Length

Duration in a physical term means a length of time (of speech) and often uses second (s) or millisecond (ms) of the time unity. Duration is correlated to rhythm (a regular repeated pattern of sounds), tempo (speed of a speech or an utterance), and timing (a repeated rhythm). Different from fundamental frequency and intensity, duration needs segmentation to measure it.

Lexical category distinction, such as lexical (content) words and grammatical (function) words, can influence different word duration. Bell Brenier, Gregory, Girand and Jurafsky [2009] revealed in their study of telephone conversation corpus in American English that word durations can be influenced in terms of word frequency, repetition, and predictability. For the grammatical words, pronunciations are shorter, after considering frequency and predictability. As for the lexical words, frequent content words have shorter durations. The content word durations are also shorter if they are repeated. Also in French it can be seen that grammatical words shorter duration, and lower f_0 and intensity [Vaissière and Michaud, 2006, p.54].

Different situations cause also different rate of speech according to speaking styles (read speech, conversation, spontaneous speech, etc.). Even within a conversation speech, the rate of speech may change between friends with a familiar situation or unknown persons with a formal speaking way. This is because of the different context such as public or private. Generally the speaking rate of read speech is slower and the utterance is better articulated than the spontaneous speech. We can say the same thing for the public speech. In the case of spontaneous and conversational speech, the rate of speech is faster or more variable.

From the viewpoint of the question raised in this work, i.e. the role of prosodic cues in identifying words and syntagms in continuous speech, the duration is a significant parameter as speech rate is associated to the prosodic articulation of different speech levels [Zellner, 1998]. The speech rate plays an unquestionable role in the realization of other prosodic parameters and is considered in our work through the analyses of different speaking styles, from prepared (that is slow speaking rate) to conversational (that is faster speaking rate and more variation).

2.2.4 Formant/Timbre

Formants characterize a resonance of the human vocal tract, especially vowels. A formant is a range of frequency (frequency band) that is particularly increased sound strength by the vocal tract. A formant is a concentration of acoustic energy around a particular frequency in the sound wave. If amplitudes are big, the sound would be a vowel. Timbre is correlated to perceptual term of formants. Timbre is related to tone quality and color which allows us to distinguish between two different sounds of the same pitch at the same amplitude.

Formants are useful to detect vowels. The first three formants represent vocalic characteristics: the first formant (F1) is correlated to open/close (vocalic) sounds. If the F1 value increases, the vocalic sound becomes more open (e.g. /a/). The second formant (F2) corresponds to an axis front/back vowel. The higher F2 value becomes the fronter vowel (e.g. /i/). The third formant (F3) is concerned in rounded/unrounded with its lip position like between /i/ and /y/ [Gendrot *et al.*, 2008]. Figure 2.3 represents the French vocalic triangles of the average of F1 and F2 from male (left) and female (right) speakers of the ESTER corpus composed of francophone broadcast news (see in section 3.1.1 for the detail) as the way of [Gendrot and Adda-Decker, 2005]. This vocalic triangle figure is demonstrated with regard to 4 groups of duration: 30-40 ms, 50-60 ms, 70-90 ms, and more than 100 ms. Shorter a vocalic duration is, in more center it locates. This figure reveals that acoustic realization changes are caused due to duration variation.

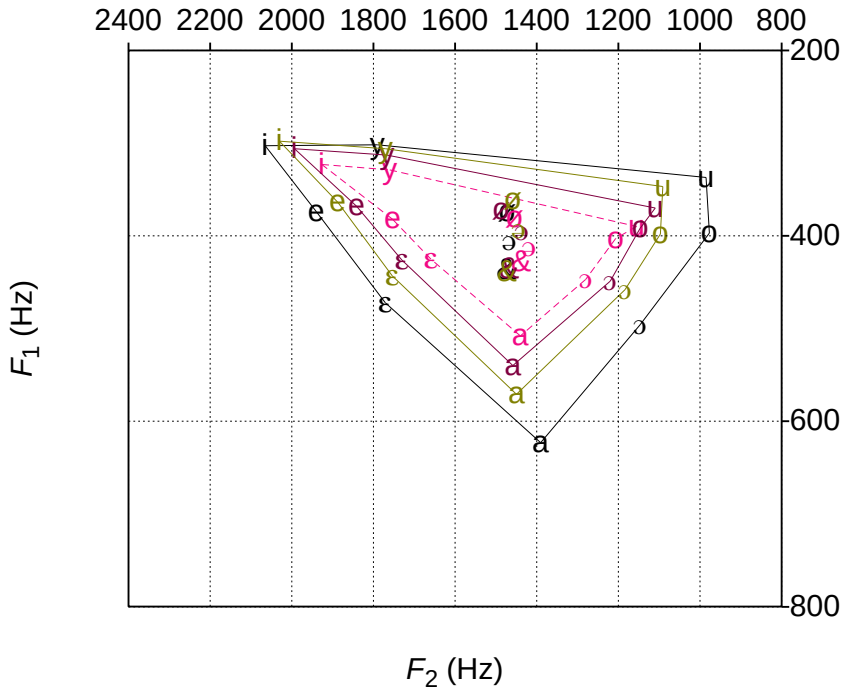


Figure 2.3: Vocal triangles in French of male speakers from the ESTER corpus according to duration. Duration ranges are: from center line 30-40 ms (pink), 50-60 ms (brown), 70-90 ms (green), and more than 100 ms (black). The '&' symbol signifies hesitation.

2.2.5 Pauses

If we say 'pause', there are two types of pause: one is silent pause and another is filled pause (filler). A silent pause plays a role of the speech structure and the speaker changes during the conversation (cf. [Candea, 2000, p.21], also see in [Campione, 2001, p.195]). Breath also is a part of silent pauses. It is quite rare that there is a breath pause within a word unless there is a physiological problem [Candea, 2000, p.21]. In the literature, the authors tend to fix the minimum duration threshold of silent at 200 ms because this 200 ms of silent pause is perceptible and countable by humans [Candea, 2000; Campione, 2001]. Duez [1982] measured silent pauses which was "any interval of the oscillographic trace where the amplitude is indistinguishable from the background noise". And she used the threshold of silent pause between 180 and 250 ms. Lacheret and Victorri [2002] employed the threshold of the minimum silent pause duration at 300 ms as intonation period boundaries.

Filled pause is a kind of disfluencies recognized as 'hesitations' like 'uh, um, er' for English, '*euh*' in French, and '*ano*, *e*, *eto*' in Japanese [Watanabe *et al.*, 2005] that the hesitations differs in between languages [Vasilescu and Adda-Decker, 2006]. In French, vocal lengthening at the end of a word is also considered as hesitation, or filled pause [Candea, 2000, p.24]. Filled pauses occur most of the time in the spontaneous speech than the read speech.

2.3 Prosodic structure

Prosodic units are hierarchically structured in various levels of phrasing from the largest level, which is the utterance level, to the smallest one, i.e. syllable or mora levels. In the framework of the Strict Layer Hypothesis proposed by [Selkirk, 1986] (also described in [Nespor and Vogel, 1986]), phonological constraints on prosodic structure are given as a single constraint requiring that a prosodic constituent of level C^i immediately dominates only constituents of the next lower level in the prosodic hierarchy, C^{i-1}. An example of the prosodic structure was hierarchically illustrated in Table 2.3:

Table 2.3: Prosodic structure proposed by [Selkirk, 1986, p.384]: Utt (utterance), IPh (intonational phrase), PPh (phonological phrase), PWd (prosodic word), Ft (foot), and Syl (syllable).

(_____)	Utt
(_____) (_____)	IPh
(_____)(_____) (_____)	PPh
(_____)(____) (_____) (_____) (____)	PWd
(_) (____)(____) (_) (_____) (____) (_) (____)	Ft
(_) (_) (_) (_) (_) (_) (_) (_) (_) (_) (_) (_)	Syl

The autosegmental-metrical (AM) studies [Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986; Ladd, 1996] use 2 tone levels (High and Low) to produce pitch accent, phonological phrase and intonational phrase boundaries combining these two tonal levels. [Silverman *et al.*, 1992] adapted this AM theory to develop the intonational transcription system called ToBI (Tone and Break Indices) using symbolic coding of intonation. This system have been originally developed for the intonation of American English, then enlarged to other languages such as German, Japanese, Korean, Greek, Catalan, Portuguese, Serbian, Mandarin, Cantonese, Spanish, etc.[7]

2.3.1 Prosodic structure of French

Here above we summarized a number of studies which reveal the prosodic structuration of a language. However, such studies lack of generalization as they are often build on English language specificities. In the following we address the question of the French distinctiveness in terms of stress system.

Delattre [1966b] investigated 10 basic frequent intonation curves corresponding to 4 voice height levels from 1 (lowest) to 4 (highest). These 10 curves are divided in 7 distinctive classes: A) minor continuation rise; B) major continuation rise; C) question; D) implication; E) finality; F) interrogation, command, exclamation; and G) parenthesis, echo. These 7 classes are categorized into 3 groups with f_0 curves (rise, falling, or static) as follows:

Rise continuative melody: A) minor continuation rise, B) major continuation rise, C) question, and D) implication;

[7] See in http://www.ling.ohio-state.edu/~tobi/.

Falling melody: E) finality, F) interrogation, command, exclamation;

Appendix melody: G) parenthesis and echo.

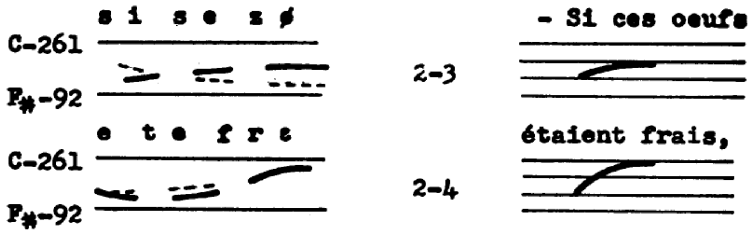


Figure 2.4: Example of major (top) and minor (bottom) continuation contours from [Delattre, 1966b].

In the work of [Delattre, 1966b], most known and cited f_0 curves are minor and major continuations. Major continuation rise (2-4 of voice height movement) is higher than minor continuation rise (2-3) (cf. Figure 2.4). And major continuation may have preceded minor continuation(s).

Rossi [1981] described intonation at morpheme levels and proposed 6 possibilities of syntactic interpretations:

Major continuative intoneme (CT): raising tonal rupture at the intonation level 4 (extreme high) with vowel lengthening of 100%;

Minor continuative intoneme (ct): raising tonal rupture at the intonation level 3 with vowel lengthening of 50%;

Non-terminal "calling" intoneme (CA) or (CT+): perceptible glissando, vowel lengthening of 100%;

Major conclusive (terminal) intoneme (CC): melody dropping at the 1 or 2 levels with glissando of intensity of -10 dB;

Minor conclusive (terminal) intoneme (cc): melody dropping of tone;

Parenthetical intoneme (PAR): static melody of level 1 during several syllables.

A model of intonation as superposition of levels appeared as particularly plausible in French [Vaissière, 1997; Vaissière and Michaud, 2006]. This model settles up to several hierarchical layers of speech units from syllable to sentence and prosodic paragraph levels: syllable and rhyme, foot, prosodic word, prosodic phrase, melodic phrase, breath group, sentence, prosodic paragraph. Between syllable and the largest unit the prosodic paragraph, the models settles the hierarchical realization of prosodic words, then prosodic syntagms and finally breath group. Vaissière underlines the particular salience of intonational phenomena in French, due to the absence of lexically distinctive stress (by opposition with English).

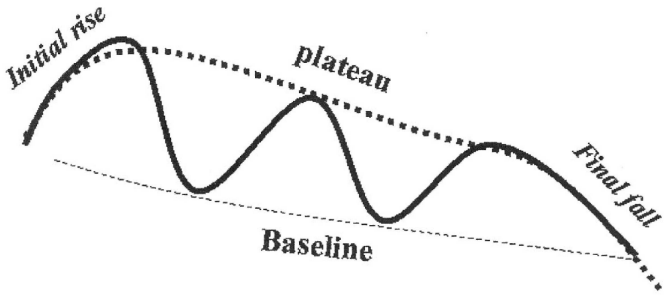


Figure 2.5: General f_0 curve of an affirmative statement from [Vaissière and Michaud, 2006].

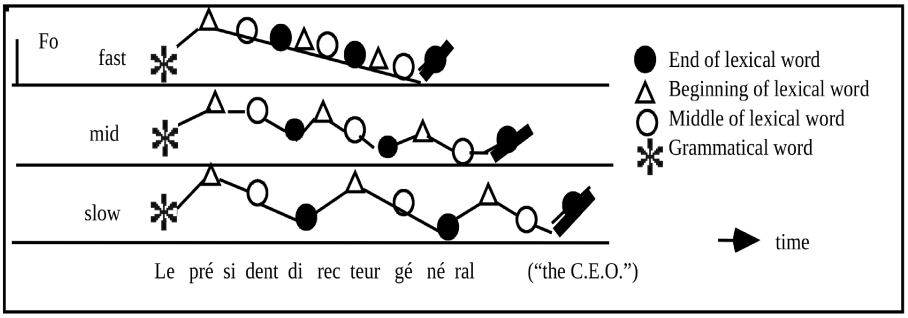


Figure 2.6: Influence of speaking rate on the division of the breath group into prosodic words extracted from [Vaissière and Michaud, 2006].

In the framework of the prosodic model of intonation as superposition, the breath group may be seen as a benchmark unit due to its physiological basis. For instance, a short sentence usually corresponds to one breath group, a longer one consists in two or more breath groups. The breath group is characterized by typical f_0, duration and intonation patterns. It may occur in a sentence or at the end of a sentence. The long prosodic syntagms are divided in prosodic phrase. In terms of pitch modulations, a breath group, whether sentence-final or not, is acoustically characterized at its beginning by a resetting of the baseline, an initial rise, generally ending at the beginning of end of the first content word, and by the return to the baseline (cf; Figure 2.5). The breath groups are divided in prosodic phrase, characterized by increasing f_0 and final lengthening. A prosodic word corresponds roughly to a content word. The alternation of lexical words with grammatical words (the latter realized less strongly with lower f_0) plays a role in French prosody: the final lengthening of the content word is a cue for boundary identification.

However these alternative f_0 rise and fall fluctuations between grammatical words and content words are influenced by speech rate that may vary f_0 fluctuations within the same phrase (see Figure 2.6).

Martin [1975; 2010] presented phonological contours using a hierarchical representation for each prosodic (accentual) group so as to indicate prosodic structure. Hierarchical structure is composed

of C4 < C3 < C2 < C1 < C0, where C0 is conclusive terminal contour and is located at the top of the prosodic hierarchy (cf. Figure 2.7). Binary acoustic and/or perceptual characteristics were added to describe more finely f_0 contours: ±rise, ±ample, ±convex, ±high, etc.

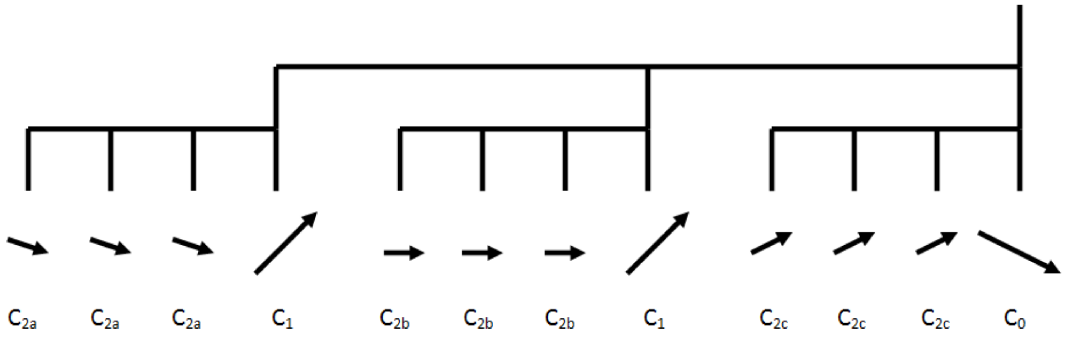


Figure 2.7: Various prosodic contours encode correspond levels in the prosodic structure extracted from [Martin, 2010].

Hirst and Di Cristo [Hirst and Di Cristo, 1984; Di Cristo, 1998] employed the term of tonal unit (TU) and intonation unit (IU). The TU is a small stress group unit with L(ow) and H(igh) tone. The IU is a higher level unit and represents a final phrase boundary tone with L or H. But temporal factors which may define rhythmic structure were not considered in these two units. Hence a unit containing one primary (final) stress with temporal factors was needed. This unit is called rhythmic unit (RU) or prosodic word (PW) which is intermediate between the TU and the IU. Here are the examples in [Di Cristo, 1998] with (|), (||), and ([...]) indicating (TU), (PW), and (IU) respectively:
[Sa SE| créTAIRE] [m'a TÉ| léphoNÉ] (His secretary phoned me.)
[Mon FILS|| et son voiSIN] [se sont DIS| puTÉS] (My son and his neighbor had an argument.)

Hirst and Di Cristo developed a language independent intonation transcription system called INTSINT (INternational Transcription System for INTonation) [Hirst and Di Cristo, 1998]. This system allows annotating prosodic events using some symbols as follows: H (higher), S (same), L (lower), U (upstep), D (downstep), T (top), M (middle), B (bottom).

Mertens [1987; 2006] defined the intonation group (*groupe intonatif*, GI) as "a sequence of one or more syllables in which the last full syllable carries final stress". The intonation group is the central element of the prosodic structure. Mertens adopted 4 pitch levels proposed by Dooren and Eyden [1982]: low (L), high (H), extra high (H+), and extra low (L-). A major interval of two intervals, for example between L to H, can be represented about 4 semitones [Mertens, 2009]. The pitch movement within internal interval is also marked as raised (/) and lowered (\) movement. These pitch level combination can introduce intonation boundaries as following:
minor boundary: /LL, \LL, LL;
major boundary: HH, H/H, LH;
terminal boundary: L-L-, LL-, HL-.
In [Mertens, 1987], Mertens also described other prosodic characteristics such as stress (pri-

mary, secondary), syllable lengthening (lengthening, longer lengthening), pause (short, long), and breath. Primary stress indicates final stress (*accent final*, AF), and secondary stress describes initial stress (*accent initial*, AI) in [Mertens, 2006]. Only final stress can make a right hand boundary of the intonation unit. Thus an intonation group is composed in the following way:

$GI = ((unstr)(AI))(unstr)\ AF\ (appendix)$

where GI corresponds to an intonation group, $unstr$ is unstressed series, AI indicates initial stress, AF is final stress, and $appendix$ (cf. [Mertens, 2006]). An appendix is added for the words which have a flat pitch contour with a low pitch level without stress at the final part of the utterance such as *en quelque sorte* ('in a manner, kind of'). Table 2.4 illustrated the distribution of tones in the maximal intonation group of IG.

Table 2.4: Distribution of tones of intonation group presented in [Mertens, 2006]

unstr		AI	unstr		AF	appendix
l	l	H	l	l	L-L-	l-...l-
h	h	L	h	h	H+H+	
					HL-	
					H/H	h...h
					/HH	
					\HH	
					HL	
					LH	
					HH	
					/LL	
					LL	
					\LL	

Jun and Fougeron [2000; 2002], claimed that stress in French is a property of a unit larger than the word, that is, the accentual phrase (AP). In a similar vein, Post [2000] proposed the phonological phrase (PP) instead of AP. Typically, the AP includes one or more content words with any preceding clitic, with a varying number of syllables. Now, the last syllable of an AP is typically lengthened and receives special prominence, in the sense that it is not only marked by its lengthening, but equally by a rise in its fundamental frequency (f_0). This late final stress is sometimes referred to as 'primary stress' or 'primary accent' (noted as LH*) and this primary stress is obligatory. An early f_0 rise (LHi) is sometimes found at the beginning of the AP and marks what is called a *secondary stress* or *secondary accent*. These two-rise (early rise, late rise) intonational patterns can be modeled as a series of high and low tones according to an autosegmental-metrical (AM) model (cf. [Beckman and Pierrehumbert, 1986]; also see [Prieto *et al.*, 2010]). This initial rise is optional. The pattern of an AP can be /LHiLH*/ whit L represents low pitch, Hi indicates secondary or initial stress, and H* describes the primary of final stress (See Figure 2.8). An intonation phrase (IP) is a highest prosodic level, and each IP is composed of one or more APs. An IP is demarcated by a phrase final boundary tone such as H%, L%.

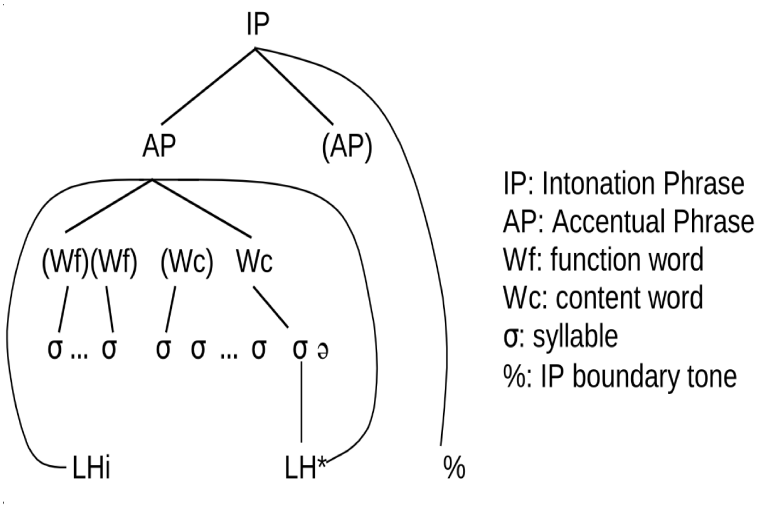


Figure 2.8: Hierarchical structure of French intonation and the affiliation of tone to syllable/structure extracted from [Jun and Fougeron, 2002].

2.4 Prosody in perception

Prosody is associated to many functions in speech. Prosodic features reflect information about the timing, amplitude and frequency spectrum of an utterance. Indeed, these are the very dimensions of sound itself, as was pointed out by Lehiste [1970] and reiterated almost three decades later by Culter, Dahan, and Donselaar [1997]. Therefore, it seems highly conceivable that human listeners exploit cues related to prosody in order to determine the location of word boundaries in continuous speech sounds (cf. [Prieto *et al.*, 2010]).

[Spinelli *et al.*, 2007] pointed out that word boundaries are marked by blank spaces in written language, while clear and obvious cues are not assigned to word begins and ends in spoken language. How do listeners detect word segmentation? Do they make use of specific parameters such as segmental and/or prosodic parameters?

As the use of prosodic characteristics differ from one language to another, the segmentation procedures that come into play during spoken word processing should also vary as a function of the listener's language experience ([Mattys *et al.*, 1999; Bagou *et al.*, 2002; Spinelli *et al.*, 2003; Kim and Cho, 2009; Warner *et al.*, 2010; Kim *et al.*, 2010]). For instance, English is known to be a stress language, as the vast majority of English content words have lexical stress on the initial syllables. In a series of behavioral experiments, Cutler and Norris [1988] demonstrated that English-speaking listeners make use of this knowledge when they need to detect words embedded in nonsense bisyllabic words (see also [Cutler, 1997; Cutler *et al.*, 1997]).

The French language does not possess contrastive word stress unlike English (cf. [Welby, 2007]). Rather, stress in French is a property of a unit larger than the word, that is, the accentual phrase (see Jun & Fougeron [2000; 2002]), the phonological phrase (cf. [Post, 2000]), the intonation group [Mertens, 1987], prosodic word [Vaissière, 1997; Di Cristo, 1998], etc.

Several studies provide evidence that French listeners are sensitive to the prosodic patterns of their

language and use such patterns in word segmentation. In a perceptual study conducted by Banel and Bacri [1994], it was found that listeners used the lengthening associated with phrase-final syllables as a potential cue to a word end. With the study of Banel and Bacri, listeners were tested with the ambiguous [ba.gage] as two words: one is *bagage* "luggage" as one word and the other is *bas gage* with two words (*bas* meaning 'low', and *gage* meaning 'pledge'). When the two syllables had a long-short pattern, listeners interpreted the item as two words, whereas they interpreted the item as one word when they had a short-long pattern. This meant that a phrase-final syllable was likely to be lengthened and a phrase boundary will not occur in the middle of a word.

In a similar vein, Rietveld [1980] studied the phonetic differences in minimal pairs in French such as *le comtat saccagé* ('the devastated country') and *le comte a saccagé* ('the count has laid waste'). Different prosodic cues such as f_0, energy levels, and durational patterns were identified that enable listeners to distinguish between these pairs. The author showed that listeners could perceptually differentiate the meanings of the minimal pairs, thereby retrieving the intended meaning of the minimal pairs.

More recently, Welby [2007] examined whether the optional early rise in French is used as a perceptual cue for lexical segmentation. This cue had already been exploited in early ASR studies in French to identify word beginnings (cf. [Vaissière, 1976; Vaissière and Le Corre, 1976]). In the first experiment, participants listened to noise-masked targets such as *ballon* ('balloon') followed by a two-syllable content words (*de mes manteaux*, 'of my coats') or by a three-syllable content word (*de mémentos*, 'of reminders') that differ in segmentation and in presence of an early rise. It was demonstrated that listeners interpreted early rise as acoustic markers of content word beginnings. In the second experiment, the alignment of the early rise was being manipulated in strings of nonword sequences such as *mélamondine*. Upon hearing these sequences, listeners were more likely to perceive two words when the early rise started at the second syllable whereas they perceived one content (non-)word when it started at the first syllable. On the basis of the empirical findings, the author concluded that French listeners use intonational cues to locate word beginnings.

In a similar vein, Spinelli, Welby and Schaegis [2007] demonstrated that listeners were well capable to perceptually discriminate between ambiguous sequences such as *la fiche* ('the file') and *l'affiche* ('the poster'). Listeners performed an off-line identification task[8] of phonemically identical sequences and managed to retrieve the correct segmentation.

In a recent follow-up study, Spinelli, Grimault, Meunier, and Welby [2010] examined the potential role of f_0 as an intonational cue in lexical access and controlled for the confounding influence of other prosodic factors by resynthesizing the f_0 of the /a/ vowel in sequences such as *la fiche* that were used in a previous study [Spinelli *et al.*, 2007]. The empirical findings of this study suggested that raising the f_0 facilitated on-line activation of vowel-initial target words.

From the viewpoint of the word segmentation issue, pitch patterns appear to provide reliable cues for lexical segmentation in French: in particular they indicate word beginnings in French [Welby, 2007]. Empirical studies have then shown that French has an optional "early rise" in fundamental frequency (f_0) at the beginning of a content word. The author tested the role of this rise in segmentation by human listeners and found evidence of the use of the parameters in word segmentation.

[8] Off-line task is to measure a result of a response from a question whears on-line task is to measure priming effect like reaction time between prime (begining of a stimulus) and response time of a target.

From the above, it thus becomes clear that human listeners can fairly easily segment spoken input on the basis of sophisticated acoustic and prosodic information, and this from an early age on. For instance, it has been shown that infants are sensitive to the rhythm patterns of their native language at birth (cf. [Ramus, 2002]).

Automatic lexical segmentation, however, has proved to be much more difficult. Automatic speech recognition (ASR) systems typically locate word boundaries in continuous speech on the basis of word and word co-occurrence information. Indeed, virtually every large vocabulary recognition system relies on the *n-gram model* for representing word sequence probabilities, in particular the trigram model. From this perspective, speech recognition can be formulated as a search problem which implies the computation of likelihoods of all possible word sequences (see [Ostendorf *et al.*, 2003]). ASR systems thus locate word boundaries by taking into account distributional language-specific properties. Indeed, specific acoustic cues, in particular prosodic ones, related to word-boundary location may be implicitly accounted for due to cross-word triphone models. The distributional cues stem in this case from the lexical level, but not the prelexical one, since ASR systems dispose of a priori knowledge about the lexicon of the language.

However, from the above-mentioned behavioral studies in human language processing, it has become clear that the contribution of prosodic processing to word-boundary location, and thus to lexical access, should operate prior to any role played by prosody in the lexical access process itself. Therefore, at least for human language processing, the question of word boundary location has long been considered as part of the prelexical processing of speech [Cutler *et al.*, 1997].

2.5 Conclusion

This chapter has reviewed the prosodic organization of speech along with the main prosodic parameters in relation with the question addressed: how do humans make use of prosodic events to segment speech. By extension, the role of such cues in speech technology has been considered as well.

Prosody is composed of pluriparametric components: three main parameters are widely considered in the literature that is fundamental frequency (f_0), intensity, and duration. Others parameters such as formants and pauses may be also linked to prosodic effects. The term of prosody is broadly used including tone, stress, accent, accentuation at the lexical level and intonation at the non-lexical or postlexical level. Prosody at the lexical level contribute to differentiate among several meanings of a same sequence of phones/phonemes via tones (Mandarin, etc.), pitch accent (Japanese and Swedish), or stress accent (English). Intonation can be useful to structure syntactic representations, and to convey pragmatic, attitudinal and emotional information.

French prosodic characteristics are often linked to the final syllable accentuation by lengthening and f_0 rise of a content word of a prosodic boundary (e.g. accentual phrase (AP) by Jun and Fougeron [2002], prosodic word by Vaissière [1997], rhythmic unit by Di Cristo [1998], intonation group by Mertens [1987], etc. The terms are different according to authors), while functional words express low f_0. With these characteristics, French is considered as fixed stress language.

Psycholinguistic experiments also revealed that the prosodic patterns play an important role as cues for word segmentation with f_0 rise and lengthening at the final syllable of a content word

[Banel and Bacri, 1994]. An optional early f_0 rise or also called secondary rise contributes to locate a word beginning [Welby, 2007].

Both corpus-based and psycholinguistic prosodic studies demonstrated that the information of low and high pitch, and of final syllable f_0 rise and lengthening help to locate word boundaries. The sequence of phonemes /lezar/ is interpreted as "*les arts* (the arts)" with low pitch at the functional word *les* and high pitch at "*arts*". The same phonemic sequence is also decoded as the word "*lézard* (lizard)", bisyllabic content word, with f_0 rise and lengthening at final syllable as the other one, but f_0 for the first syllable may not be as low as the function word. Prosodic parameters can be reliable cues to discriminate words for both humans and machines.

In the literature, most of prosodic analyses have been dedicated to the lexical or postlexical levels. However, fine-grained prosodic analyses at the micro-prosodic level (segmental phonetic level) [Di Cristo and Hirst, 1986] within n-syllabic words have been conducted as well. Hence, we dedicated a special focus to the investigation of prosodic patterns in French (see chapter 5).

Part II

Realized works

Chapter 3

Corpora and methodology

Speech is not a homogeneous phenomenon. Physical differences between male and female, children and adults prove that female voice is higher than male voice and children voice is higher than adults' one. Even from a same person, speech can be different in terms of communication contexts (reading texts, presentation, conversation, accents, etc.). The characteristics of reading texts are clear utterance, slow tempo and quite grammatical sentences, while conversation speech is opposite, especially between friends or family who know very well to each other. We can see high tempo and less careful pronunciation. Sentences are less grammatical with spontaneous speech characteristics as disfluencies (filled, e.g. "*euh*", "um, uh", and empty pauses) and discourse markers ("*donc*, *alors*, etc."; "so, then, etc."). During conversation between unknown people, one is likely to speak clearly so as to make understand a listener. And this is contrary to between close persons (hypo-/hyper systems [Lindblom, 1963]). We were interested in studying these different types of speech. In this purpose we investigated and compared acoustic and prosodic particularities of such types of speech in large prepared and spontaneous speech corpora.

Our studies are based on two different types of speech: prepared (broadcast news) and spontaneous (conversation) speech (section 3.1). These two corpora are automatically segmented in phoneme thanks to the automatic alignment system developed at LIMSI. Section 3.2 describes the procedure of acoustic and prosodic parameter extractions from the audio corpora and the alignment system.

3.1 Corpora

The prepared speech type corpus, the French acronym for "**É***valuation des* **S***ystèmes de* **T***ranscription d'***É***missions* **R***adiophoniques* (ESTER, Evaluation of Radio Broadcast Rich Transcription Systems)", is presented in Section 3.1.1. This corpus gives several French-speaking radio broadcast news and most of speech is prepared speech type by professional speakers. The speech is fluent and has less disfluencies than spontaneous speech.

For the contrast of prepared speech, we also would like to investigate spontaneous speech corpus. The "**P***honologie du* **F***rançais* **C***ontemporain* (PFC, Phonology of Contemporary French)" project has established large French speech database from different regions and countries. The PFC corpus includes several speech types like word list and text reading for read speech, and con-

versational speech between unknown and known people for spontaneous speech. The PFC project is described in Section 3.1.2.

The details of the used corpus for each study will be described in the corresponding chapter.

3.1.1 ESTER corpus

The ESTER campaign aimed at evaluating automatic broadcast news transcription systems for the French language with its three tasks: transcription, segmentation and information extraction. There were two evaluation campaigns.

First evaluation campaign, called in this thesis **ESTER1** [Gravier *et al.*, 2004b; Gravier *et al.*, 2004a; Galliano *et al.*, 2005; Galliano *et al.*, 2006] was financed by the program inter-ministerial TECHNOLANGUE and organized by the **A***ssociation* **F***rancophone de la* **C***ommunication* **P***arlée* (AFCP, French-speaking Speech Communication Association), the **D***élégation* **G***énérale de l'***A***rmement* (DGA, General Delegation for Ordnance), and the **E**valuations and **L**anguage resources **D**istribution **A**gency (ELDA). This French corpus ESTER1 consists in recordings of broadcast news shows from six different francophone (French and Moroccan) radio stations (*France Inter*, *Radio France International*, *France Info*, *Radio Télévision Marocaine*, *France Culture*, and *Radio Classique*). For the campaign 100 hours of manually transcribed and 1,677 hours of non transcribed corpus were recorded. The ESTER1 campaign was composed of two phases: from March 2003 to March 2004 and from March 2004 to March 2005.

Second campaign, called **ESTER2** [Galliano *et al.*, 2009], aimed at measuring the progress since the first campaign. The ESTER2 campaign started on January 2008 and ended on April 2009. This ESTER2 campaign was also organized by DGA, AFCP, and ELDA. For ESTER2, about 300 hours of corpus was transcribed and about 1,600 hours of corpus data was not transcribed. ESTER2 corpus consisted of almost the same radio channels from ESTER1, plus *Africa number one* (Africa1), *Radio Congo* and *TVME* (which was changed from *Radio Télévision Marocaine*). The ESTER2 corpus has a more variety of speaking styles and accents. In spite of various speech variations, the participants kept as good results as the ESTER1 campaign or even better results for the transcription task. The interest in the named entity detection task was significantly increased with its number of participants for the ESTER2 than for the ESTER1. These phenomena revealed that researchers are interested in not only automatic transcription, but also other tasks over transcription.

3.1.1.1 ESTER campaign tasks

The main tasks of ESTER1 and ESTER2 campaigns were composed of three tasks: segmentation, transcription and information extraction. For the segmentation task, three different goals were evaluated: sound event tracking (speech/music), speaker diarization, and speaker tracking. Sound event tracking is to identify parts of the document containing music or speech. Speaker diarization consists in detecting speaker turns and grouping speech uttered by the same speaker. Speaker tracking aims at detecting parts of the document that have been uttered by a given speaker known beforehand.

The transcription task is composed of producing the orthographic transcription from the waveform. Two goals were implemented: real time and unconstrained transcription. For the real time transcription evaluation, participants were asked to run a system which could process the 8 hours of the development set in a time less than or equal to 8 hours. Unconstrained transcription evaluation did not have any time constraint.

Information extraction task goals were to extract higher level information useful for indexing or document retrieval purposes. Thus a prospective named entity detection task was carried out with 8 main categories (persons, locations, organizations, socio-political groups, amounts, time, products and facilities) and more than 30 sub-categories for the ESTER1 campaign and with 7 main categories (persons, locations, organizations, human products, amounts, time and functions) and 38 sub-categories for the ESTER2 campaign.

For our studies, we used some part of the corpus, mainly transcribed corpus.

3.1.2 PFC corpus

For the other type of speech than prepared speech, we used PFC corpus [Delais-Roussarie and Durand, 2003; Durand *et al.*, 2002; Durand *et al.*, 2003; Durand *et al.*, 2005]. The PFC project is an international project directed by Jacques Durand (ERSS, University of Toulouse-Le Mirail), Bernard Laks (MoDyCo, Paris West University Nanterre La Défense) and Chantal Lyche (University of Oslo and of Tromsø). The PFC project aimed at establishing a large contemporary French database recorded in French-speaking countries or regions. The PFC site[1] describes 72 investigation points and 33 investigation points are fully collected[2]. Speech recording made by one common protocol including a 94-word list including 10 minimal pairs, text reading, directed conversation between a subject and an interviewer (formal style) and free conversation between two or more persons who are close to each other (informal style). Averages of 10 speakers are recorded in each investigation point considering the balance of gender and age.

For our studies, we used some part of PFC corpus, mainly directed interviews and free conversations for spontaneous speech style.

3.2 Methodology

To study and analyze acoustic and prosodic parameters, we made use of the LIMSI automatic alignment system to measure phones and word durations, and to get pause information. We give a brief overview of the automatic alignment system in Section 3.2.1.

Second, to extract fundamental frequency (f_0), first three formants (F1, F2, F3), and intensity from speech corpora, the PRAAT [Boersma and Weenink, 2008] software was used. Section 3.2.2 describes how we extracted these parameters.

[1] http://www.projet-pfc.net
[2] on 21st July 2010

3.2.1 Automatic speech alignment system

Our transcribed speech corpora were aligned in phones (here phones correspond to phonemes and allophones) thanks to the LIMSI's automatic speech alignment system based on the automatic speech recognition (ASR) system [Lamel and Gauvain, 2003; Gauvain and Lamel, 2003; Gauvain *et al.*, 2005] (see section 1.1.3 for the ASR system). As speech data were transcribed in words with a certain duration boundary, the alignment system forces to segment the transcribed data into phones using an acoustic model and a pronunciation dictionary. So the alignment system needs lexical constraints. Figure 3.1 illustrates the procedure of the alignment system. This system has been used for former studies [Adda-Decker and Lamel, 1999; Gendrot and Adda-Decker, 2005].

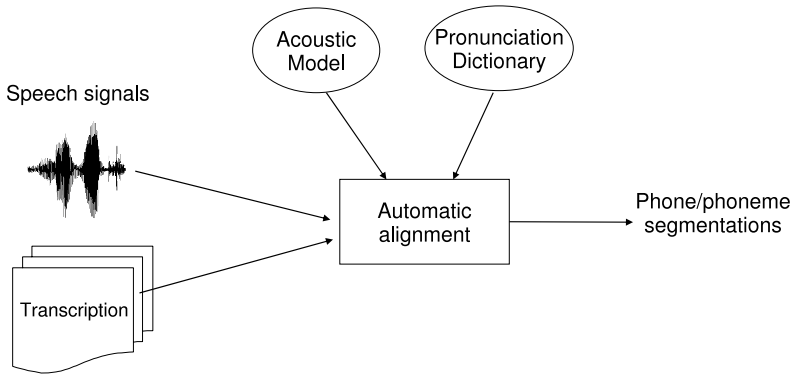


Figure 3.1: Automatic speech alignment.

At acoustic level, the **acoustic model** (AM) is used for modeling human voice or other sound like music, noise, breath etc. so as to find a certain phone corresponding with which sound wave form. Such a system is called phone-based system. Here phones correspond to phonemes and also to allophones (several articulatory realizations for a phoneme). A three-state left-to-right Hidden Markov Model (HMM) is used for our studies as for most of the ASR systems. First state is onset, then second state locates at middle, and last state is in the end part. Each state corresponds to at least one acoustic segment of 10 ms. Thus the minimum phone duration can be 30 ms with three states. Phone models are able to consider left and right context (neighbor phones), and this is called triphone. Or phone models may just investigate a single phone context (right- or left-context), or even without taking account of contexts (context-independent model). In our study, the context-independent (CI) model is applied in order to representing more phone variabilities. These variabilities are adjusted thanks to pronunciation dictionary to transform phone chains into words. The CI model may include major pronunciations (canonical pronunciations) and variants (e.g. schwa, vowel weak, etc.). As the alignment system is limited to a minimum 30 ms of phone duration, the CI model could include neighbor consonants.

The **pronunciation dictionary** is used to model word pronunciation including variants. From acoustic-level representations yielded by AM, through pronunciation model, word sequences will be outputted. For example, the word *est* ("is") has four possible representations: its canoni-

cal pronunciation /ɛ/ and its variant with liaison[3] /ɛt/, and with different height vowel /e/ and its liaison /et/. Pronunciation dictionary also admits an optional schwa realization: for example the word *seize* ("sixteen") /sɛz/ has optional schwa at the end of the word finishing by a consonant /sɛzə/ [Adda-Decker and Lamel, 1998; Adda-Decker and Lamel, 1999; Adda-Decker *et al.*, 1999b].

Table 3.1 presents 36-phones or symbols for French used in the LIMSI ASR system. 36-phone set is composed of 13 vowels including 3 nasal vowels, 17 consonants, 3 semi-vowels, and 3 special symbols of outside from International Phonetic Alphabet (IPA) phone (silence, breath noise and hesitation).

Table 3.1: Phoneme inventory of the LIMSI automatic alignment system, with LIMSI and IPA symbols.

	Vowel										Nasal vowel			Semi-vowel		
LIMSI	i	e	E	y	@	x	a	c	o	u	I	A	O	h	w	j
IPA	i	e	ɛ	y	ø	ə	a	ɔ	o	u	ɛ̃	ɑ̃	ɔ̃	ɥ	w	j

	Consonant																
LIMSI	p	b	t	d	k	g	f	v	s	z	S	Z	m	n	N	l	r
IPA	p	b	t	d	k	ɡ	f	v	s	z	ʃ	ʒ	m	n	ɲ	l	ʁ

	Silence	Breath	Hesitation
LIMSI	.	H	&

3.2.2 Extraction f_0, F1, F2, F3 and intensity

We made use of the standard settings of the PRAAT software to extract f_0, F1, F2, F3 and intensity. Measurements were carried out on a frame by frame basis of 5 milliseconds (ms). These measurements were then aligned with phone segment boundaries issued by the ASR system to calculate mean phone values etc.

Fundamental frequency (f_0) of speech corresponds to vibration of vocal cords or vocal folds measured in Hertz (Hz) which means number of oscillations per second. f_0 determines voice height and the higher rate of f_0 presents higher voice height. The f_0 information allows us to distinguish between voiced and unvoiced sounds, speaker's gender (male/female) and generation (child/adult). The first three formants represent vocalic characteristics: the first formant (F1) is correlated to open/close (vocalic) sounds. If the F1 value increases, the vocalic sound becomes more open (e.g. /a/). The second formant (F2) corresponds to a front/back axis for vowel. The higher F2 value becomes the front vowel (e.g. /i/). The third formant (F3) is generally described as rounded/unrounded with its lip position like between /i/ and /y/ [Gendrot *et al.*, 2008]. In-

[3] According to [Boula de Mareüil *et al.*, 2003], "French liaison consists in producing a normally mute consonant before a word starting with a vowel, a mute h or some glides". But this rule is not always demanded since liaison is sometimes obligatory, optional, or prohibited in the contexts.

tensity refers to the amplitude of sound waveforms to describe sound strength and correlates with loudness in a perceptual term. And the amplitude is determined by the vibration of sound pressure.

3.3 Summary and Conclusion

This chapter has presented the investigated corpora and the methodology for our studies. First of all, two different types of speech corpora, prepared (ESTER) and spontaneous (PFC) speech, were described.

The two ESTER evaluation campaigns helped to develop ASR systems in French through different tasks: segmentation, transcription, and information extraction. The ESTER campaigns also contributed creating large transcribed audio corpus which allowed us to investigate our studies. The particularities of the ESTER corpus are that speakers are almost professional using prepared text reading. Thus good articulations and fluent speech are expected. The PFC corpus aims at collecting French-speaking audio date of different speaking styles, genders, ages, and accents by one common protocol. In comparison with the prepared ESTER corpus, the conversational speech parts of the PFC corpus are chosen in our studies. As speakers of the PFC corpus are not professional like the ESTER corpus and we use parts of spontaneous speech, we can presume that speech in the investigated PFC corpus is not fluent with disfluencies and speech may also have variations with some factors such as accents, genders, and ages. Two different speaking styles (prepared and spontaneous speech) will allow us to clarify the difference between of them at several levels such as acoustic and prosodic levels.

The used transcribed speech corpora are automatically segmented in phone/phoneme owing to the forced alignment system using a pronunciation dictionary. Segmented phones/phonemes give us their durations which allow us to compute their segment values of fundamental frequency (f_0), first three formants, and intensity extracted by the PRAAT software. From these extracted acoustic values, we will study pronunciation variation in terms of phonetic/prosodic details.

Chapter 4

Classification for homophone words

Many automatic speech recognition (ASR) errors in French arise from frequent homophone or almost homophone words. A question of interest in this chapter is whether homophone words such as "*et*" ("and") and "*est*" ("to be"), for which ASR systems mainly rely on language model (LM) weights, can be discriminated by acoustic and prosodic properties, and not accounted for in the acoustic phone model. To answer this question, we investigate two complementary approaches: automatic classification and perceptual tests. For the automatic classification, two speaking types (prepared and conversational) are used to compare speaking style differences. Then we conducted a perceptual transcription test to verify if humans are able to discriminate these homophone words with n-gram constraints similar to those of n-gram LM or if they use acoustic and prosodic parameters to identify these words.

Section 4.1 gives a short overview of the transcription errors of homophone words. Automatic classification in section 4.2 presents acoustic analyses concerning prosody, prosodic attribute selection to discriminate the homophone words, and automatic classification results. Description, method and results of perceptual transcription tests are presented in section 4.3. Finally, the proposed study also contributes to describe and compare factors of automatic and perceptual confusability in section 4.4 as conclusion.

4.1 Automatic transcription errors

Automatic speech recognition (ASR) errors often arise from: Out-Of-Vocabulary (OOV) words, and (near) homophones. In the case of OOV, words are unknown by the ASR system such as proper names, or rarely used verb tenses and subjunctive forms. Thus errors happen in the ASR system. Homophones or multi-word homophones, i.e. phonemically the same but different words, induce acoustic confusability. So higher level information (e.g. neighbor word context like n-gram LM) is needed to solve acoustic confusability. However higher level information is not sometimes sufficient with a limited context of 3- or 4-gram LM because contexts can also be ambiguous. These doubly confusable words [Goldwater *et al.*, 2010] can lead more errors. Or phonetic detail [Hawkins and Local, 2007] may help to distinguish these confusable (near) homophone words. We adopt here a loose definition of near-homophone words as proposed in [Cutler, 2005] for pseudo-homophones and adopted by [Vasilescu *et al.*, 2011]. Vasilescu *et al.* conducted perceptual

experiments on near-homophone transcription by humans and machines in line with our own experiments: Cutler defines pseudo-homophony as the inability to distinguish minimal pairs in L2 language which sound the same in L1 language of the speaker, e.g. wright/light. The definition is extended here to such lexical items which may "sound identically" for an ASR system as they differ in no more than two phonemes. Such acoustic proximity makes them near-homophones.

At the Evaluation of Radio Broadcast Rich Transcription Systems (ESTER[1]) campaign, the LIMSI automatic transcription system obtained about 11% of WER (word error rate) [Galliano *et al.*, 2005; Galliano *et al.*, 2006]. The French language is particularly challenging for automatic transcription, because it admits a large number of homophones, especially different verb forms: e.g. the verb *aller* "to go" (infinitive form), *allé* (past participle for masculine singular), *allées* (past participle for feminine plural)... These words' pronunciations are the same /ale/. A large number of errors were also led by the grammatical words, which are the most frequent and are often monosyllabic as *et*, *est*, *à*, *a*, *un*, *que*, *qui*, *il*, *y*, etc. [Adda-Decker, 2006; Huet *et al.*, 2010]. Such words are often less carefully pronounced (i.e. hypo-articulated [Lindblom, 1990]). Whereas the overall WER of the LIMSI system [Gauvain *et al.*, 2005] in ESTER1 evaluation is below 12%, error rates from the 20 most frequent words contribute to more than one fourth of these transcription errors (cf. Table 4.1).

Table 4.1: List of 20 most frequent lexical forms ranked by their occurrences (left) and by their intra-class error rate (right). The error rate is composed of substitutions, deletions and insertions. The numbers between parentheses do not count their insertions [Adda-Decker, 2006].

form	*#occ*	*rank*	*form*	*%err (-%ins)*	*rank*
de	5355	1	et	25.4 (17.7)	4
la	2684	2	est	20.0 (17.1)	8
le	3011	3	a	19.5 (10.3)	14
et	1927	4	il	18.8 (16.2)	15
à	1887	5	à	15.6 (10.2)	5
l'	1840	6	un	13.1 (9.6)	11
les	1800	7	que	9.8 (7.6)	16
est	1367	8	qui	9.6 (7.0)	19
des	1378	9	en	9.6 (7.3)	10
en	1315	10	l'	9.5 (8.3)	6
un	1311	11	les	9.0 (8.3)	7
d'	1116	12	le	8.7 (6.2)	3
du	1101	13	des	8.5 (7.4)	9
a	1815	14	d'	7.8 (6.5)	12
il	916	15	de	6.7 (3.9)	1
que	913	16	une	5.8 (4.3)	18
pour	882	17	dans	5.0 (4.6)	20
une	790	18	pour	4.4 (2.2)	17
qui	797	19	du	4.3 (3.6)	13
dans	724	20	la	3.4 (2.4)	2

[1]Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques

In this chapter, we focus on two frequent monophone (near) homophones in pairs of *et* (conjunction)/*est*(verb *être*, "to be") for both automatic homophone classification in Section 4.2 and perceptual transcription test in Section 4.3, and another two frequent homophones in pairs of *à* (preposition)/*a* (verb *avoir* "to have") for automatic homophone classification. These homophone pairs are among the most frequent words in French, and are also often confused during the automatic transcription. Among the frequent words, *et* and *est* are the most error-prone items: 25.4% of *et* "and" and 20% of *est* "is" (verb "to be") occurrences, and 19.5% of *a* "has" (verb "to have") and 15.6% of *à* "in, to" were misrecognized in ESTER1 [Adda-Decker, 2006]. We notice that the canonical pronunciation of *est* corresponds to a mid-open vowel [ɛ], but in fluent speech its actual realization tends to become a closed [e], which then becomes homophone with the pronunciation of *et*. Concerning the word *est*, the pronunciation dictionary contains several variants with or without liaison: [ɛ], [e], [ɛt], and [et]. However, we only focused on realizations [ɛ] and [e] for our study, learning aside realizations with liaison constraints.

We notice that these two verbs *est* and *a* may be considered as a class of content words. However the special case of *est* and *a* entails these words as function words (auxiliary verbs) than as content words (verbs). Here are the examples:

For the verb "*est*",
Il est à Paris. "He is in Paris." (Full verb)
Il est arrivé à Paris. "He arrived in Paris." (Auxiliary verb)

For the verb "*a*",
Il a un chapeau. "He has a hat." (Full verb)
Il a obtenu un prix. "He got a prize." (Auxiliary verb)

These two verbs can locate in an internal position of a prosodic word/phrase while conjunction and preposition words can be an initial position of a prosodic word. The examples are presented below with the prosodic word boundaries "|". According to [Vaissière and Michaud, 2006], prosodic word boundaries occur in "final lengthening at the end of the first word, a strengthening of the beginning of the following word, or an f_0 fluctuation aligned with the edge of one of the words". One prosodic word contains about 3 or 4 syllables in the careful speech compared to 7 or 8 syllables in the fast speaking rate [Vaissière, 1971].

(Il) **a** *(faim)* | **et** *(elle a soif).* "He is hungry and she is thirsty."
(Il) **est** *(parti)* | **à** *(Paris).* "He went to Paris."

These homophone pairs locate in syntactically different positions in a phrase, so it will be interesting to find acoustic and prosodic detail that can differentiate these (near) homophones. As humans perform much better word recognition, it will also be interesting to compare the performance of homophone recognition between machines and humans with the same condition as the ASR system. The next section will present acoustic and prosodic analyses of the investigated homophones.

4.2 Automatic classification

In this section, we focus on the automatic classification of the two frequent homophone pairs *et* (conjunction)/*est* (verb *être*, "to be") and *à* (preposition)/*a* (verb *avoir* "to have") and the two different speech types (prepared/conversational) to explore differences in acoustic-prosodic information to discriminate the homophone words without considering classical acoustic parameters such as cepstral parameter nor acoustic HMM models that are used by ASR systems. Automatic classification of ambiguous homophone words is based on selection of acoustic and prosodic parameters.

We hypothesized that prosodic information would help to contribute discriminating certain homophone types, especially if they have different syntactic (hetero-syntactic) classes[2].

The questions that we are interested in are the following. First, although theoretically homophone, is there some fine phonetic/prosodic detail to discriminate between our homophone pairs? Second, what impact of speaking style? We hypothesized that the *à*/*a* pair (homophone) more difficult than the *est*/*et* pair (almost homophone). Spontaneous speaking style includes more "prosodic information" than prepared speaking style which is beneficial to discrimination on the basis of prosodic parameters.

4.2.1 Corpora for automatic classification

Table 4.2: Homophone word occurrences.

	ESTER (prepared, 66 hours)		PFC (spontaneous, 11 hours)	
word	#occ.	phone	#occ.	phone
à	20.4k	/a/	3.6k	/a/
a	11.3k	/a/	3.4k	/a/
et	19.1k	/e/	5.0k	/e/
est	14.5k	[ɛ]5.0k, [e]9.5k	6.2k	[ɛ]1.9k, [e]4.3k

As described earlier in chapter 3 (cf. section 3.1), we made use of two different corpora. Homophone words *et*/*est* and *à*/*a* have been extracted (cf. Table 4.2) from about 65.8 hours of speech coming from the ESTER1 corpus (section 3.1.1) and from 11 hours of spontaneous speech from the PFC corpus (section 3.1.2, see Table 4.3). Table 4.3 lists the number of speakers and their speaking time excluding silence in terms of investigated points for the PFC corpus. The speakers are from 11 investigated places in France (all except Nyon) and in Suisse (Nyon). The number of male speakers and female speakers are quite balanced: 94 male speakers and 90 female speakers. Total speaking time is 6 hours for males and 5.3 hours for females. For the ESTER1 corpus that we used for this study, the speaker gender and identity was not included at the time of our measurements. Approximately number of speakers are two third of male speakers and one third of female speakers [Galliano *et al.*, 2006].

[2]Here we mean grammatically different classes, e.g. noun, article, verb, etc.

Table 4.3: PFC corpus speaker description

Investigated points	#Guided conversation speakers (#M + #F)	#Free conversation speakers (#M + #F)
Aveyron-Paris*	11 (5 + 6)	8 (4 + 4)
Biarritz	9 (5 + 4)	5 (3 + 2)
Brunoy	10 (6 + 4)	10 (6 + 4)
Dijon	7 (2 + 5)	6 (2 + 4)
Douzens	10 (5 + 5)	6 (3 + 3)
Lacaune	12 (5 + 7)	12 (5 + 7)
Lyon-Villeurbanne	10 (5 + 5)	9 (5 + 4)
Nyon	11 (7 + 4)	5 (3 + 2)
Roanne	7 (4 + 3)	8 (4 + 4)
Rodez	7 (3 + 4)	6 (3 + 3)
Vendee	7 (4 + 3)	7 (4 + 3)
Total	101 (51 + 50)	82 (43 + 40)
Total hours	6.9h (3.6h + 3.3h)	4.4h (2.4h + 2.0h)

* Aveyronian speakers living in Paris since many years.

4.2.2 Measurements of acoustic parameters

Differences between grammatical and lexical words may be more fine-grained differences between part-of-speech (POS) classes and different positions within prosodic words/phrases. Duration can be associated to speaking style or a word belonging to a lexical or functional class [Adda-Decker, 2006]. Conversational/spontaneous speech tends to have a less stable pace including a larger proportion of fast and slow segment [Adda-Decker and Snoeren, 2011]. Furthermore, functional words tend to be more quickly or less carefully uttered than content words.

Pause contexts may be a cue to differentiate between conjunction/preposition and verbs. Since verbs locate within a prosodic word while two grammatical words are at the beginning of a prosodic word, two grammatical words are more potential to be preceded by pause. f_0 is low for grammatical words and high for lexical words [Vaissière, 1991].

These parameters will be used for the attribute definition. Selected parameters concern duration, fundamental frequency (f_0), the first three formants (F1, F2, and F3), intensity, and surrounding context information (preceding/following pauses of the target word).

Acoustic and prosodic parameters have been defined and automatically extracted thanks to the LIMSI automatic speech alignment system [Gauvain *et al.*, 2005] and to the PRAAT software [Boersma and Weenink, 2008]. We made use of PRAAT software to extract f_0, F1, F2, F3 and intensity, and of the LIMSI automatic speech alignment system to extract duration and pauses.

For each aligned phone segment, corresponding to one of the four target words and their context, f_0, F1, F2, F3, and intensity measurement have carried out every 5 ms. As minimum duration of

a segment is 30 ms, a phonemic segment includes at least six points of measurement. For each phone segment, mean values have been computed for the parameters f_0 and the first three formants (F1, F2 and F3) over all voiced frames of the segment.

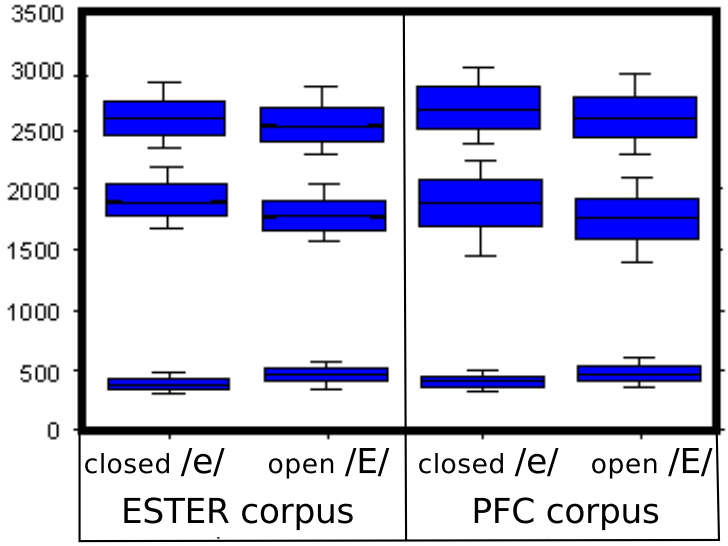


Figure 4.1: First three formant dispersion in box plot for /e/ and /ɛ/ phonemes in the ESTER corpus (left) and the PFC interview corpus (right). First formant is displayed at the bottom positions. Second formant is presented in middle and third one is the top. Y axis displays formant values in Hz.

For the word *est*, the canonic pronunciation [ɛ] (opened /E/) and its variant [e] (closed /E/) are including in the pronunciation dictionary of the ASR system. These two phonemes are perceptibly as well as acoustically different. General differences between phonemes are concerned with target words in analyzed corpora. As explained above, we can observe differences between [ɛ] and [e]. Firstly, at the F1 (bottom), higher F1 values are achieved for [ɛ] because of its more opened articulation. Higher F2 values are shown in the more front vowel [e]. Also higher F3 values are demonstrated in the vowel [e].

4.2.3 Considered parameters

Prior to the automatic classification, we need to define and measure the acoustic and prosodic parameters potentially able to differentiate the homophone word pairs. Beyond f_0 and formant measurements, duration, voicing characteristics and pauses before and after the homophone target words have been considered.

4.2.3.1 Duration

Phone duration measurements are given by automatic alignment boundaries. We did not carry out manual correction of these boundaries. Figures 4.2 (*et*/*est*) and 4.3 (*à*/*a*) represent homophone

pair duration distributions for the two selected speaking styles: prepared speech ESTER (top) and spontaneous speech PFC (bottom). Durations range from 30 to 200 ms. To facilitate the figure comparisons, each represented line sums up 100% of occurrences.

et/*est* pair

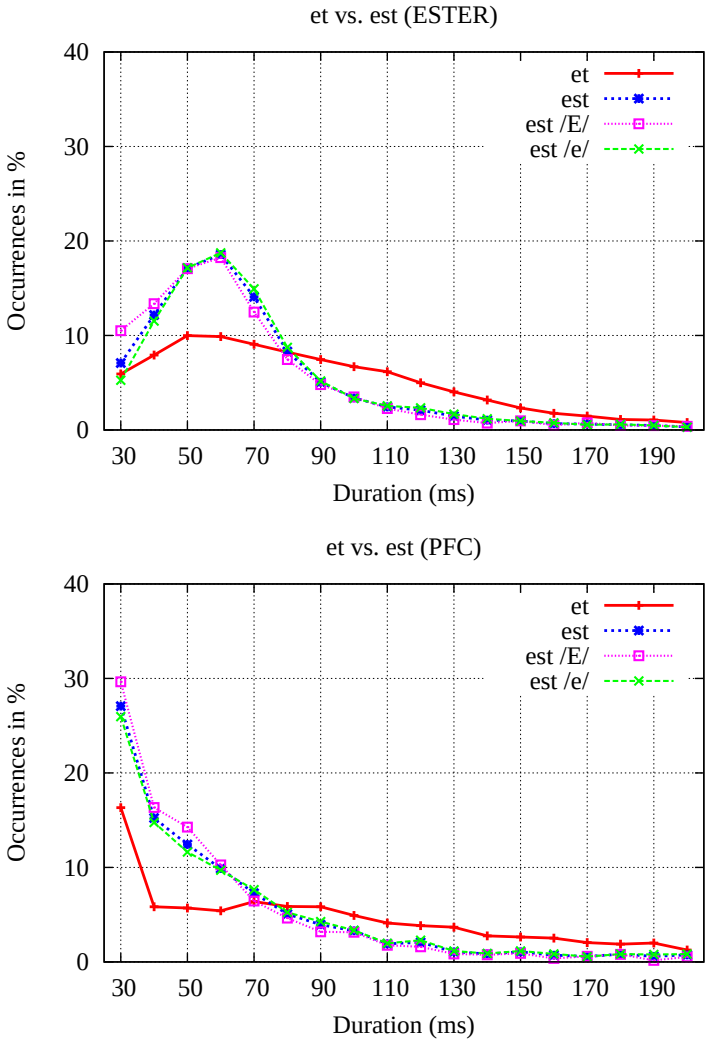


Figure 4.2: Homophone duration distributions of *et* (red $+$) and *est* (blue $*$, /ɛ/ pink $\square$, /e/ green $\times$). **Top**: ESTER corpus. **Bottom**: PFC corpus.

Duration's distribution comparison for the homophone words ***et*** and ***est*** (Figure 4.2) shows differences between the two target words within each pair. For the word *est*, three curves are presented: first curve (in blue, asterisk $*$) pools the two pronunciations [ɛ] (canonical pronunciation) and [e] (variant, but majority). Second curve (in pink, square $\square$) illustrates aligned [ɛ] realizations and third curve (in green, times $\times$) represents phone [e].

On average for the two corpora, the conjunction *et* lasts longer than the verb *est*. Especially, after 80 ms, the percentage of *et* is more important than the one of its homophone. The three curves of *est* remain similar for each of the two corpora.

Now let us describe the features for each corpus. For the ESTER corpus (top), the conjunction *et* in red line has a relatively flat distribution, including in particular more segments with durations above 80 ms, whereas *est* in blue line has an almost bell-shaped distribution centered on 60 ms. The **PFC** corpus figure (bottom) also displays a flat distribution excepting 30 ms for the conjunction *et* in red line. The word durations are shorter in the PFC corpus (max. 30 ms, bottom of Figures 4.2) than in the ESTER corpus (max. 60–70 ms).

à/*a* pair

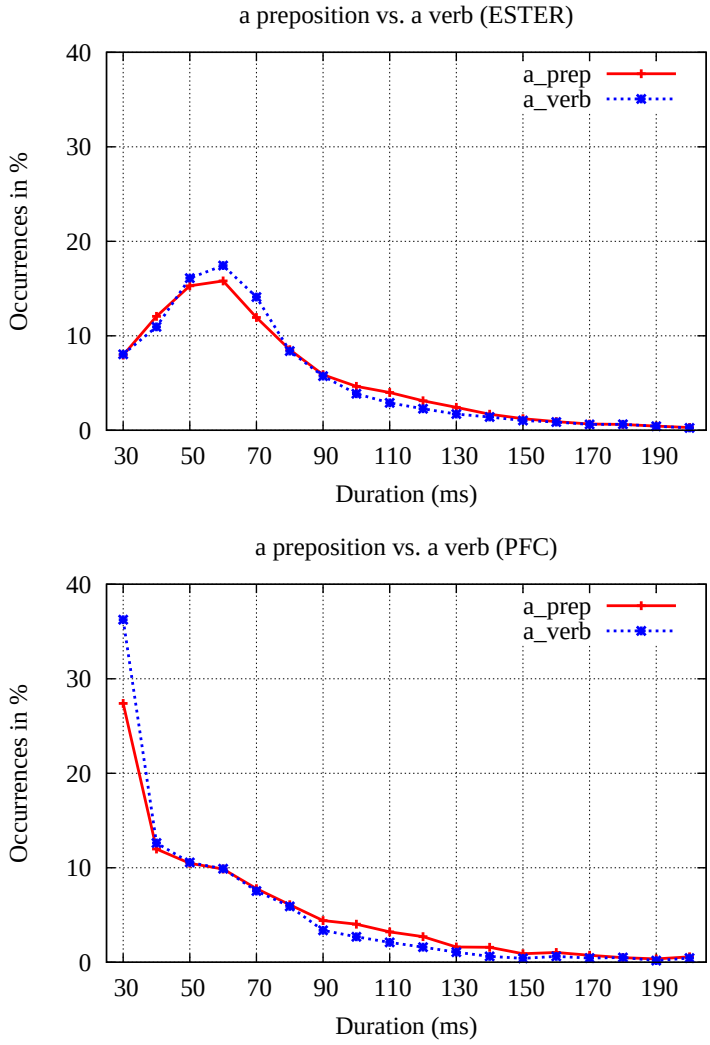


Figure 4.3: Homophone duration distributions of *à* (red $+$) vs. *a* (blue $*$). **Top**: ESTER corpus. **Bottom**: PFC corpus.

Figure 4.3 presents duration distribution for the pair ***à*** (preposition)***/a*** (verb). The word *à* is illustrated by red line with + point and *a* by blue dotted line with ∗ point. In comparison with *et*/*est* pair (Figure 4.2), we can not observe important differences between the two words of the *à*/*a* pair. However we can note in the PFC corpus (bottom) that about 8% of difference concern the minimum duration of 30 ms.

A significant difference for the two pairs of homophone words and the two corpora is that grammatical or functional word duration, e.g. conjunction and preposition, is longer than (auxiliary) verbs *est* and *a*. This information may possibly contribute to differentiate such phonemically ambiguous words.

4.2.3.2 Left-right pause co-occurrences

Pauses play an important role in the process of automatic prosodic information extraction, especially in spontaneous speech [Lacheret-Dujour and Beaugendre, 1999, p.220]. We aimed at evaluating the relationship between the pause and the investigated homophone words. We hypothesize that if *et*/*à* locate at an initial position in a prosodic word/phrase, then pause may co-occurs more frequently with *et*/*à* than the (auxiliary) verbs *est*/*a*. In the study of [Candea, 2000, p.22], she defined (silent) pause as significant interruption of more than 200 ms. This pause duration criterion is based on the former studies such as Duez [1991] (between 180 and 250 ms with an average of 200 ms), and Lacheret and Victorri [2002] (300 ms as threshold of pause). As for hesitation (filled pause) in French, the durations of hesitation generally reach between 150 and 500 ms [Candea, 2000, p.26]. But in our studies, we defined the "pause" class in larger ranges: silences, breaths and filled pauses, i.e. hesitations which are automatically aligned by LIMSI system. And we did not take into consideration minimum duration limits for pauses in our studies. Since there are short phone/phoneme segment durations which can be aligned by the alignment system, we hypothesized that there may also be some effects for short pause durations. In this case, the term "pause" is likely to be a kind of "rests" (interval of silence) in musical terms including very short rests. We thus examined their left-right co-occurrences with the target words: *et*/*est* pair and *à*/*a* pair.

Table 4.4: Left and right pause (silence, breath, hesitation) occurrences (in %) of the target homophone words.

Words	*et* (conjunction)		*est* (verb *être* "be")		*à* (preposition)		*a* (verb *avoir* "have")	
Corpus	ESTER	PFC	ESTER	PFC	ESTER	PFC	ESTER	PFC
Left pause	**49**%	**65**%	9%	8%	**23**%	**23**%	11%	7%
Right pause	7%	17%	5%	9%	3%	10%	6%	11%

Table 4.4 lists the percentage of occurrences of left and right pauses of two homophone pairs and two corpora. The main difference between the conjunction (***et***) and the verb (***est***) concerns the amount of pause occurrences, in particular left pauses (49% for the ESTER corpus and 65% for the PFC corpus). These results suggest that the verb *est* is less frequently preceded by a pause than the conjunction *et*. And this pause phenomenon is applied to two different corpus styles.

This suggests that speakers generally introduce a caesura more often before the conjunction than before the verb. This tends to confirm the initial position in a prosodic word/phrase.

Concerning the *à*/*a* pair, comparable differences are observed: pauses are more frequent before the functional word *à* than before the verb *a*, but differences are less important than the *et*/*est* pair.

The principal difference between functional words (*et* and *à*) and lexical words (*est* and *a*) concerns pause occurrence particularly left pause and lightly right pause of the target word. This is because *et*/*à* may be at the initial position of a prosodic word/phrase while *est*/*a* words locate in an internal prosodic word/phrase. Thus the verbs (*est* and *a*) are rarely preceded by pause, contrary to the functional words (*et*/*à*).

4.2.3.3 Fundamental frequency (f_0)

The two homophone word pairs correspond to different parts of speech (POS): conjunction/preposition vs. (auxiliary) verb. This distinction may entail differences in the prosodic realization of words, e.g. the duration of the words and the fundamental frequency (f_0). The words can be distinguished according to the grammatical category, e.g. functional/grammatical words (determiner, pronoun, preposition, auxiliary verb, complement, conjunction, etc.) or lexical/content words (noun, verb, adjective, adverb, etc.) in English [Selkirk, 1996]. One may hypothesize that a verb inside a prosodic word/phrase is differently realized in terms of average f_0 from a conjunction or preposition occurring at the beginning of a prosodic word/phrase and serving in isolating syntactic blocs. Furthermore, the voicing may vary according to the position of the lexical item within a prosodic word/phrase, e.g. the voicing may be partial at the beginning of the prosodic word/phrase in particular when the prosodic word/phrase is preceded by caesuras (break) or pauses.

The voicing ratio is computed as described in Equation 4.1 and corresponds to the percentage of non null f_0 values. For each aligned phone segment, corresponding to one of the four target words, f_0 measurement have carried out every 5 ms. As minimum duration of a segment is 30 ms, a phone segment includes at least six points of measurement. For each phone segment, a voicing ratio was computed as the ratio between the number of voiced frames (f_0>0 Hz) and the total number of frames (Equation 4.1) to minimize the f_0 and formant measurement errors. It corresponds to a simple filtering.

$$Pv = \frac{\#voiced\ frames}{\#all\ frames\ of\ a\ phone\ segment} \tag{4.1}$$

To examine the extracted voicing ratio measures, three classes have been defined:

1. Devoicing: % of voicing $< 20\%$;
2. Partial voicing: % of voicing between $20 \leq 80\%$;
3. Voicing: % of voicing $\geq 80\%$.

The voicing assimilation of consonant (C1#C2) study with broadcast news speech in [Hallé and Adda-Decker, 2007] showed that if both two consecutive consonants are theoretically voiced, the

consonants C1 and C2 could be voiced more than 70% regardless of phone durations (more than 60 ms). For the duration category of 60–120 ms, voicing ratios of two consonants were achieved more than 80%. As we treated vowels, which are considered voiced sounds, we hypothesize that most of occurrences will belong to the "voicing" class. As vowels are theoretically voiced, the "devoicing" class is expected to be close to 0. Phone segments in this class are indicative of either non standard phones with, for example, vowel elision, alignment problems, or noisy signal. In the "partial voicing" class, vowels in prosodic word at the initial position may be partially voiced due to "glottal stops", voice onset time delay, etc.

Voicing ratio: *et*/*est* pair

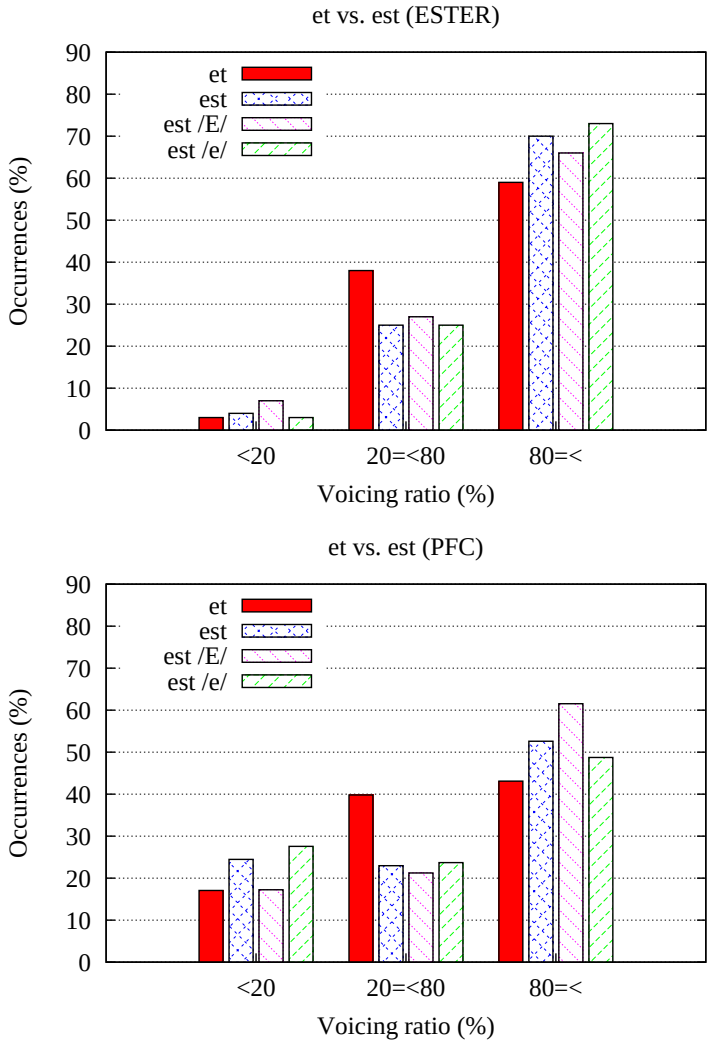


Figure 4.4: Bar charts of homophone pair *et*/*est* occurrence distribution according to the voicing ratio. *et* (red) and *est* (blue, /ɛ/ pink, /e/ green). **Top:** ESTER. **Bottom:** PFC.

For each voicing ratio class, results are given firstly for ***et*** (solid red bar) and ***est*** (blue cross-hatched bar) pair in Figure 4.4. For the *est*, two bars are added to separate [ɛ] realizations (bar with slanted stripe pink lines) from [e] realizations (bar with slanted stripe green lines) pronunciation. The results for the ESTER corpus are presented at the top of Figure 4.4 and those of the PFC corpus are below. To produce comparable results for the different conditions, absolute counts are transformed in relative rates which sum up to 100% to each condition.

As expected, "**devoicing** ($< 20\%$)" class contains a small amount of data. The ratio is very low for the ESTER corpus (top) and more important for PFC (bottom) in which the ratio of the word *est* is a little higher. In the "**partial voicing** ($20 \leq 80\%$)" class *et* is more frequent than *est* (10–15% more than the verb *est*). Conversely, *est* is more frequent in the "**voicing** ($\geq 80\%$)" category. The verb *est* ratio is higher of about 10% than that of the conjunction *et*. We can observe that the verb *est* is globally more often voiced than the conjunction.

If we compare the two speaking styles, we can observe that the general voicing ratio is lower for the PFC corpus. One may hypothesize that speaking style plays a role in obtaining such differences, as we observed the shorter durations for spontaneous speech (cf. bottom Figures 4.2 and 4.3). And we can hypothesize that spontaneous speech can be characterized by hypo-articulation [Lindblom, 1990] with vowel elision, that can lead to low voicing ratio due to unclear pronunciation.

Voicing ratio: *à*/*a* pair

Figure 4.5 illustrates the voicing ratio of the pair ***à*** (the preposition "to") and ***a*** (the (auxiliary) verb *avoir* "to have") for the ESTER corpus (top) and the PFC corpus (bottom). The results of the word *à* are described in red full bars and those of the word *a* are in blue cross-hatched bars.

Similar to the results of the conjunction *et*, the preposition *à* is more prone to belong to the "**partial voicing**" class. Reciprocally, the verb *a* is more represented in the "**voicing**" class, even though this tendency is less strong than for the *et*/*est* pair. It can be interpreted that in the two different speaking type corpora and for the two homophone pairs, the verbs locating within a prosodic word are more frequently voiced than the preposition or the conjunction situating at prosodic word boundaries. The voicing ratio in the "voicing" class is less important for the PFC corpus than for the ESTER corpus. This can again be linked to the speaking style. Indeed, very short durations are more frequently occurring in spontaneous speech.

Correlation between voicing ratio and duration

From these voicing ratio results, we wonder if voicing ratio can be concerned with phone durations like short duration can be more seen in devoicing class and long duration for the voicing class, inversely. Thus we investigate the correlation between voicing ratio and duration.

Table 4.5 presents the short duration impact on voicing ratio according to each proportion of voicing ratio classes. As we expected, there are more short duration (30–40 ms) rates in the devoicing class. The verbs have higher rate of short duration than the conjunction and preposition words. The duration distribution is very flat for the partial voicing class, especially for the conjunction

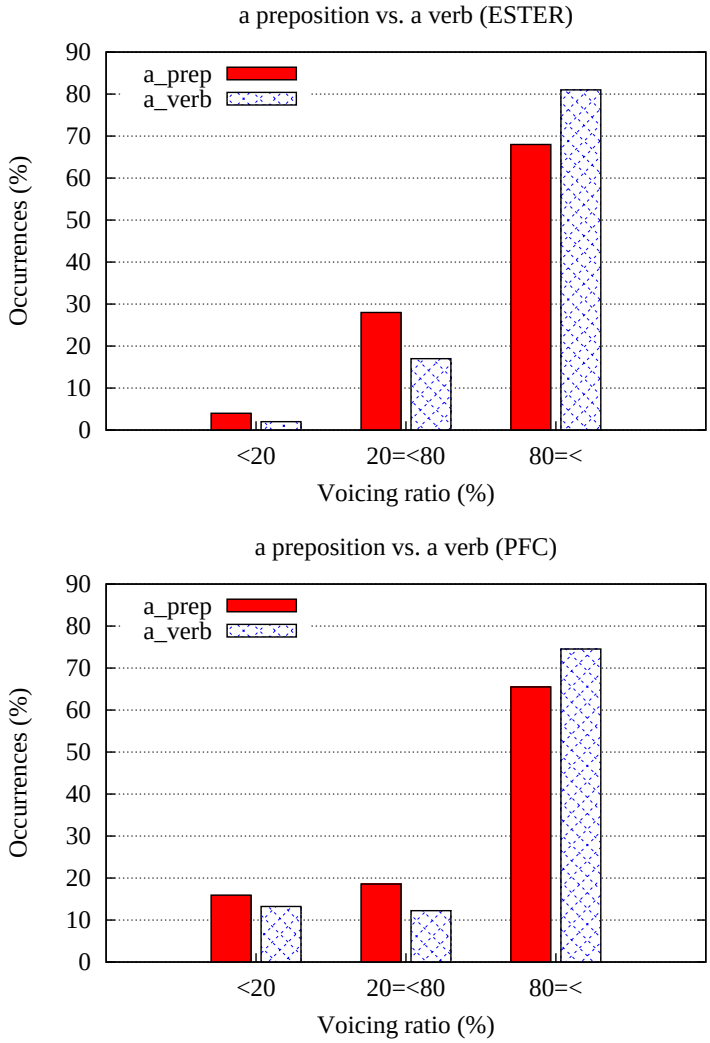


Figure 4.5: Bar charts of homophone pair *à*/*a* occurrence distribution according to the voicing ratio. Bar charts show: *à* (red), *a* (blue) **Top:** ESTER. **Bottom:** PFC.

and preposition. Short duration proportion is very low for the function words even in the PFC corpus which has generally high rate of short duration. Even there kept quite high distribution after 100 ms. For the voicing class, short duration proportions are lower than in devoicing class, but higher than partial voicing class. The duration distributions of the voicing class are more similar to the distribution figures showed in Figures 4.2 and 4.3.

The duration impact on voicing ratio reveals that short duration can be more seen in the devoicing class and less for the other classes. However, we cannot neglect that less rates in the partial voicing ratio are shown that short duration proportion is less than both of the two other classes. If short duration influence voicing ratio, the results can be that more voicing ratio is, less short duration proportion appears. Hence, the question is raised that not only duration impact influences voicing ratio, but also pause impact may do as we hypothesized that vowels in prosodic words/phrases at

Table 4.5: Short duration (30–40 ms) distributions in % corresponding to each voicing ratio class.

Words	*et vs. est*				*à vs. a*			
Corpus	ESTER		PFC		ESTER		PFC	
	et	*est*	*et*	*est*	*à*	*a*	*à*	*a*
<20	32%	64%	44%	63%	20%	29%	55%	73%
20≤80	4%	19%	7%	30%	6%	7%	16%	31%
≥80	19%	17%	28%	38%	26%	21%	43%	48%

the initial position of a prosodic word/phrase can be partially voiced because of voice onset time delay. With this hypothesis, we will look into the link between voicing ratio and pauses.

Correlation between voicing ratio and pauses

Table 4.6: Preceding (**top**) and following (**bottom**) pause distributions in % corresponding to each voicing ratio class.

Preceding pause								
Words	*et vs. est*				*à vs. a*			
Corpus	ESTER		PFC		ESTER		PFC	
	et	*est*	*et*	*est*	*à*	*a*	*à*	*a*
<20	71%	8%	69%	14%	56%	39%	44%	20%
20≤80	79%	23%	80%	9%	61%	47%	43%	9%
≥80	28%	3%	50%	5%	6%	4%	12%	4%

Following pause								
Words	*et vs. est*				*à vs. a*			
Corpus	ESTER		PFC		ESTER		PFC	
	et	*est*	*et*	*est*	*à*	*a*	*à*	*a*
<20	7%	5%	19%	11%	7%	7%	18%	17%
20≤80	7%	7%	16%	9%	6%	13%	11%	13%
≥80	7%	5%	17%	7%	2%	4%	8%	9%

Table 4.6 demonstrates the proportion of target word occurrences preceded or followed by a pause in each class. In the line with the general study of pause in preceding section (cf. section 4.2.3.2), preceding pauses are more significant than following pauses and conjunction/preposition than verbs.

In the devoicing class of preceding pause, it is noticed that conjunction and preposition are more preceded by a pause than the verb words. And the verb '*a*' is higher ratio than '*est*'. The ratios of this class are much higher than average pause percentage (see Table 4.4) excepting the verb *est*. In the "partial voicing" class of preceding pause, as much high as or higher ratios than the devoicing class are observed excluding *est* and *a* verbs for the PFC corpus. This may be because of the difference of speaking styles. Much less pauses are observed in "voicing class" than other

classes, even though the conjunction *et* is often preceded by a pause. This reveals that if a vowel is full voiced, this may mostly locate within a prosodic word/phrase.

As for the link to following pauses, as seen in the previous general study of pause, it is noted that following pauses are less important than preceding pauses. It is observed that the proportion of pauses is slightly less in the voicing class.

From the study of the correlation between voicing ratio and pauses, it is revealed that not only short duration may influence voicing ratio, but also preceding pauses can play a role of voicing quality.

Profiles of f_0

Vaissière [1991, p.112] claims that lexical words contrast with function words by pitch and tend to be uttered on the high register and function words on the low register. Then we measured the average f_0 values to verify how f_0 values are influenced according to different grammatical categories and speaking styles. We also would like to investigate how f_0 values vary with voicing ratio. We noticed that the results of voicing ratio inferior to 20% ("devoicing" class) are not "reliable" due to few tokens.

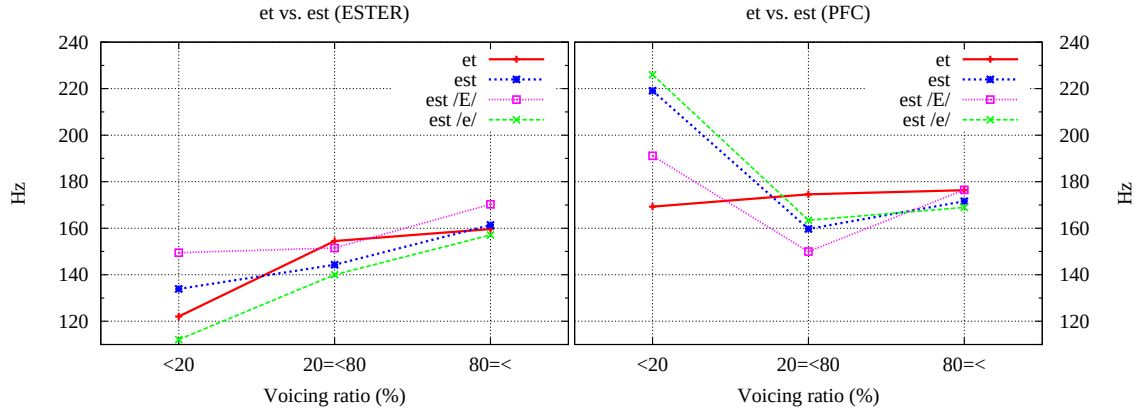


Figure 4.6: Average f_0 for homophone pair *et* and *est* according to the voicing ratio. *et* (red $+$) and *est* (blue $*$, /ɛ/ pink $\square$, /e/ green $\times$). **Left:** ESTER. **Right:** PFC.

Figure 4.6 shows the average f_0 results for the pair ***et/est*** according to voicing ratio: ESTER corpus (left) and PFC corpus (right). As the duration distribution figure in Figure 4.2, for the word *est*, three curves are illustrated: first curve, in blue line with asterisk $*$, gathered the two pronunciations [ɛ] (canonical pronunciation) and [e] (variant, but majority). Second curve (in pink line with square $\square$) describes phone [ɛ] and third curve (in green line with times $\times$) represents phone [e]. For the ESTER corpus (left), f_0 values rise when voicing ratio increases for the two words. For the PFC corpus (right), the average f_0 values of the word *est* in the "devoicing" class is much higher than the two other classes. It can be presumed that f_0 is both unstable and unreliable in the "devoicing" class. Overall, average f_0 values are a little higher in the "voicing" class than in the "partial voicing" class. The mean f_0 of the word *et* is almost stable in all three voicing

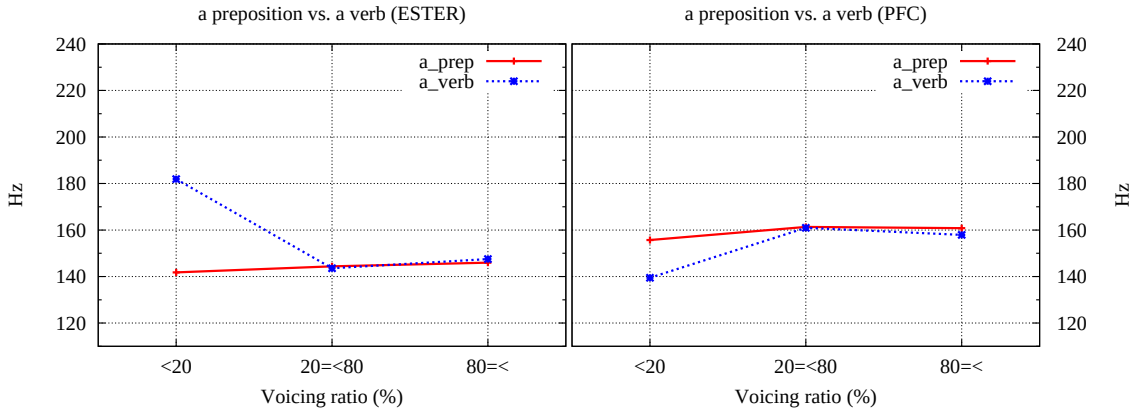


Figure 4.7: Average f_0 for homophone pair *à* (red $+$) vs. *a* (blue $*$) according to the voicing ratio. **Left:** ESTER. **Right:** PFC.

ratio classes. Contrary to our expectation, the average f_0 values of these two words did not show remarkable difference between them.

The comparison of the pair ***à/a*** is illustrated in Figure 4.7 for the ESTER corpus (left) and the PFC corpus (right). Like the duration distribution in Figure 4.3, the word *à* is represented by a red line with $+$ point and *a* by a blue dotted line with $*$ point. The stable line for the preposition *à* can be observed in three classes for the two corpora. The values are stable in the "**partial voicing**" and "**voicing**" classes. The two f_0 values of *à* and *a* are showed almost the same profiles in the "partial voicing" and "voicing" classes. For the word *a*, we can observe the average f_0 in the "**devoicing**" class is much higher for the ESTER corpus and lower for the PFC corpus. But this result is not reliable.

Average f_0 results do not show significant differences between the homophone pairs. But voicing ratio analyses revealed that the lexical category words (verbs) remained more fully voiced than preposition/conjunction words.

4.2.3.4 Discussion on acoustic analyses and measurements

Different acoustic realizations of the two frequent homophone pairs have been examined. From these results of this subsection, we can observe that the acoustic and prosodic characteristics of two homophone pairs, studied from about more than ten hours of speech and thousands of occurrences for each examined word, present some differences. Parameters such as duration and voicing ratio allow (partially) distinguishing our homophone words. The comparison of duration distributions showed that the word *et* tends to last longer (with a flatter distribution) than the verb *est*. The duration differences were less remarkable for the *à*/*a* pair than for the *et*/*est* pair. However, differences in voicing have been noticed for both pairs. The functional words (*et*, *à*) have weaker voicing ratios than the verbs (*est*, *a*). Co-occurrence of left and right pause with the target words further contributes to distinguish *et*/*à* from *est*/*a* due to their position within prosodic words/phrases. Thus we could wonder if this kind of acoustic attributes could be useful to automatically discriminate such word pairs. Following these considerations, we defined a set of attributes to characterize the

target words and explore their relevance using data mining techniques. In the next section (cf. section 4.2.4), we describe, first of all, the investigated acoustic and prosodic descriptors before presenting the adopted method to discriminate the analyzed homophone pairs.

4.2.4 Automatic homophone classification

The previous corpus-based acoustic measurement results encouraged us in establishing a set of acoustic and prosodic parameters to represent differences between the two homophone pairs. In this section we address the matter of the automatic separability of the two pairs thanks to appropriate acoustic and prosodic attributes. Automatic classification tests for the homophone pairs were conducted using acoustic and prosodic parameters derived from the results of the preceding section 4.2.3. Then 62 attributes were defined and tested using a wide range of classifiers (25 algorithms among these Bayesian classifiers, Decision Trees, Support Vector Machine, etc.) implemented in the data mining software WEKA [Witten and Frank, 2005] developed at the University of Waikato in New Zealand.

4.2.4.1 Classification experiment using cross-validation

Two distinct data sets were established for classification: one is a **training** set for classifiers and the other is a **test** set proper. Our data include different speakers, genders, accents and speaking time per speakers. It suggests a variability which may complex the selection of the training and test sets. The cross-validation method can solve this problem of data set choice. In our study, the classification experiments were estimated using a "K-fold cross-validation". It consists in the following: first, all data are divided into K-folds. Second, one of K-folds is used for the test and $K-1$ folds are used for training. Then, this process is repeated K times. Each of the K results is gathered and scores are averaged to produce the final result. In our study, we used 10-folds. This number is commonly used for the K-fold cross-validation.

4.2.4.2 Attribute definition

62 acoustic and prosodic attributes have been defined for the automatic classification. They were chosen in order to model both the target word (**intra-phonemic** attributes) and its relation to the context (**inter-phonemic** attributes).

Intra-phonemic attributes (40): duration, f_0, voicing ratio, first three formants (F1, F2, F3), intensity. Except duration, global mean values by segments and begin, middle, end values are computed. We also calculated the differences (Δ) between begin-middle, middle-end and begin-end for the f_0, three formants and intensity (cf. Figure 4.8). Thus 1 attribute for duration, 4 attributes for voicing ratio (global, begin, middle, and end), and 7 attributes for f_0, three formants, and intensity (mean, begin, middle, end, and 3 Δ values: Δbegin-mid, Δmid-end, Δbegin-end).

Inter-phonemic attributes (22): duration, f_0, three formants, intensity, pauses. Inter-phonemic attributes give dynamic information about our target word's context. Measurements are carried out issuing the closest left/right vowel segments and our target vocalic segment (see Figure 4.9). Duration attributes were measured as following: the difference between a center segment duration

of a target word and a center segment duration of a previous/following vowel, even though there are consonants or pauses between these vowels. For f_0, formants, and intensity, different values (noted as Δ) were calculated as the difference between the mean values of the target word vowel and the previous/following vowel. In addition, the difference between two mean values of previous and following vowels of the target word was considered as well. Finally, left-right pause attributes were added looking into segments before and after. So inter-phonemic attributes are composed of 3 Δ duration (Δtarget-preceding, Δfollowing-target, Δfollowing-preceding), 3 Δ values for f_0, 3 formants, and intensity, 2 left pauses and 2 right pauses.

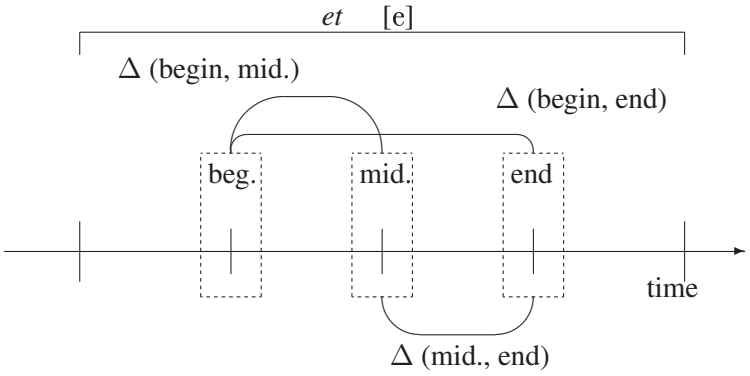


Figure 4.8: Intra-phonemic measurements.

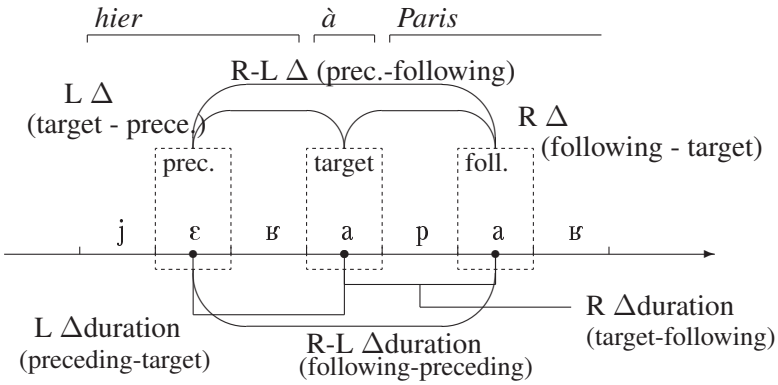


Figure 4.9: Inter-phonemic measurements with the words: *hier à Paris* ("yesterday in Paris").

4.2.4.3 Classification for all 62 attributes

62 attributes were thus available for automatic classification tests. Among our 62 attributes, we categorized these attributes firstly using linguistic knowledge: formants (30 attributes), prosody (32 attributes), intra-segmental attributes (40 attributes), and inter-segmental attributes (22 attributes).

The results of correct word identification are presented in Table 4.7: ***et/est*** pair results on the **top** and ***à/a*** pair on the **bottom**. Identification scores are classified in terms of the algorithm giving the

Table 4.7: Comparison of homophone word classification (in %) according to 62 attribute types. The best algorithm classification result, mean on 10 best algorithms and mean on 25 algorithms are presented. The employed attribute number for each category is demonstrated in parentheses. **Top**: *et*/*est* pair, ESTER on the left, PFC on the right. **Bottom**: *à*/*a* pair, ESTER on the left, PFC on the right.

Words	*et* vs. *est*					
Corpus	ESTER			PFC		
	best	10 best	mean	best	10 best	mean
all (62)	79.8	77.8	**71.3**	83.1	81.1	**76.3**
formants (30)	67.5	65.9	62.3	66.6	65.3	62.7
prosody (32)	79.5	77.7	**70.9**	82.4	81.0	**77.3**
intra- (40)	73.2	71.3	65.7	71.7	70.4	67.0
inter- (22)	75.7	74.4	**69.2**	81.2	80.5	**77.0**

Words	*à* vs. *a*					
Corpus	ESTER			PFC		
	best	10 best	mean	best	10 best	mean
all (62)	72.9	71.4	66.3	69.4	66.4	61.6
formants (30)	69.0	67.7	64.3	62.7	61.2	58.5
prosody (32)	72.3	70.6	65.6	67.7	65.9	60.7
intra- (40)	68.9	68.0	64.0	60.0	59.3	57.0
inter- (22)	71.0	70.1	65.5	65.9	65.1	60.1

best score among 25 tested algorithms, the mean of 10 best score algorithms and the mean of all 25 tested algorithms. 14 of 20 results in the columns of 'best' in Table 4.7 chose Logistic Model Trees (LMT) [Landwehr *et al.*, 2005] as best algorithm.

The results in Table 4.7 show that the *et*/*est* pair (top) is much better classified than the *à*/*a* pair (bottom). Comparing to the mean results of all 62 attributes (**all (62)**) between two pairs, we can observe 5% of difference for the ESTER corpus (71.3% vs. 66.3%) and about 15% for the PFC corpus (76.3% vs. 61.6%). These results can be relevant to the parameter analyses of "pause" co-occurrences, duration, and f_0 in section 4.2.3 in which we noticed that the *et*/*est* pair was better distinguishable than the *à*/*a* pair. It can also be related to the fact that one third of the occurrences of the verb *est* are not real homophone (canonical pronunciation /ɛ/) with *et*, resulting in more discriminable attributes. The results of *et*/*est* are particularly promising for the PFC corpus because spontaneous speech in general presents more errors during automatic transcription processing.

The prosodic attribute results (**prosody (32)**) remain almost identical in comparison with the all 62 attribute results (all (62)) for both pairs and both corpora (±2%). The inter-phonemic attributes (**inter- (22)**) also give almost as good results as all 62 attributes and prosodic attributes (±2%) excepting the *et*/*est* pair of ESTER corpus ('best' and '10 best' ±4%). These differences are bigger for the *et*/*est* pair than the *à*/*a* pair. The results from formants (**formants (30)**) and intra-

phonemic (**intra- (40)**) attributes are less efficient, especially for the *et*/*est* pair. By contrast, the differences between *à* and *a* words are not as large as for the *et*/*est* pair. Prosodic and inter-phonemic parameters are particularly interesting to distinguish our homophone words.

4.2.4.4 Attribute Selection

We also hypothesized that among these 62 attributes, some are more relevant than the others as revealed from the results from 62 parameters: prosodic and inter-phonemic parameters. However these results did not exactly show which prosodic or inter-phonemic parameters are promising. So we performed automatic attribute selection on our corpora to find and choose suitable attributes for improving the performance of learning algorithms [Guyon and Elisseeff, 2003]. For the attribute selection, we referred to the study of identification of foreign-accented French using data mining techniques [Vieru-Dimulescu *et al.*, 2007]. Vieru-Dimulescu *et al.* performed 15 attribute selection algorithms implemented in the WEKA software. To sort out the different ranking scores from the different algorithms, Vieru-Dimulescu *et al.* used the following equation:

$$score = \frac{p}{n} \sum_{i=1}^{n} (C - rank_i) \tag{4.2}$$

where $rank_i$ is the attribute rank in the i-th algorithm, p is the number of algorithms which select this attribute, n is the number of algorithms used, and C is a constant fixed to 100.

We applied this equation to our attribute selection. In our study, six attribute selection algorithms implemented in Weka (e.g. Gain Ratio, Information Gain, oneR, Relief, Symmetrical Uncertainty, Chi-squared statistic) were used in a 10-fold cross-validation scheme with Ranker search method, which orders the attributes according to their relevance. Thus the equation can be presented as following:

$$score = \frac{p}{6} \sum_{i=1}^{6} (C - rank_i) \tag{4.3}$$

The results of 15 best attributes for each pair are shown in Table 4.8 for the ESTER corpus and in Table 4.9 for the PFC corpus. Each selected attribute is presented with its attached classes such as formants, prosody, intra-phoneme, and inter-phoneme. And 15 best attributes selected by all of both of two corpora and of two pairs are presented in Table 4.10.

As mentioned above, among our 62 attributes, 32 attributes may be considered as prosodic parameters and the other 30 attributes are relevant to formants. Observing these 15 selected attributes, we can notice that about two third of them concern prosody parameters in both pairs and corpora: pause, intensity, duration, f_0. This points out the importance of prosodic parameters. Through both pairs and corpora, the most discriminating parameter is "**L**(eft) **pause**" which shows the role of the pause before the target word (phone). This parameter is the first of the three categories except the *à*/*a* pair of the ESTER corpus ranked at the fourth. This result is in line with our acoustic measurements (see Table 4.4). Intensity variation parameters also play an important role, since we can observe two attributes concerning intensity (Δbegin-end, Δbegin-middle) within 10 best

Table 4.8: 15 best attributes of the ESTER corpus (*intra-phonemic* (in italic) and **inter-phonemic** (in bold) selected by WEKA attribute selection algorithms with their concerning classes (formant, prosody, intra-phonemic, inter-phonemic).

Corpus	ESTER									
Words	*et* vs. *est*					*à* vs. *a*				
	Attributes	For.	Pro.	Intra	Inter	Attributes	For.	Pro.	Intra	Inter
1	**L pause**		×		×	*f_0 beg v. ratio*		×	×	
2	*Δint. beg-end*		×	×		**R Δduration**		×		×
3	*int. beg*		×	×		*F2 beg*	×		×	
4	*Δint. mid-end*		×	×		**L pause**		×		×
5	*duration*		×		×	*F2 mid.*	×		×	
6	*Δint. beg-mid.*		×	×		*f_0 v. ratio*		×	×	
7	*F2 end*	×		×		*f_0 mid. v. ratio*		×	×	
8	*F2 mid.*	×		×		*F1 beg*	×		×	
9	*Δf_0 beg-end*		×	×		*Δint. beg-end*		×	×	
10	*f_0 v. ratio*		×	×		*Δint. beg-mid.*		×	×	
11	*int. mean*		×	×		*f_0 beg*		×	×	
12	**L ΔF2**	×			×	*ΔF2 beg-end*	×		×	
13	*F2 mean*	×		×		*Δf_0 beg-end*		×	×	
14	**L Δint.**		×		×	*F2 mean*	×		×	
15	*ΔF2 beg-end*	×		×		**L-R Δduration**		×		×

Table 4.9: 15 best attributes of the PFC corpus (*intra-phonemic* (in italic) and **inter-phonemic** (in bold) selected by WEKA attribute selection algorithms with their concerning classes (formant, prosody, intra-phonemic, inter-phonemic).

Corpus	PFC									
Words	*et* vs. *est*					*à* vs. *a*				
	Attributes	For.	Pro.	Intra	Inter	Attributes	For.	Pro.	Intra	Inter
1	**L pause**		×		×	**L pause**		×		×
2	*duration*		×	×		*Δint. beg-end*		×	×	
3	*Δint. beg-end*		×	×		**L Δint.**		×		×
4	*Δint. beg-mid*		×	×		*Δint. beg-mid.*		×	×	
5	*Δint. mid-end*		×	×		**L ΔF2**	×			×
6	*ΔF2 beg-end*		×	×		*F2 beg*	×		×	
7	*f_0 v. ratio*		×	×		**LΔDuration**		×		×
8	*F2 mean*	×		×		*F1 mean*	×		×	
9	**L Δduration**		×		×	**R ΔDuration**		×		×
10	*F2 end*	×		×		*F1 beg*	×		×	
11	**L-R Δduration**		×		×	*F1 end*	×		×	
12	*Δf_0 beg-end*		×	×		*f_0 beg v. ratio*		×	×	
13	*F2 mid.*	×		×		*F1 mid.*	×		×	
14	*ΔF2 beg-mid.*	×		×		*Δint. mid.-end*		×	×	
15	*Δf_0 beg-mid.*		×	×		**L-R Δduration**		×		×

Table 4.10: 15 best attributes of the two corpora (ESTER and PFC) and the two pairs (*intra-phonemic* (in italic) and **inter-phonemic** (in bold) selected by WEKA attribute selection algorithms with their concerning classes (formant, prosody, intra-phonemic, inter-phonemic).

Corpus	2 corpora (ESTER & PFC)				
Words	2 pairs: *et* vs. *est* and *à* vs. *a*				
	Attributes	For.	Pro.	Intra	Inter
1	**L pause**		×		×
2	Δ*int. beg-end*		×	×	
3	Δ*int. beg-mid.*		×	×	
4	*F2 mid.*	×		×	
5	Δ*int. mid-end*		×	×	
6	*F2 beg*	×		×	
7	Δf_0 *beg-end*		×	×	
8	*F2 mean*	×		×	
9	*F2 end*	×		×	
10	Δ*F2 beg-end*	×		×	
11	f_0 *v. ratio*		×	×	
12	Δf_0 *beg-mid.*		×	×	
13	f_0 *beg v. ratio*		×	×	
14	*duration*		×	×	
15	**L ΔF2**	×			×

attributes in the ESTER and PFC corpora for both pairs. It shows that intensity dynamic attributes (noted as Δ*int.*) within a phoneme are more important than intensity static attributes (mean, begin, middle, end). Two intra-phonemic intensity static attributes were worthless selected for the *et*/*est* pair in the ESTER corpus.

With respect to the *et*/*est* pair, we may note that the second formant (F2) intra and inter phone/phoneme attributes frequently appear in both corpora ESTER and PFC: 5 attributes for each corpus. The phonemic duration attribute is also selected which is in line with our earlier measurements (cf. Figure 4.2). The f_0 voicing ratio is listed for the two corpora as well which also corroborate the importance of the measured voicing ratio differences (see Figure 4.4). Inter-phonemic duration attributes are not negligible for the PFC corpus.

Concerning the *à*/*a* pair, not only F2 attributes appear among the 15 best attributes, but also F1 ones, mainly occurring in the PFC corpus. Differences between two corpora are that the PFC corpus has more inter-phonemic attributes (6 attributes among 15). For the PFC corpus inter-phonemic duration attributes are important whereas for the ESTER corpus, attributes linked to f_0 voice quality appear to be key attributes (5 attributes).

We can notice that 11 attributes are common between the *et*/*est* pair of two corpora and 8 attributes for the *à*/*a* pair. But just 3 attributes are shared between both pairs and both corpora: **Left pause**, Δ*intensity begin-end* and Δ*intensity begin-middle* which may be related to the homophone pairs' shared opposition (verbs vs. conjunction/preposition). However, different words feature different attributes, so it seems difficult to find a common attribute set to discriminate all word pairs.

29 attributes in 62 are appeared in Tables 4.8, 4.9, and 4.10 as the 15 best selected attributes. On the contrary, 33 attributes were not selected none of the pairs nor the corpora. 4 attributes concerning with duration were involved with the selected attributes. However, none of the third formant (F3) attributes were chosen which reach 10 attributes. As mentioned earlier, F3 contributes to identify a round or unrounded lip position like between /i/–/y/, so F3 values might be not helpful to distinguish our investigated homophones. Intra- and inter- dynamic (Δ) F1 values were not included in the selection. As for F2, most of attributes are chosen in the selection except 3 attributes (Δ*F2 mid-end*, **R** Δ**F2**, **L-R** Δ**F2**). For the first three formants, we can say that intra-phonemic attributes are more important than inter-phonemic attributes. 7 attributes in 14 concerning f_0 were not considered in the 15 best selections: 3 static values (mean, middle, end), voicing ratio of end, intra-dynamic value (middle-end), and inter-dynamic values: (**L** Δ, **R** Δ). Voicing ratio may play a more important role than raw static f_0 values. As for intensity values, intra-static values (middle and end), inter-dynamic values (**R** Δ**int.** and **L-R** Δ**int.**) were less salient to be selected. However Δintra-phonemic values may be major cues. Just left immediate pause is taken consideration into the best 15 selected attributes and not right pause.

4.2.4.5 Results for different attribute types

From the attribute selection, we defined two categories for classification: 15 best attributes for each pair and each corpus (15 attributes), and 15 best attributes from all of pairs and corpora (15 attributes). The results are shown in Table 4.11 in comparison with the results from all 62 attributes and their different categories (prosody, formants, inter-phonemic and intra-phonemic parameters).

The results of correct word identification are presented in Table 4.11: ***et*/*est*** pair results on the **top** and ***à*/*a*** pair on the **bottom**. Identification scores are classified in terms of the algorithm giving the best score among 25 tested algorithms, the mean of 10 best score algorithms and the mean of all 25 tested algorithms. 22 of 28 results in the columns of 'best' in Table 4.11 chose Logistic Model Trees (LMT) [Landwehr *et al.*, 2005] as best algorithm.

The results in Table 4.11 show that the *et*/*est* pair (top) is much better classified than the *à*/*a* pair (bottom). 15 selected attributes (**15 best att. (15)**) and 15 selected attributes (**15 all best att. (15)**) results for each corpus and each pair show similar results as inter-phonemic and prosodic attribute ones. By contrast, the differences between the *à* and *a* words are not as remarkable as for the *et*/*est* pair. Even though F2 attributes are selected in the 15 best attributes for each pair and each corpus, 30 formant attributes do not seem to play a prominent role to classify our homophone words. Table 4.11 shows that the set of 15 best, on the set of inter-phonemic attributes or the set of prosodic attributes, produce almost equivalent results to those from all 62 attributes. Thus, prosodic and inter-phonemic parameters are particularly interesting to distinguish our homophone words.

4.2.4.6 Discussion on automatic homophone classification

We defined 62 intra- and inter- phonemic acoustic and prosodic measures potentially relevant for the automatic classification of two homophone pairs and we tested 25 different algorithms implemented in the WEKA software. Results are promising: classification scores of correct iden-

Table 4.11: Comparison of homophone word classification (in %) according to attribute types. The best algorithm classification result, mean on 10 best algorithms and mean on 25 algorithms are presented. The employed attribute number for each category is demonstrated in parentheses. **Top**: *et*/*est* pair, ESTER on the left, PFC on the right. **Bottom**: *à*/*a* pair, ESTER on the left, PFC on the right.

Words	*et* vs. *est*					
Corpus	ESTER			PFC		
	best	10 best	mean	best	10 best	mean
all (62)	79.8	77.8	**71.3**	83.1	81.1	**76.3**
formants (30)	67.5	65.9	62.3	66.6	65.3	62.7
prosody (32)	79.5	77.7	**70.9**	82.4	81.0	**77.3**
intra- (40)	73.2	71.3	65.7	71.7	70.4	67.0
inter- (22)	75.7	74.4	**69.2**	81.2	80.5	**77.0**
15 best att. (15)	77.6	76.4	**70.5**	81.4	80.5	**76.9**
15 all best att. (15)	76.1	75.0	**69.5**	80.4	80.3	**76.7**

Words	*à* vs. *a*					
Corpus	ESTER			PFC		
	best	10 best	mean	best	10 best	mean
all (62)	72.9	71.4	66.3	69.4	66.4	61.6
formants (30)	69.0	67.7	64.3	62.7	61.2	58.5
prosody (32)	72.3	70.6	65.6	67.7	65.9	60.7
intra- (40)	68.9	68.0	64.0	60.0	59.3	57.0
inter- (22)	71.0	70.1	65.5	65.9	65.1	60.1
15 best att. (15)	70.9	69.7	65.5	67.5	65.4	61.2
15 all best att. (15)	68.9	67.8	64.2	62.1	60.9	58.4

tification for *et* vs. *est* are in the range of 62 – 71% on average for the ESTER corpus and 63 – 77% on average for the PFC corpus. The *à*/*a* pair appears to be less discriminable: mean identification around 64 – 66% for the ESTER corpus and 57 – 62% for the PFC corpus were observed. The random results of our pairs could be 50%, thus the results of the *à*/*a* pair are not far from the random results. The automatic classification results illustrate that: 1) the attributes concerning prosody (intensity, duration, f_0, pause) are better than the first three formant results in line with the homophone hypothesis; 2) dynamic (Δ) features perform better than static features. The attribute selection confirms the important attributes to discriminate our homophones such as left pause, Δintensity, F2, Δf_0, voicing ratio, and duration. These selected features revealed as good results as the results from all 62 attributes. An acoustic realization seems to vary slightly (fine phonetic detail) as a function of part-of-speech and/or position in a prosodic word/phrase.

4.3 Perceptual transcription test

The previous section was concerned with automatic classification for homophone words in which prosodic parameters were important cues to discriminate between homophonic words. In this section, we would like to verify if humans also use acoustic-prosodic parameters to discriminate homophone words or if they tend to use context information similar to n-gram language models (LMs) for ASR systems.

During the last decade, several speech studies have established that human accuracy significantly outperforms machine accuracy on transcription tasks. These observations are particularly true when a large embedding context (complete and long sentences) is provided. They highlight that aspects of variation, such as pronunciation variants, noise, disfluencies, ungrammatical sentences, accents, which still remain important challenges for current automatic speech recognition (ASR) systems, are well managed by human listeners. Word error rates (WER) of an order of magnitude higher were reported for ASR systems as compared to human listeners on English sentences taken from read continuous speech (CSR'94 spoke 10 and CSR'95 Hub3) databases under various SNR (signal-to-noise ratio) and microphone conditions [Deshmukh *et al.*, 1996]. A similar gap in performance between humans and automatic decoders has been reported for spontaneous speech [Lippmann, 1997]. An interesting study [Shinozaki and Furui, 2003] in Japanese aimed at reproducing contextual information conditions of automatic speech decoders for human perception experiments. Stimuli comprising one target word embedded in a one word left/right context allow simulating word bigram networks as after used by automatic decoders. In this very limited context condition, results indicate degraded human performances compared to the previous studies: error rate gap between humans an ASR systems no longer corresponds to an order of magnitude. Nonetheless they remain roughly half those of the recognizers. The comparison of these different studies highlights the importance of lexical context for accurate human transcription; the information is not exclusively locally grasped from the acoustic signal.

In line with [Shinozaki and Furui, 2003], if such parameters have perceptual salience, the perceptual test aimed at verifying the reliability of the parameters allowing discriminating the homophones and in particular the context. We investigate a case study involving the *et*/*est* homophone pair which is among the most common errors encountered in automatic transcription of French: the confusion between, and more generally speaking, the erroneous transcription of two homophonic words *et* ("and") and *est* ("to be") as section 4.2. But the other pair *à* and *a* was not considered with perceptual test. The frequency of the below studied items *et* and *est* can be related to their polysemy and propensity to occur in a large variety of contexts. However, the two words correspond to different part of speech, i.e. coordinative conjunction (*et*) and third person singular present-tense of the verb "to be" (*est*). Consequently, they occupy distinct positions within prosodic words/phrases and more largely, within sentences. These differences in terms of grammatical behavior enable to believe the existence of acoustic and prosodic peculiarities of the two words which might possibly help humans to disambiguate them.

4.3.1 Corpus for perceptual evaluation

For perceptual test, we made use of the French Technolangue-ESTER corpus [Gravier *et al.*, 2004a] described in section 3.1.1 of chapter 3, consisting in recordings of broadcast news shows

from different francophone (French and Moroccan) radio stations. Transcription errors were extracted from the automatic transcriptions produced by the LIMSI speech recognition system developed for the 2005 ESTER evaluation [Gauvain *et al.*, 2005]. The ASR system made use of 4-gram language models (LMs) and context-dependent acoustic phone models.

4.3.2 Perceptual evaluation

The perceptual experimentation on the automatic transcription errors of the *et*/*est* homophones has been conducted with the aim of clarifying whether human word perception confirms or outperforms automatic decoding of the target words in a 7-gram (i.e. 4-gram left and 4-gram right) word context. The selected stimuli are based on ASR errors suggesting local ambiguity at least for the ASR system by such items. Table 4.12 below, shows some typical examples of transcription errors involving the target words *et* and *est*. The excerpts shown contain the target word in the middle of a 7-gram and are surrounded by three left and right neighboring words, thus integrating the maximum scope of the language model for the target word transcription. In many situations however, the ASR system backs off to lower n-grams, resulting in less than 7 words.

In particular, two questions have been addressed: (1) are the human transcriptions on the homophones *et*/*est* more accurate than the automatic ones in conditions corresponding to contextual n-gram constraints similar to those of automatic speech decoding; (2) if humans are more competitive, which of the linguistic levels of information (syntactic, semantic, prosodic, voice quality) may have potentially contributed.

Table 4.12: Examples of 7-gram stimuli with different types of errors: *et*/*est* confusion (Ex.1 “cold fever is the viral disease”), *est* within a syntagm substituted by another word (Ex. 2 “on the salaries is so formidable that”), *est* deletion (Ex. 3 “politics today it is essential to go into detail”).

Ex. 1	
REF	rhume de cerveau **est** la maladie virale
HYP	rhume de cerveau **et** la maladie virale
Ex. 2	
REF	sur les salaries **est si** formidable que
HYP	sur les salaries **ici** formidable que
Ex. 3	
REF	politique aujourd’hui il **est** essentiel d’approfondir
HYP	politique aujourd’hui il essentiel d’approfondir

4.3.2.1 Test material selection

Stimuli comprising the target *et*/*est* homophones in limited n-gram contexts are selected. The test material consisted in 83 chunks extracted from the ESTER development corpus (dev04). We call chunk a 7-word string with the target word as center (Table 4.12). Forced alignment of the reference manual transcriptions is carried out and selected chunks are extracted automatically. In

Table 4.12, **REF** means the reference which is transcribed by humans and **HYP** is the hypothesis which is the result obtained by the output of the ASR system.

The choice of 7-gram chunks aims at providing the human subjects with as much information around the target word as used by a 4-gram LM-based transcription system in optimal conditions. Stimuli mainly contain an erroneously transcribed *et* or *est* in central position (68 stimuli). They also illustrate different types of errors observed in the ESTER development corpus: insertions, deletions, substitutions of the target words only or of the target words together with surrounding words (target word within a syntagm). Selected errors aim at covering all the erroneous transcription case figures encountered in the ESTER corpus (as illustrated in Table 4.12, Ex. 1, 2, 3 above). Some distracting items consisting in 7-gram chunks correctly transcribed as well as different target words were also added (see Table 4.13).

Table 4.13: Examples of 7-grams distracting stimulus. Different target word stimulus (Ex. 1 "recreate a bourgeois interior, the decoration not") and correctly transcribed stimulus (Ex. 2 "incredible sadness, this is a monster this").

Ex. 1	
REF	recréer un intérieur bourgeois le décor ne
HYP	recréer un intérieur bourgeois le décor ne
Ex. 2	
REF	tristesse inouïe c'**est** un ogre c'
HYP	tristesse inouïe c'**est** un ogre c'

Table 4.14 sums up the different types of stimuli corresponding to contexts giving rise or not to automatic transcription errors. Among 83 stimuli, 5 distractor stimuli do not include *et*/*est* words in the middle. The *et*/*est* words are correctly transcribed by the ASR system for 10 stimuli. 20 stimuli have symmetric ASR confusions of *et*/*est*. 48 stimuli are composed of 6 stimuli of 4 error types (insertion, deletion, substitution of a target word, substitution within a sytagm) and 2 target words (*et*/*est*).

Table 4.14: Types of automatic transcription errors illustrated by the 83 selected stimuli.

Chunks (nbr.)	**Types of errors**
5 distractors	Stimuli without *et*/*est* in the middle
10 corrects	Stimuli with *et*/*est* correctly transcribed by the system
20 *et/est symmetric confusions*	Stimuli with symmetric ASR confusions of *et*/*est*
48 other errors (6 stimuli/4 types/2 target word)	Stimuli with other errors: insertions, deletions, erroneous transcription of target word alone or within a syntagm.

4.3.2.2 Test protocol

Sixty native French subjects took part in the experiment. They were not informed of either the target words or the selection criteria, or the fixed chunk length. The 60 subjects were divided into two sub-groups according to different test conditions.

A first condition focused on the role of context information without providing the acoustic stimuli. This condition stimulates perfect homophony and allows comparing human performances to ASR's language model contributions. Thus condition is called "LM".

The second test condition corresponds to the main test: subjects listen to the stimuli transcribe them and they succeed if the central target is correct. This condition is called "AM+LM" as both acoustic and higher level context information is available.

20 subjects performed a local **language model (LM)** condition test on the 30 chunks (10 correct chunks + 20 symmetric confusion chunks) focusing on *et* vs. *est* confusion (i.e. the stimuli for which the system transcribed *et* by *est* and vice-versa to which we added the 10 correct chunks as control stimuli). They had to fill in the written version of the 30 chunks using the most plausible item *et* or *est*, as suggested by the 3-word left and right contexts. Figure 4.10 below gives a schematic representation of the written test protocol. This condition is a simplification of the ASR ambiguity processing, which has to score all possible expanded ambiguities of the uttered sequence. This test assumes perfect homophony for the target. The rationale of this test is twofold: contribution of syntactic/semantic information of the written sequence to solve ambiguity; humans' focus on local ambiguity.

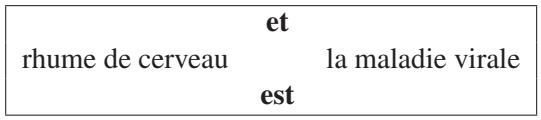


Figure 4.10: Written test corresponding to a local LM condition.

40 subjects underwent the **acoustic+language model (AM+LM)** condition test. The 83 stimuli have been submitted to two groups of 20 subjects via a web interface. Listeners were provided with the audio excerpt corresponding to the 7-gram chunk and had to transcribe the entire chunk. Each group of 20 subjects listened to and transcribed half of the stimuli. This choice was made to limit the duration of perceptual test to less than one hour: subjects were spending about 20 x RT (real-time) to transcribe a stimulus (compared to automatic transcriptions processed in 10 x RT). The two groups were comparable in terms of age and background.

4.3.2.3 Results of perceptual transcription test

Results are measured in terms of erroneous transcription of the target words compared to the reference transcriptions. Human error rates are then compared to ASR word error rates. Detailed WER rates for the different stimuli sets and conditions are reported in Table 4.15. We recall that the ASR error was the criterion for stimuli selection which entails 100% WER for these ASR stimuli.

For the **Humans AM+LM condition**, results of the perceptual test show that no errors are produced on the distractor stimuli and that a marginal error rate (1.4%) is measured on the 10 perfectly decoded stimuli by the ASR system. However on the stimuli subset corresponding to system confusions, an important increase in the human error rate can be observed. The results suggest that

Table 4.15: WER on 4 stimuli subsets in different automatic/human transcription conditions: ASR (selection criteria); LM (written test on local ambiguity); AM+LM (audio test).

	WER (word error rates)		
Stimuli	**ASR**	**Humans**	
Condition	**AM+LM**	**AM+LM**	**LM**
5 distractors	0	0	-
10 corrects	0	1.4	8.2
20 *et/est symmetric confusions*	100	25.5	27.6
48 other *et/est errors* (6 sets/4 types/2 target words)	100	16.0	-

stimuli which are difficult for ASR systems are also problematic for humans, even though their overall performance tends to be 4 times better. A statistical significance test was carried out to measure the validity of this result. The potential correlation between human and automatic transcription solutions has been checked statistically (with one factor "system answer for target word" ANOVA, using "correct" vs. "erroneous" as nominals). The factor "system answer for the target word" is statistically significant for both LM ($F(34,07)$, $p<0.0001$) and AM+LM ($F(38,22)$, $p<0.0001$) conditions.

Consequently, humans produce more errors on stimuli misrecognized by the ASR system. Reversely humans are almost error free on the correctly decoded stimuli. Humans appear to be 4-5 times more accurate than ASR system in this particular test condition.

We also checked which of the two words *et* vs. *est* is more ambiguous. An ANOVA analysis (with one factor "target word", using "*et*" and "*est*" as nominals) showed that *est* was missed more frequently by human listeners than *et* ($F(38,95)$, $p<0.001$). Finally, when looking at different types of errors for each of the target words, namely insertions, deletions and erroneous transcriptions of the target word or of the target word and the surrounding words, one may notice that the type of ASR error and the number of errors produced by the listeners are positively correlated: the more ambiguous the local context the more frequently the correct solution is missed. Consequently, humans produce more frequently errors for the stimuli for which the system missed the target word and the surrounding context than for the stimuli for which the target word has been only deleted or inserted while the other surrounding words remained correctly transcribed. This finding suggests that the local linguistic ambiguity is problematic for both the ASR system and the humans. In case of local ambiguity the transcription forces "random" choices which are prone to error both for humans and ASR systems.

The **LM test** might be considered as the easiest one, as only a local ambiguity has to be worked out, while relying on the surrounding written words. We remind that the LM test represents the written version of the stimuli focusing on the symmetric *et/est* confusion. However the lack of punctuation (due to the ASR simulation protocol) probably adds some difficulty here. We compared the results for the LM test with the LM+AM test section focusing uniquely on the symmetric *et/est* confusion. The difference between the two conditions is statistically significant (one sample t-test, $p<0.0025$, $t=2.66$, $p=0.0078$) and the LM condition generates more errors than AM+LM condition. Better results on the AM+LM condition might be related to additional

structuring information of the audio signal: the lack of punctuation does most likely not allow retrieving information on syntactic structures which might rely on prosodic cues in AM+LM condition. However, no statistical difference has been observed when comparing the ratings for each of the target words, i.e. subjects are equally competitive in processing chunks with *et* or *est* and the target words seem to be equally ambiguous in the given word strings. This information suggests that the two polysemic words produce comparable contexts in terms of intrinsic degree of ambiguity in the French language and human subjects encounter similar challenges in processing them. It suggests that for the given string length, both humans and ASR systems leave some unresolved ambiguities, even though less numerous in the case of humans (at least as observed in this perceptual experimentation).

4.3.3 Discussion on perceptual evaluation

For human perceptual test results, we observe that almost no errors occur for the distractor stimuli and for the 10 stimuli without confusions on the target words. The context entailing symmetric *et*/*est* errors for ASR are thus highly ambiguous as well as contexts for which the local ambiguity concerns the target homophone word and the close surrounding context. A comparison between the system answers and the human transcriptions reveals that humans achieve better results in terms of correct *et*/*est* ratings for those stimuli correctly transcribed by the ASR system as well.

The perceptual test reveals that even though automatic and perceptual errors correlate positively, in conditions which attempt to approximate the information available for decision with a 4-gram language model, human listeners deal with local ambiguity more efficiently than the ASR system. Perceptual results seem to support the following hypothesis: differences in ratings for similar ambiguous syntactic structures suggest that prosodic/acoustic information may help in operating the right choice in terms of target word selection. This result is in line with the observed major role of prosodic and dynamic attributes with automatic classification task.

4.4 Summary and conclusion

A question addressed in this chapter is raised from whether homophone words for which ASR systems rely on language model (LM), can be discriminated by only acoustic and prosodic parameters. From our question, the presented study in this chapter is composed of three parts. The first part is concerned with prosodic parameter analyses. The second part presents automatic classification of homophone words using investigated acoustic-prosodic parameters and data mining techniques to find which parameters are efficient to discriminate homophone words. The third part deals with the perceptual transcription test to examine if humans can discriminate homophone words with limited context information like LM and if they use acoustic and prosodic parameters to identify these words.

Acoustic analyses show that the two homophone words *et* “and” and *est* “to be” may be distinguished thanks to some relevant acoustic and prosodic parameters (duration, voicing ratio, co-occurrence of pause with left context), but it still remains difficult to discriminate the *à* “to, at”/*a* “to have” homophone pair. The first experiment in automatic classification of the two pairs using

data mining techniques highlights the role of the prosodic (f_0, duration, voicing, and intensity) and contextual information (co-occurrence of pauses) in distinguishing the target words. This is revealed by the comparison of the full 62 acoustic-prosodic parameter results with smaller attribute subsets' results. The 15 best selected parameters, the 22 inter-phonemic parameters, and the 32 prosodic parameters are as much efficient as the full 62 tested parameters for automatic classification.

The perceptual test has been conducted in order to check human subjects' capacity to correctly transcribe the two homophone words (*et*/*est*) in ambiguous contexts. Perceptual results have been measured in terms of erroneous transcription of the target words compared to the reference transcriptions. Human error rates were then compared to ASR word error rates. Human transcriptions' analysis showed that distractor stimuli were error-free. A marginal error rate has been measured on the perfectly decoded stimuli by the ASR system. Reversely, on the stimuli subset corresponding to system confusions, an important increase in the human error rate could also be observed. Results suggest that local contextual ambiguity is problematic for both the ASR system and the humans.

Our results from two homophone pairs show that the acoustic realizations of homophone word pairs may undergo specific systematic changes (fine phonetic/prosodic detail) depending on their part-of-speech (POS) or their position within prosodic words/phrases. These may then be exploited in an automatic classification task. In next chapter, chapter 5, we would like to extend our research at the morpho-syntactic level in order to explore prosodic parameters: how do prosodic parameters vary according to more detailed grammatical categories, and to the syllabic length of words and syntagms.

Chapter 5

Large-scale prosodic analyses of French words and phrases

Our lack of knowledge concerning pronunciation variants represents a bottleneck to further improvements of ASR systems across conditions and in particular across speaking styles. ASR experience shows that the introduction of a large number of pronunciation variants into the pronunciation dictionary tends to increase homophone pronunciations between different word types. As an example, we may cite "*montre*" (show) /mɔ̃tʁ/ which can be pronounced as [mɔ̃t] in a consonantal right context and adding this variant to the pronunciation dictionary seems reasonable. As a side effect, the word "*monte*" (climb) becomes a homophone of "*montre*" and only higher level context (which is represented by word n-grams in ASR systems) may be able to make the sound choice between "*montre*" and "*monte*". More pronunciation variants thus enable the system to better account for the observed variation. However, due to increased homophone rates this may result in higher word error rates, unless the language model with the word n-grams is perfectly tuned for the input speech to be transcribed. Beyond simple word homophones, French also produces a wide range of multiword homophones (e.g. "*sévère*" (severe) vs. "*ces vers*" (these worms) vs. "*c'est vert*" (it's green); "*émoi*" (agitation) vs. "*et moi*" (and me); "*l'affiche*" (the poster) vs. "*la fiche*" (the form) with shifting and varying word boundaries).

It is relatively straightforward to introduce pronunciation variants using phonological rules (e.g. schwa insertion or deletion rules, liaison rules, consonant cluster reduction rules, voicing assimilation rules...). Applying these rules to the full system vocabulary results in high pronunciation variant rates and, as we just mentioned, in increased homophone rates. For frequently observed words in the acoustic training corpora, it is possible to select the most relevant variants from the observed tokens, and even estimate probabilities for all the different variants. However, the occurrence of lexical entries in the language follows a Zipf law, which entails a small number of word types with a large number of observed tokens, and a large number of types with very few tokens in the training data. This means that reliable variant probabilities cannot be estimated for a large number of words of the vocabulary. To tackle this problem, we have to move from words to word classes, where each class comprises a large number of tokens.

The following corpus-based study is motivated by the exposed problem. However, we do not directly address the problem of pronunciation variants, but we investigate overall prosodic prop-

erties of French on a lexical level. Although more traditional prosodic investigations focus on larger than word units, which may be called prosodic words, intonation phrases, accent phrases, chunks... depending on authors and analysis levels, our work mainly focuses on lexical units (for which pronunciation variants are sought). Pronunciation variants are often due to shorter or longer pronunciations, to added or deleted segments or even syllables, which then entail different prosodic characteristics. One may wonder whether pronunciation variants are due to varying prosodic constraints or vice versa. Without taking an option to answer this question, we shift our focus from pronunciation variants to prosodic realizations of French words or more precisely, of word classes of different syllabic lengths. We saw that empirical studies of pronunciation variants based on observed tokens in the training data was limited due to Zipf's law. To overcome these basic limitations, lexical **classes** were introduced which are motivated by prosodic (length) and/or syntactic criteria. The proposed study thus examines word classes of syllabic length 1 (monosyllabic words), of syllabic length 2 (bisyllabic words)... Each class of n-syllabic words thus includes a large number of tokens providing a statistically interesting basis for further investigations. From these, different types of average profiles are derived and discussed across classes and conditions.

In this chapter, we are interested in the following:

(i) Is the proposed methodology able to capture well-known prosodic properties of French (word-final lengthening, f_0 rise...)? If so, the proposed method may also be valuable to produce more detailed results, which can at least be considered as worthfull hypotheses for further in-depth studies. In particular, we are interested in the influence of word-final schwas or with respect to multiword homophones; we are interested in prosodic cues to word boundary locations.

(ii) What differences can be measured between different speaking styles? How can these differences be interpreted in light of ASR results? We want to recall that spontaneous speech entails much higher word error rates than prepared (journalistic) broadcast speech. In spontaneous face-to-face speech, involved speakers share more context information and as a consequence less information needs to be conveyed by the acoustic channel. This may at least partially explain the higher word error rates for spontaneous speech. The proposed study aims at clarifying how speaking style differences between prepared and spontaneous French are reflected on a prosodic level.

The importance of prosodic cues was observed for homophone word classification in the preceding chapter. In this chapter, we present extended prosodic analyses, (i.e. fundamental frequency (f_0), duration and intensity) taking benefit of large speech corpora. We made use of automatic processing (lexical and phonemic alignment, f_0 extraction, part-of-speech tagging) in order to study prosodic regularities of French words via average prosodic profiles. Some influential factors are taken into consideration for prosodic measurements: word syllable length, word-final schwa, duration, and part-of-speech. The following questions are addressed for this study: can specific prosodic profiles be measured for French words using large corpora? If so, how do they vary with respect to the cited influential factors? The aim of this study is then to produce empirical evidence from large corpora concerning the raised questions, in order to contribute to our knowledge of prosodic realizations in French words and their potential to contribute to the pronunciation variation and word segmentation problems. We consider this knowledge as a first step to the elaboration of word-class specific rules for pronunciation variants.

First we will present the speech corpora and the methodology in section 5.1. Then we will show the prosody calculation results for different word categories (grammatical/lexical words) in section 5.2. Next, section 5.3 presents a comparison of `noun` with most frequent bigram morpho-syntactic categories `determiner-noun`. A conclusion is presented in section 5.4.

5.1 Corpora and methodology

A major question of interest of this study is to investigate how prosodic features, and in particular duration and f_0, vary with the syllabic length of a word. We want to recall here that one motivation of this study was to gain some insight of potential pronunciation variants in French depending on speaking style and/or speaking rate. The question is then whether internal syllables of polysyllabic words tend to be temporally reduced and/or whether their average f_0 tends to decrease. Such measurements could be considered as indicators of shortened pronunciations in these parts of the words with potential deletions of one ore more segments, although the exact interpretation of the pronunciation variants goes beyond the scope of this study. Contrasting with the homophone study presented earlier on a very limited subset of the French vocabulary, the present findings concern the whole set of observed French lexical items and as such are innovative, as to our knowledge there have been no large-scale studies on the prosodic properties of French word classes.

5.1.1 Corpora

This study makes use of 13 hours of male speech from the manually transcribed French TECHNOLANGUE-ESTER corpus (news from different Francophone radio stations, see section 3.1.1 for more details) and from the PFC (Phonology of Contemporary French) corpus, see section 3.1.2. Only the spontaneous PFC speech data including *guided* and *free* conversations from male speakers are considered (6 hours). Table 5.1 gives a word level description of the two corpora according to mono- and polysyllabic words.

5.1.2 Methodology

In the following, we propose contrastive measurements on lexical subsets of increasing length, with increasing proportions of potential prosodic phrase boundaries. Acoustic correlates, namely f_0 and durations are examined with respect to supposed influential factors: word length expressed in number of syllables, presence or absence of word-final schwa, part-of-speech (POS), speaking rate. Figure 5.1 gives a schematic overview of the processing steps on the investigated data.

f_0 and intensity measurements: As described earlier (cf. section 3.2.2), fundamental frequency (f_0) and intensity values were measured each 5 milliseconds (ms) using the standard settings of PRAAT [Boersma and Weenink, 2008]. This 5 ms step results in at least six samples per segment (as the minimum phone duration is 30 ms). A large number of samples per segment entails an increased resolution for mean f_0 measurements as well as for voicing rate measurements. Voicing ratios per segment are defined as the number of voiced frames per total number of frames.

Lexical and phonemic alignment: The audio corpus was automatically aligned by the LIMSI speech recognition system [Gauvain *et al.*, 2005] producing word and phoneme segmentation.

Table 5.1: Quantitative ESTER and PFC corpus description with regard to word tokens of word syllable length n from 0 to 4 (n = 0-4). Separate counts are given for words without (top)/with (bottom) realized final schwa. *Syll.class* n_s states n: the number of full syllables; s: absence (0)/presence (1) of final schwa.

n	*Syll. class* n_s	*Occurrences #Words* ESTER	PFC	*Examples*
0	0_0	12578	13921	l'; d'; de
1	1_0	72249	65521	vingt; reste
2	2_0	36027	20346	beaucoup
3	3_0	15994	4959	notamment
4	4_0	6053	1408	présidentielle
n	*Syll. class*	*Occurrences #Words + /ə/*		*Examples*
0	0_1	12295	5056	de; le; que
1	1_1	3918	1642	reste; test
2	2_1	2087	716	ministre
3	3_1	698	208	véritable
4	4_1	174	49	nationalistes

During the alignment, the pronunciation dictionary allows for optional word-final schwas, if the standard pronunciation ends with a consonant (no matter whether there is a word-final graphical −e or not: e.g. the word test with standard pronunciation /tɛst/ admits a variant [tɛstə]) [Adda-Decker and Lamel, 1999]. As already stated before, aligned phone segments have a minimum duration of 30 ms and the precision of boundary locations is of 10 ms.

Part Of Speech (POS) tagging: The transcribed corpus was POS-tagged using the WMATCH tool, a regular expression engine [Galibert, 2009], including a French version of TREETAGGER [Schmid, 1994], to measure the influence of different POS classes and noun phrases on f_0 realizations. We introduced new tags for speech specificities such as breath, silence, hesitation or filled pause (e.g. *euh* for French language), and disfluencies[1]. By default, no such tags exist in the original tag set, as the POS taggers were designed for written text.

Word syllable length: After speech alignment, each uttered word was annotated by its *word syllable length*, corresponding to its pronunciation vowel count. In this way, the word syllable length of population (/pɔpylasjɔ̃/) is 4, as there are 4 full vowels /ɔ/, /y/, /a/ and /ɔ̃/. Each vowel of the corpus was annotated by its word syllable rank (e.g. in the former example vowel /y/ has rank 2 of 4). Table 5.1 shows the corpus composition according to classes of word syllable length n.

Syllable length class: All words with the same word syllable length n should be grouped within the same class of syllable length n. Word-final schwas did not count for the word syllable length, however, they were used to tag words into specific subsets (see bottom of Table 5.1) as one might

[1]Disfluencies comprise repetitions, repairs, filled pauses, false starts, etc. [Shriberg, 2001]. In our work, we use the tag disfluency for words which are only partly uttered, often termed word fragments. For example, instead of saying interesting, a speaker says interes−.

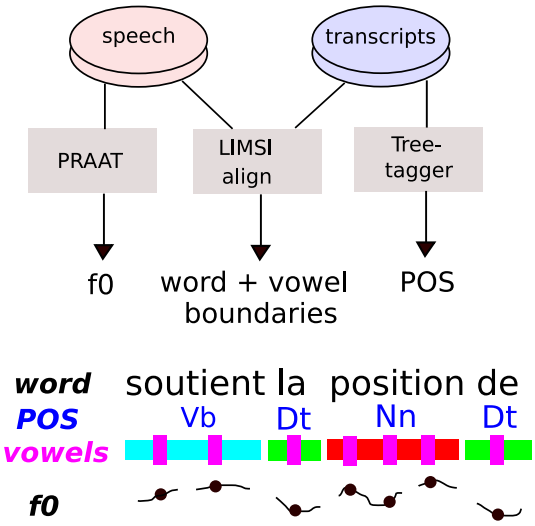


Figure 5.1: Automatic processing steps and annotation levels: each vowel is tagged by an average f_0 value and its duration, by its rank within the word, by lexical and POS information.

expect word-final schwas behave differently than word-final full vowels. For example, the word *reste* ("rest") with pronunciation [ʁɛst] was of syllable length 1 without word-final schwa, and was tagged as belonging to the *syllable length class* 1_0 (1 is for the full syllable count and 0 is for no word-final schwa). The same word pronounced with word-final schwa [ʁɛstə] goes to the *syllable length class* 1_1 (cf. *syll.class* in Table 5.1).

Words with 0 word syllable length according to the adopted representation (class `0_0`), are small function words with elided mute-e (schwa), either on the graphemic level (`l'` pronounced as /l/) or at the aligned pronunciation level (`le` pronounced as [l]). Monosyllabic words (class `1_0`) were the most frequent and word frequency then decreases with word syllabic length. Words of the same syllable class (class `n_s`: `n` full vowels; `s`: with/without word-final schwa) are merged to compute average f_0, duration, and intensity profiles.

f_0 values and profiles: f_0 profiles were computed for each syllabic word class (making use of all words belonging to *syll.class* n_s of Table 5.1) as the sequence of average f_0 values of vowels of rank k for all syllable ranks. To compute these average f_0 profiles, only vowels with voicing ratios over 70% (cf. 4.2.2) were used. The voicing ratio of a vowel is defined as the ratio of voiced frames over the total number of frames (100% for fully voiced vowels, 0% if there were no voiced frames in a vowel segment). Voicing ratio is expected to be close to 100% for all vowels. Lower rates may be due to pronunciation variants (devoiced vowels due to context, production irregularities with for example a glottal stop at the beginning or the end of the vowel, or a vowel uttered with a creaky or pathological voice) including vowel elision. Low voicing ratios may thus be indicative of automatic alignment and/or f_0 detection problems. This simple vowel selection criterion of voicing ratio over 70% resulted in a rejection rate of about 10% for the journalistic ESTER corpus and of 30% for the spontaneous PFC corpus. The discrepancy in rejection rates between prepared and spontaneous corpora suggests that there might be major changes in the acoustic realizations of these different speaking styles.

Only words with all their vowels passing the voicing criterion >70% were kept for further investigations. This selection aimed at reducing the impact of erroneous measurements, due to combined alignment and/or f_0 extraction errors. For each vowel, a mean f_0 value was computed over all voiced frames of the vocalic segment (different ways of computing a single f_0 value per vowel segment were experimented with: mean over all voiced frames, mean over three central frames resulting in very similar f_0 profiles). The f_0 values in Hz were converted to semitones (ST), with 120 Hz as baseline frequency for male voices (often considered as average male f_0 [Léon, 2007, p.51]), which was actually close to the average f_0 of our data.

Each word from the prepared corpora including orthographic/phonemic transcribed pronunciation was also tagged with a corresponding POS using WMATCH/TREETAGGER (as in Figure 5.1). Each vowel of the corpus was annotated with its mean f_0 in ST, its word class and its syllable rank within the word class and its POS. The *f_0 profile* of a word was then defined as a schematic f_0 contour connecting the f_0 values of the different vowels of increasing rank. Similarly, for a given word class (e.g. *syll.class* in Table 5.1), its average *f_0 profile* was defined as connecting average f_0 values of increasing rank, where the average f_0 value of a given syllabic rank was computed over all the vowels of this rank in the considered word syllable class. For example, given the `2_0` class of bisyllabic words without final schwa, the corresponding average f_0 profile was computed as the contour connecting the average f_0 value of the rank 1 vowels (first syllable) to the average f_0 value of the rank 2 vowels.

Intensity values and profiles: As explained above for f_0 values and profiles, intensity profiles are computed with the same voicing criterion (>70%). We limited intensity values to frames where f_0 was defined (with the idea that these frames correspond to most reliable intensity values).

A comparison of these prosodic parameters (f_0, duration and intensity) is presented for `grammatical` and `lexical` words in section 5.2, and for `nouns` and `noun phrases` in section 5.3.

5.2 Lexical *versus* grammatical words

As a first investigation, we computed the average f_0 profiles for our different syllabic length word classes. According to earlier studies on French prosody [Vaissière, 1991, p.112], grammatical (function) words are uttered on a lower register than lexical (content) words. Different average profiles may thus be expected for lexical words as compared to grammatical words. For each syllabic length word class, we separated **lexical** words from **grammatical** words. The words were thus divided into two categories according to their POS-tags:
grammatical words: article, conjunction, preposition, pronoun, etc.
lexical words: noun, name, adjective, adverb, verb, etc.

Table 5.2(a) shows the quantitative description of lexical words for each corpus and Table 5.2(b) addresses grammatical words. We applied a minimum word frequency criterion (#word tokens >100) in order to ensure representative data to estimate average profiles. Due to this criterion, all profiles are limited to at most 4-syllabic lexical words. From the quantitative data of Tables 5.2(a) and 5.2(b), several observations can be drawn:

(i) Journalistic broadcast speech (ESTER) includes higher rates of polysyllabic content words than PFC.

(ii) Both corpora include a low percentage of words with word-final schwas. Average profiles can be produced only for shorter words (up to 3 syllables for ESTER; at most 2 syllables for PFC).

(iii) Grammatical words tend to be short and average profiles can be produced only for shorter words (up to bisyllabic for ESTER; monosyllabic for PFC).

Table 5.2: Quantitative ESTER and PFC corpora descriptions of lexical (left (a)) and grammatical (right (b)) words with regard to word tokens of syllabic length `n` from 1 to 4 for words without final-schwa and from 0 to 1 for words with final-schwa. `n_s` states `n`: the number of full syllables; `s`: absence (0)/presence (1) of final schwa.

(a) Quantitative ESTER and PFC corpora description of lexical words.

Lexical	n_s	ESTER	PFC
without final-schwa	1_0	30888	29583
	2_0	33715	18391
	3_0	15960	4854
	4_0	6036	1390
with final-schwa	1_1	2755	1147
	2_1	1999	691
	3_1	693	206

(b) Quantitative ESTER and PFC corpora description of grammatical words.

Grammatical	n_s	ESTER	PFC
without final-schwa	1_0	40919	32382
	2_0	2237	1791
with final-schwa	0_1	11795	4949
	1_1	1158	496

After categorizing words into two groups (lexical and grammatical words), we computed and schematized the average f_0, duration and intensity of lexical and grammatical words. These average profiles may then be checked for the known f_0 rise and longer duration on "final" syllables in French [Rietveld, 1980; Di Cristo, 1998; Welby, 2006; Welby, 2007] as well as for comparisons between speaking styles. We want to add a caveat at this point: all word endings do not correspond to "prosodic unit" endings and thus need not show specific duration lengthening or particular f_0 and intensity patterns. However, the probability of word-final syllables coinciding with prosodic unit endings grows with increasing word syllabic length,

As French tends to produce word final accentuation, we chose to adapt the graphical displays of our profiles by producing right-justified curves: the values corresponding to the last syllables (of any `m`-syllabic word class) are displayed together at the final `n`-th position of the longest `n`-syllabic words. Preceding syllables are noted (on the X-axis) as `n-1` for penultimate syllables, `n-2` may indicate the initial syllable of bisyllabic words or more generally the antepenultimate syllable of longer words... The prosodic profiles are then compared across the two speaking styles of corpora (**prepared** and **spontaneous**). The profiles of words ending by **final-schwa** or not were also investigated to examine the impact of word-final schwa on the prosodic profile patterns according to f_0, duration, and intensity respectively.

5.2.1 f_0 profiles

First, we will present average f_0 profiles of lexical words for the different syllabic lengths, starting with the journalistic speech, before switching to the spontaneous speech data. Grammatical words

are shortly addressed for comparison. The latter are not supposed to be in "prosodic unit" final position – they tend to be in initial (e.g. determiners and prepositions) or central positions – and should hence produce (unmarked) profiles without specific prosodic word ending marks.

Lexical words

First, we present the results from the prepared journalistic broadcast speech of the **ESTER** corpus. Among the specificities of this type of data, one may cite the richness of vocabulary with many polysyllabic words, in particular nouns and proper names; the produced speech is meant for a broad and diverse audience who doesn't share instantaneous context with the speakers (as opposed to our spontaneous speech corpus of face-to-face conversations of PFC). The information content is high and the acoustic signal is supposed to contain many phonetic and prosodic cues to enable distant (in place and time) listeners to correctly decode such a difficult, highly informative acoustic signal.

Figure 5.2 exhibits mean f_0 profiles of the ESTER corpus **without** final-schwa with all profiles right-justified on the n-th syllabic position.

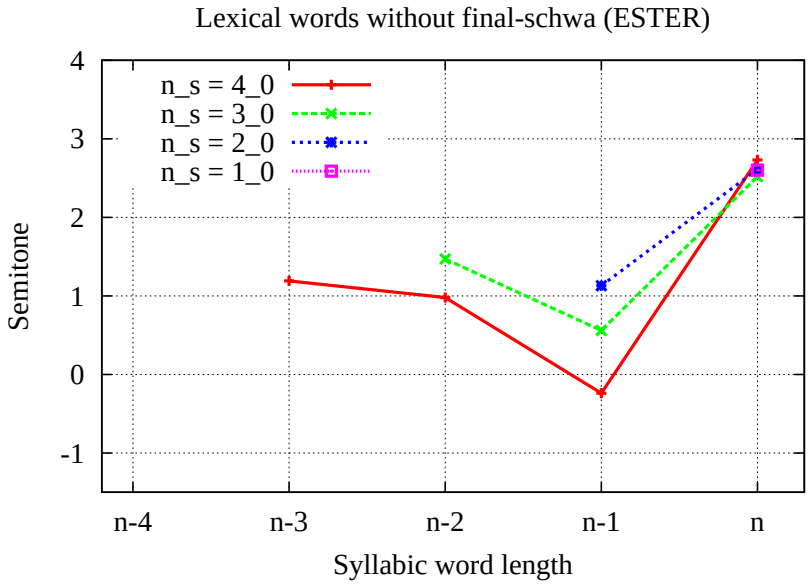


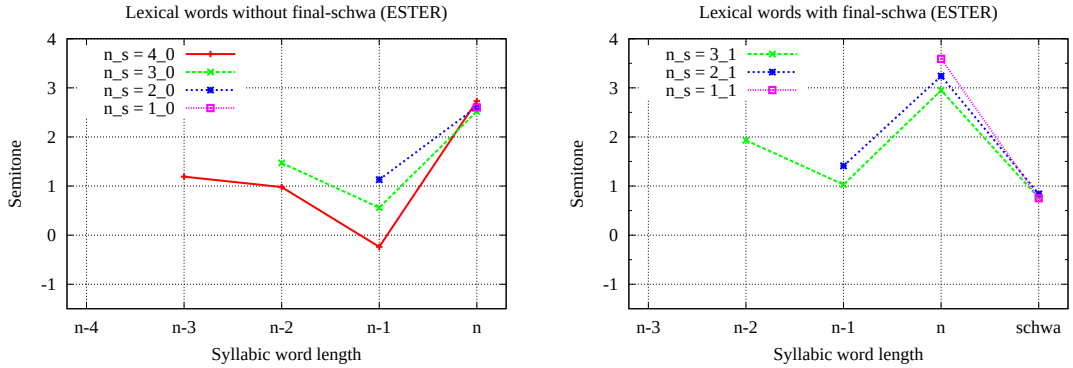
Figure 5.2: Lexical word f_0 profiles of n_0 word classes (without final-schwa) from prepared journalistic speech (ESTER corpus).

Lexical words of the ESTER corpus show the following average characteristics (Figure 5.2):

(i) A final f_0 rise for all n_0 word classes (word length from 1 to 4);
(ii) The initial syllable tends to have the highest f_0 when considering all but final syllables.

These observations are consistent with our general knowledge of French prosody. We may thus consider that the proposed method is sound to look for major prosodic specificities of French. However, fine prosodic detail may require a better human expert control of the material as well as more advanced tuning and measuring techniques. We can further add the following observations:

(iii) A minimum f_0 on penultimate syllables for all n_0 word classes;
(iv) A maximum Δf_0 (rise) between penultimate and final syllables;
(v) The Δf_0 (rise) between penultimate and final syllables tends to grow with n (syllabic word length). This may be linked to the previous caveat: the probability of word-final syllables actually being prosodic unit endings grows with increasing word syllabic length.



(a) Lexical words **without** final-schwa for ESTER corpus. (b) Lexical words **with** final-schwa for ESTER corpus.

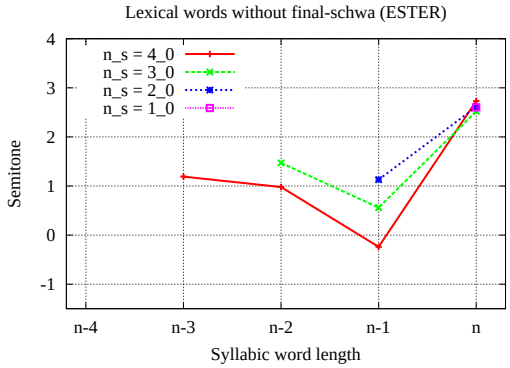
Figure 5.3: Lexical word f_0 profiles with/without final-schwa of ESTER corpus.

In Figure 5.3, average f_0 profiles are added on the right for lexical words with word-final schwas. It is noticeable that the two sets of f_0 profile patterns remain quite similar up to the final syllabic length `n` in spite of the presence (or absence) of word-final-schwas and the very different number of samples per populations. However, a word-final schwa entails an important average f_0 drop on the very final schwa vowel. Hence, our measurements for lexical words with final schwas may be summarized as follows (Figure 5.3 right):
(i) A maximal Δf_0 (drop) between the final syllable `n` and the following final schwa corresponding to 2–3 semitones (ST);
(ii) Average f_0 profiles with a word final-schwa are globally in a slightly higher register (higher mean f_0 values) as compared to the average profiles of lexical words without final schwa.

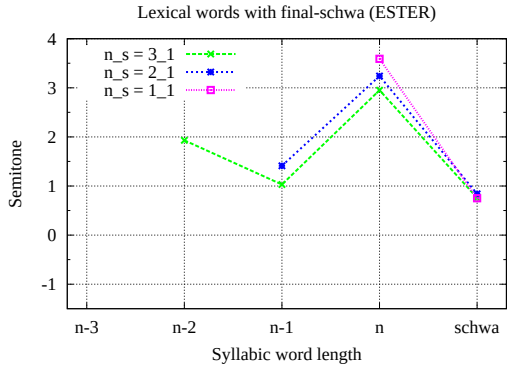
Common points from both figures can be summarized in the following (Figure 5.3 left and right):

(i) The mean f_0 is highest for the word-final syllables (reaching about 2.8 ST for lexical words without schwa and about 3.2 ST for lexical words with final schwa);
(ii) The f_0 difference between final `n` and penultimate (`n-1`) consecutive vowels tends to increase with word syllabic length;
(iii) Mean monosyllabic f_0 is at least as high as that of the final syllable of longer syllabic words;
(iv) Initial accentuation remains relatively weak on mean f_0 contours.

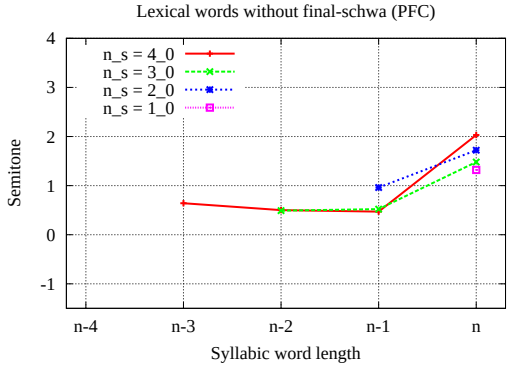
We now turn to spontaneous speech with the face-to-face conversations of the **PFC** corpus to check whether similar average f_0 profiles are achieved for this different speaking style. In general, spontaneous speech contains a relatively limited vocabulary, as also suggested by the low proportion of long polysyllabic words in our quantitative description of the corpus vocabularies (cf. Table 5.1). Our hypothesis of more shared context between speakers in these conversations, as well as the



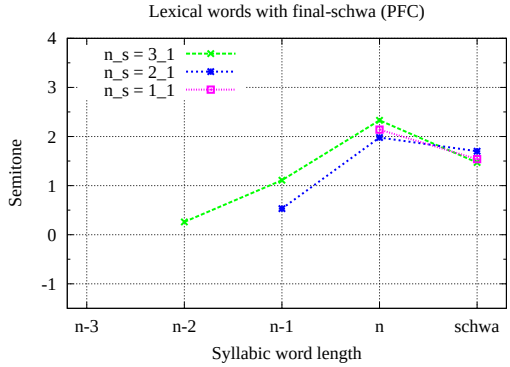
(a) Lexical words **without** final-schwa for ESTER corpus.



(b) Lexical words **with** final-schwa for ESTER corpus.



(c) Lexical words **without** final-schwa for PFC corpus.



(d) Lexical words **with** final-schwa for PFC corpus.

Figure 5.4: Lexical word f_0 profiles of ESTER corpus (**top**) and PFC corpus (**bottom**).

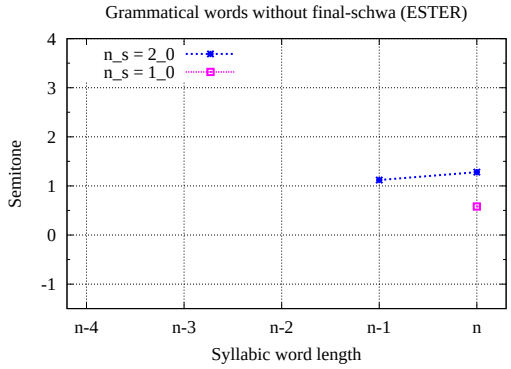
evidence of degraded ASR results for this kind of speech suggest that spontaneous speech may include weaker acoustic-prosodic cues as compared to professional journalistic speech.

To facilitate comparisons, we recall in Figure 5.4 the already presented prepared speech profiles of the ESTER data, before introducing the new spontaneous speech PFC profiles. Concerning the spontaneous speech (PFC) data, Figure 5.4(c) displays average f_0 profiles for lexical word classes **without** final-schwa and Figure 5.4(d) those **with** final-schwa. At a first glance, it can be observed that the average f_0 profiles from the PFC corpus are flattened as compared to those of the ESTER corpus. Nonetheless, the f_0 profiles exhibit some similar patterns as the ESTER profiles, i.e. highest mean f_0 values for final syllable `n`, an f_0 drop on final schwas, and a tendency of highest f_0 values on initial syllables for all but final syllables, at least for the word classes without final schwa. The only relevant class with final schwa, the 3_1 class contains only 206 tokens. It is interesting to note that the f_0 drop on the word-final schwa is much weaker for the PFC corpus. Given these results for spontaneous speech, one may question whether the flattened profiles are really due to limited f_0 excursions on a majority of spontaneously uttered words or whether this result is due to a method bias. In particular, the averaging within a given class of potentially very different profile patterns may entail flattened average profiles (individual f_0 profiles might present important variations with differently placed rises and drops across tokens

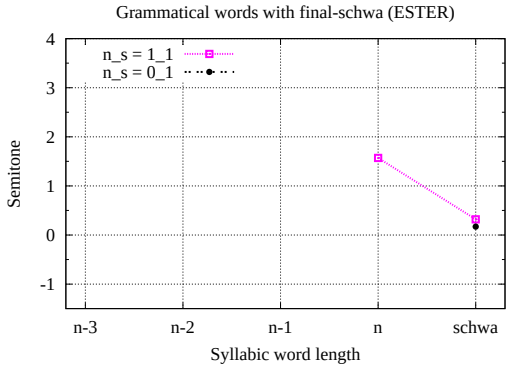
of a given word class). This hypothesis is especially interesting, as our PFC data include different regional accents from North and South of France. The question of a potential method bias is addressed in a later subsection dealing with intervocalic measurements (see subsection 5.3.4) and we may anticipate that the future results go in favor of effectively flattened profiles for larger proportions of spontaneous speech as compared to journalistic speech.

To summarize, we may say that the obtained f_0 profiles from the two speech corpora representing different speaking styles have confirmed the intonational patterns for French already known in the literature (i.e. tendency of f_0 rise on final syllables in lexical word) which were often investigated using small corpora of read speech or artifact corpora with a small number of utterances. A more detailed and less (or not yet) described finding concerns the correlation between word final schwa and a slight but global f_0 profile increase. We consider this as an interesting result of our proposed method and to be further checked in the future on larger data sets. We will now rapidly turn to f_0 profiles of grammatical words.

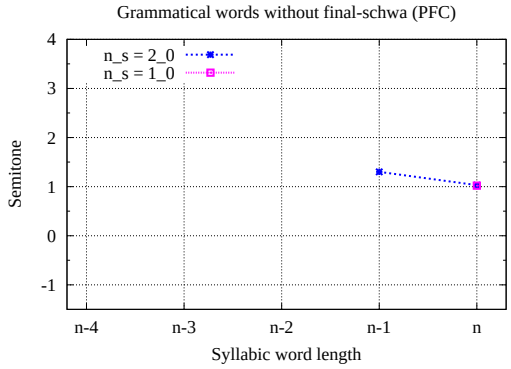
Grammatical words



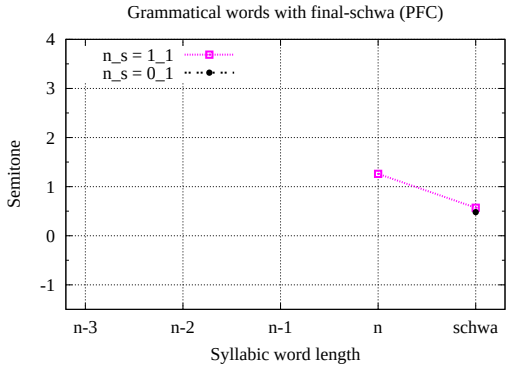
(a) Grammatical words **without** final-schwa for ESTER.



(b) Grammatical words **with** final-schwa for ESTER.



(c) Grammatical words **without** final-schwa for PFC.



(d) Grammatical words **with** final-schwa for PFC.

Figure 5.5: Grammatical word f_0 profiles of ESTER corpus (**top**) and PFC corpus (**bottom**).

Grammatical word profiles are of limited interest as they involve only short words of maximum two syllables. Variations with syllable rank are limited to simple rises or drops (with a stable option in between) for bisyllabic words. When comparing to lexical words, one may notice, as expected, that the average f_0 values of **grammatical words** (Figure 5.5) of word-final syllables are lower than those of lexical words (Figure 5.4) for both ESTER and PFC corpora. Furthermore, f_0 profiles are fairly similar between the ESTER corpus and the PFC corpus for both words without and with final-schwa. Vaissière claimed in [1991, p.112] that lexical words are uttered on the high register and grammatical words on the low register. Thus, average f_0 contours of grammatical words feature flatter curves than the lexical word ones especially for words **without** final-schwa (two left sub-figures). Regarding to words **with** final-schwa, Figure 5.5(b) of ESTER shows 1.2 ST of difference between final-schwa and final syllable (n) and this difference is smaller than that for lexical words which tends to be more than 2 ST (Figure 5.4(b)). For the PFC corpus (Figure 5.5(d)), the average difference is about 0.7 ST between final-schwa and final syllable as for lexical word f_0 profiles (Figure 5.4(d)). Here also, we can notice the effect of different speaking styles that spontaneous speech produces flatter profiles. Major observations for grammatical words may be summarized as follows (see Table 5.5):

(i) The average f_0 for word-final syllables is close to 1 semitone (rather 3 and 2 semitones for ESTER and PFC lexical words respectively);
(ii) No (strong) rise may be observed between penultimate and final syllables;
(iii) The average f_0 drops on word-final schwas (and is actually close to 0 semitones).

We have presented average lexical f_0 profiles in terms of: lexical and grammatical words, absence or presence of word-final schwa, prepared and spontaneous speaking styles. Concerning **lexical and grammatical** words, lexical word classes include more polysyllabic words whereas grammatical word classes are likely to be mono- or bisyllabic. For lexical words, average f_0 values of final syllables n are highest in comparison to preceding syllables and to word-final schwas. The average f_0 of the final syllable of monosyllabic lexical words tends to be as high as those of the other n-syllabic word-final syllables. From a purely probabilistic point of view, this might be against intuition, as monosyllabic words are less prone to be prosodic-word final and thus less prone to benefit from a final f_0 rise. This raises interesting questions of specific prosodic realizations of mono-syllabic words to be investigated in future studies. Further, we could observe an increase of Δf_0 between penultimate and final syllables with increasing syllabic word length. As expected, for grammatical words, f_0 profiles at final syllable n are not as high as for lexical words. Also, f_0 profiles of bisyllabic grammatical words are flattened as compared to lexical words. Concerning the presence or absence of **word-final schwas**, the comparisons reveal slightly higher average f_0 profiles for words with word-final schwas for both corpora and for both lexical and grammatical words. The profiles for different **speaking styles** suggest that spontaneous speech tends to have flatter f_0 realizations for lexical words. This may indicate that less acoustic-prosodic cues can be found in spontaneous speech which in turn might contribute to increased ASR difficulties on spontaneous speech.

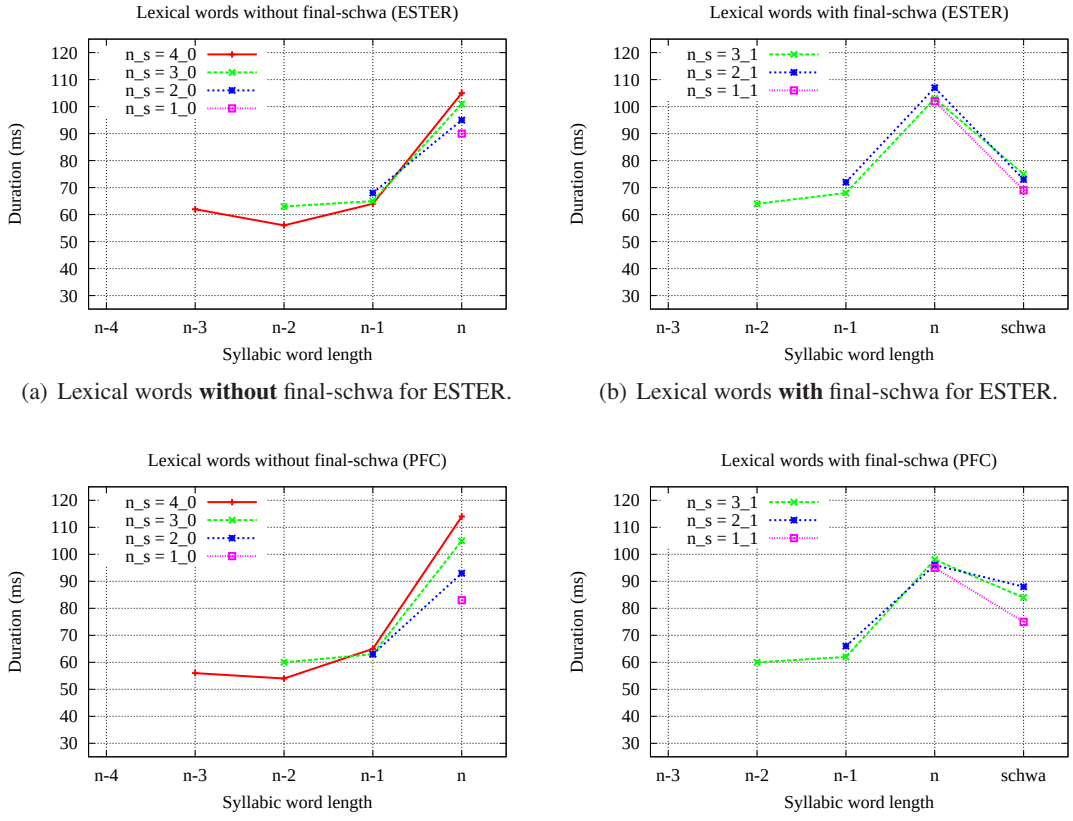
5.2.2 Duration profiles

As for word-final f_0 rise, lengthened durations on lexical word endings are among the major prosodic characteristics of French [Delattre, 1965; Delattre, 1966a] and longer durations could be

measured for French word-final syllables [Adda-Decker *et al.*, 2008] from large corpora. Our duration profiles are thus expected to indicate a duration increase for the word-final syllable n. Will duration produce profiles similar to those obtained for f_0 measurements or are there very different tendencies? In particular, one may wonder whether f_0 rises correspond to duration lengthening and, in return, f_0 drops to duration cuts, or whether any major f_0 variation entails duration lengthening no matter whether f_0 rises or drops, or whether both are not correlated at all. In the following, we often refer to average syllable durations, whereas we actually measured average vowel durations and this may be considered as a misuse of language. However, as vowel and syllable durations are strongly correlated, we have not systematically corrected syllable duration by vowel duration.

Firstly, the impact of lexical words in the ESTER corpus will be discussed before presenting the PFC corpus results. Then we will summarize the duration comparison of two speaking style.

Lexical words



(a) Lexical words **without** final-schwa for ESTER.

(b) Lexical words **with** final-schwa for ESTER.

(c) Lexical words **without** final-schwa for PFC.

(d) Lexical words **with** final-schwa for PFC.

Figure 5.6: Vocalic duration profiles for lexical words of ESTER corpus (**top**) and PFC corpus (**bottom**).

Figure 5.6 presents the average vocalic duration profiles of lexical words without (left)/with (right) final-schwa of ESTER corpus (top) and PFC corpus (bottom).

Firstly we will discuss the duration profiles of the **ESTER** corpus. The two top sub-figures in Figure 5.6 highlight the following specificities:

(i) On average, the final syllable (vowel) n is much longer than the preceding syllables (close to 50% relative: from about 60 ms to about 100 ms);

(ii) Mean final syllable (vowel) durations without final schwa are comprised between 90–105 ms; those with final schwa are between 102–107 ms. As earlier observed for f_0, our results suggest that duration slightly increases for words with final schwa. This result is not completely intuitive, as the presence of a word-final schwa is *per se* a means of increasing word duration;

(iii) The final syllable (vowel) n durations vary as a function of syllabic word length and tend to increase with syllabic word length. This result is not completely intuitive either, as a higher number of syllables of a word already implies a longer duration;

(iv) The presence of word-final schwas tends to decrease the variation in word-final syllable duration;

(v) Final schwa durations drop importantly as compared to the mean durations of the full final vowels. This is the same tendency as observed earlier for the f_0 profiles;

(vi) Non-final syllables tend to have similar short durations around 60 ms with somewhat shorter durations on internal (not penultimate) syllables.

Comparing to the previously established f_0 profiles, we find some similarities here: the final rise on the word-final syllable and in case of uttered word-final schwa, a duration drop for this schwa. However, the preceding (initial and word-internal) syllables show rather flattened duration curves around 60 ms (a bit more for the final-schwa word classes). One might consider this result in favor of the isochronous syllable duration theory [Pike, 1945] for French: when measuring average syllable (vowel) durations on non word-final syllables we find very close mean values across successive syllables (vowels).

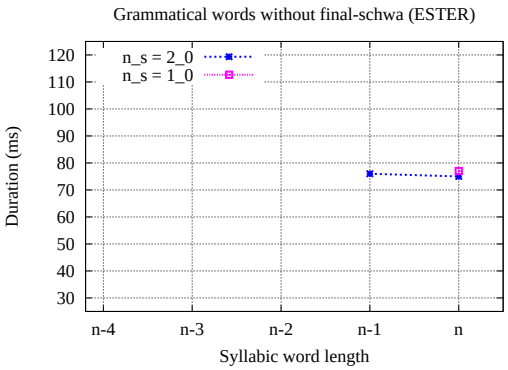
When looking at the results of our spontaneous speech data, we can observe that the average profiles are very close to those of prepared speech. Speaking style thus does not impact duration profiles as much as f_0 profiles (which significantly flattened from prepared to spontaneous speech).

For the **PFC** corpus, final syllable (vowel) n durations are in a range between 83–114 ms without final schwa and between 95–98 ms with final-schwa. Words without final schwa represent a wider range of different long durations for final syllables and durations rise with the word syllabic length. As compared to prepared speech, final schwa durations do not drop so much here and remain quite longer than penultimate or antepenultimate syllables. Recalling f_0 profiles for schwa-final spontaneous speech, a similar tendency was observed: in our data, spontaneous speech tends to keep higher f_0 values and higher durations on word-final schwas than in prepared speech.

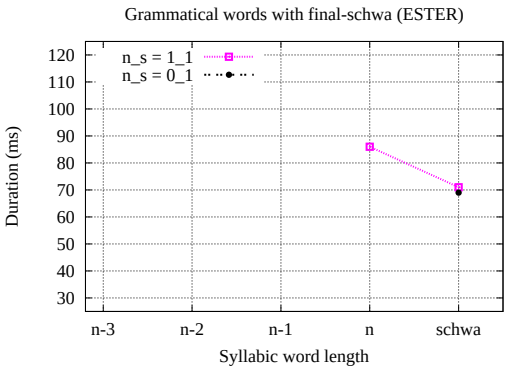
With respect to our earlier questions of correlation between f_0 and duration rises and drops, we may observe that major f_0 rises entail major duration lengthening, however f_0 drops may or may not entail duration cuts: penultimate syllables of prepared speech feature a major f_0 drop as compared to their preceding syllable, whereas for durations, the penultimate syllables tend to be slightly longer than the preceding ones.

Grammatical words

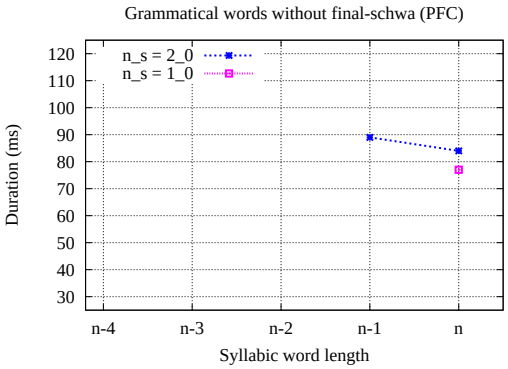
Grammatical words are known to be often underarticulated and may even disappear. Thus, they are often involved in ASR recognition errors, even though they are highly frequent and observed in a large variety of contexts to be well predicted by the language models. Recall that examined grammatical words are mono- or bisyllabic.



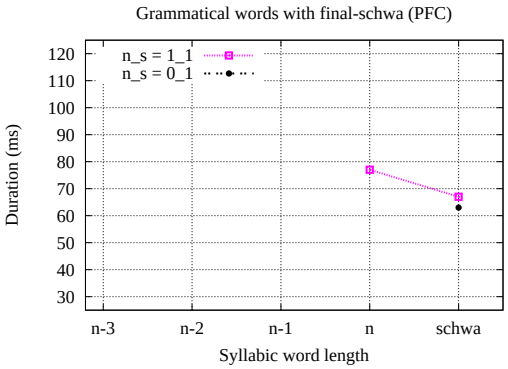
(a) Grammatical words **without** final-schwa for ESTER.



(b) Grammatical words **with** final-schwa for ESTER.



(c) Grammatical words **without** final-schwa for PFC.



(d) Grammatical words **with** final-schwa for PFC.

Figure 5.7: Vocalic duration profiles for grammatical words of ESTER corpus (**top**) and PFC corpus (**bottom**).

Figure 5.7 presents grammatical word vocalic duration profiles without/with final-schwa. Here, we do not observe important final duration increases on syllable `n`, but relatively stable durations with even, a slight drop from the first to the second syllable for bisyllabic function words. These results are consistent with the hypothesis of observing grammatical words at the beginning or in the middle of prosodic units, but not in final prosodic word positions. One might expect that average durations would be close to those of non-final syllables of lexical words. However, results show that syllable `n` durations of grammatical words are not as short as non-final syllables of lexical words (cf. Figure 5.6, most durations are about 70 ms). One reason why final vowels are not so short as expected is that the grammatical words of this study include a large proportion of nasal vowels[2]. As nasal vowels are at least 20 ms longer than oral vowels, these results

[2]like `dans` /dɑ̃/, `en` /ɑ̃/, `sans` /sɑ̃/, `on` /ɔ̃/, etc.

in somewhat increased average final syllabic durations. A further explanation may be due to hesitation phenomena which may be hided in a grammatical word lengthening, especially for spontaneous speech. Final-schwa duration tends to be shorter than the full vowel durations (Figures 5.7(b) and 5.7(d)). However, only few occurrences were available here and results are given for completeness (but are not very reliable).

To summarize the present subsection, we may say that duration profiles were presented for lexical/grammatical words, without/with final-schwa, and speaking styles (prepared/spontaneous). **Lexical and grammatical** word comparisons showed that lexical words yield an important final syllabic duration increase as opposed to grammatical words for which this tendency was not observed. The preceding (initial and word-internal) syllables of lexical words show rather flattened duration curves around 60 ms (a bit more for the final-schwa word classes). One might consider this result in favor of the isochronous syllable duration theory [Pike, 1945] for French. An interesting result, contrasting with the f_0 profiles, is that duration profiles stay very similar for both speaking styles (prepared and spontaneous). Average durations of final syllables n without final-schwa (for lexical words) increase with word syllabic length, whereas preceding syllable durations tend to slightly decrease. The comparison of lexical words finishing or not by **word-final schwa** show that the duration variation range of final syllables n is large without final schwa (and larger for spontaneous speech than for prepared speech), whereas word final schwas tend to normalize the mean duration of the last full vowel. Final-schwa durations in PFC corpus are longer than the ESTER corpus in which final-schwa durations were as long as the penultimate or antepenultimate syllables. These results also suggest that spontaneous speech yields more variation in speech rhythm.

5.2.3 Intensity profiles

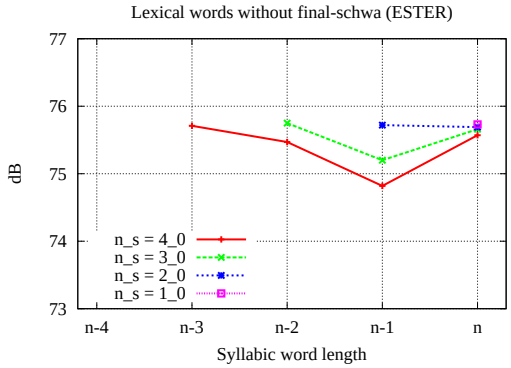
The average of f_0 and duration profiles were presented above and the measured profiles revealed final syllable accentuation for these two prosodic parameters. Here, a third prosodic parameter "intensity" will be investigated. According to [Delattre, 1966a], strong intensity does not appear at the end of a word in French and even weaker intensity is found at the end of a phrase. We may thus expect some kind of declension line on measured intensity as air pressure globally decreases along breath groups. As before, a question may be whether intensity profiles are similar to those of f_0 where important increases on final syllables could be observed (even though declension lines are also a well known topic for f_0).

Here we will establish intensity profiles from large speech corpora and compare, as in the earlier subsections, lexical and grammatical words, words without and with final-schwa, and prepared and conversational speaking styles.

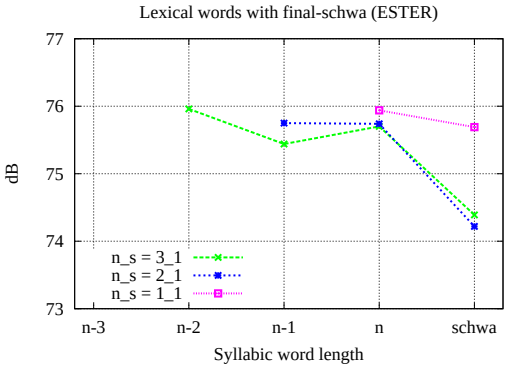
The raw intensity data was averaged. Lexical word intensity profiles are illustrated in Figure 5.8 and grammatical word profiles in Figure 5.9.

Lexical words

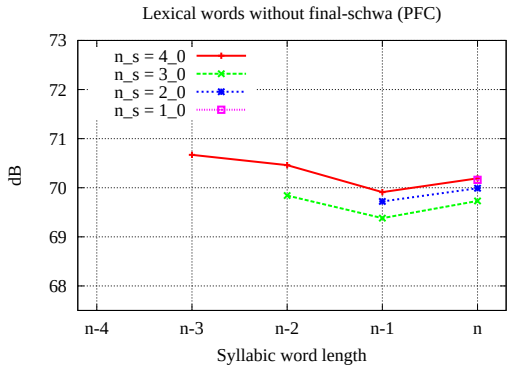
The **ESTER** corpus profiles are illustrated in the two top sub-figures of Figure 5.8 without final-schwa (left) and with final-schwa (right). It can be observed that intensity profiles remain relatively



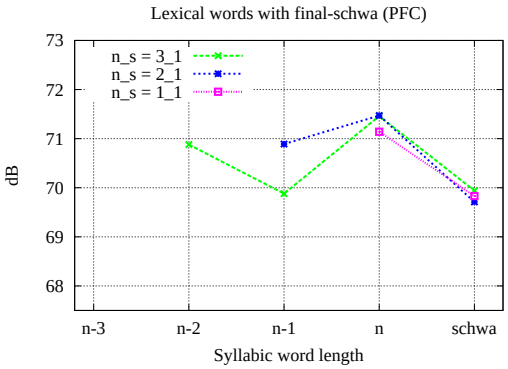
(a) Lexical words **without** final-schwa for ESTER.



(b) Lexical words **with** final-schwa for ESTER.



(c) Lexical words **without** final-schwa for PFC.



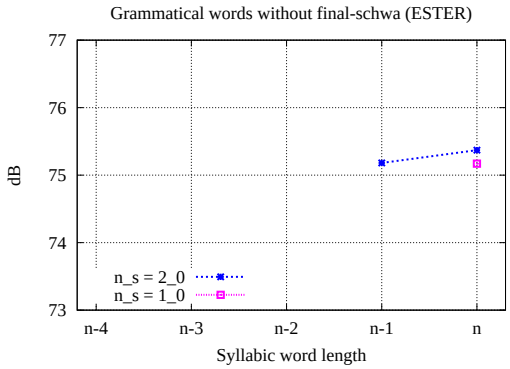
(d) Lexical words **with** final-schwa for PFC.

Figure 5.8: Vocalic intensity profiles for lexical words of ESTER corpus (**top**) and PFC corpus (**bottom**).

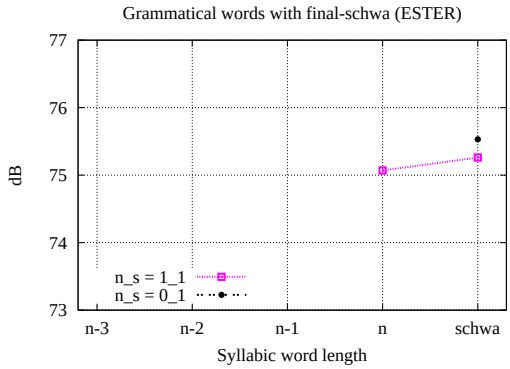
even as compared to the f_0 profiles. A relatively strong intensity drop is observed on final schwa vowels for polysyllabic words. Overall, the intensity profile shapes are similar to those of the f_0 profile patterns. However, intensity values of final syllables `n` are not highest, but at best as much as high as first syllables of a word. Polysyllabic word profiles show an intensity rise between the penultimate syllables (`n-1`) and final (`n`) syllables.

Concerning spontaneous speech with the **PFC** corpus, we may observe more flattened profiles (as earlier for the f_0 profiles). There is a very slight intensity increase on the final syllables which remain below initial syllable intensity (for words without schwa). A higher intensity rise is measured for words with final-schwa and, as already noted for f_0 and duration, the average profiles are slightly raised in the presence of word-final schwa (as compared to the left figures without final schwa). It is worth mentioning that the profile pattern corresponding to trisyllabic lexical words in presence of the word final-schwa is very similar to the f_0 profiles obtained for the same lexical word category taken from the ESTER corpus (cf. Figure 5.4(b)). To sum up, one may notice that the overall intensity declension observed for increasing word syllables is stopped at the final syllable for both speaking styles. Final syllable intensity values increase to become approximately the same as those of the first syllables. Low intensity values in word-final schwa are observed for both corpora.

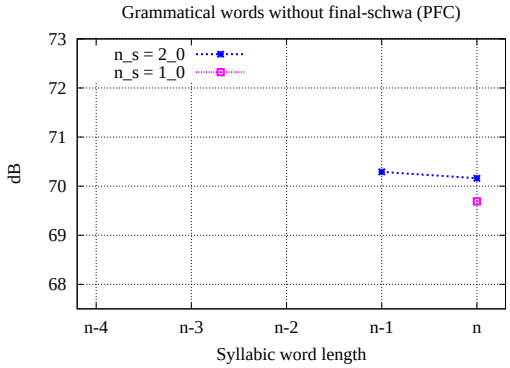
Grammatical words



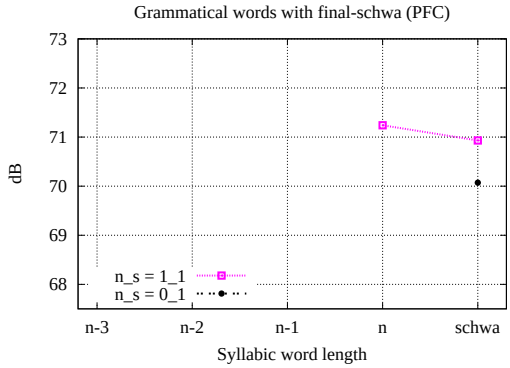
(a) Grammatical words **without** final-schwa for ESTER.



(b) Grammatical words **with** final-schwa for ESTER.



(c) Grammatical words **without** final-schwa for PFC.



(d) Grammatical words **with** final-schwa for PFC.

Figure 5.9: Vocalic intensity profiles for grammatical words of ESTER corpus (**top**) and PFC corpus (**bottom**).

Grammatical word intensity profiles are presented in Figure 5.9. We will only shortly comment on these profiles. In particular, for the final-schwa condition, the sample population remains small:

(i) Intensity values are relatively stable across the different conditions;

(ii) Average intensity values of grammatical words tend to be lower than those of word-initial and final syllables of lexical words;

(iii) No intensity drop is observed on grammatical words' final schwas (as opposed to lexical words).

We can summarize intensity results from both lexical/grammatical words and prepared/spontaneous speaking styles in that intensity values of final syllables `n` are almost the same as those of the first syllables. Intensity profiles have shapes which are relatively similar to those of f_0, but they differ nonetheless from both f_0 and duration profiles in that the latter ones have clearly marked final syllables (highest and longest). As cited in the literature, intensity is not a major parameter for final accentuation of a word in French.

5.2.4 Short *versus* long duration impact

In this section, we would like to investigate the impact of overall word duration on f_0 profiles. The question is whether a change in overall duration entails a rise or a drop of f_0 profiles. Short duration words may be due to a speaker who on purpose accelerates his speaking rate while keeping a clear and distinct articulation. Our corpora are not controlled for such speakers. We make another hypothesis for locally short duration words (in relation with ASR recognition errors). Short duration segments often result from automatic speech alignment during which canonical pronunciations of the pronunciation dictionary have to be matched to temporally reduced acoustic forms. These are not necessarily articulated very fast, but with a smaller number of segments or even syllables. Our hypothesis is here that higher speaking rates of this type or shorter overall word durations may be correlated with lower f_0 and possibly with less Δf_0. For this special focus on duration, we limited our investigations to lexical words without final-schwa as this condition provides both the most data and polysyllabic words. Lexical words were separated into two groups: **fast** rate words and **slow** rate words.

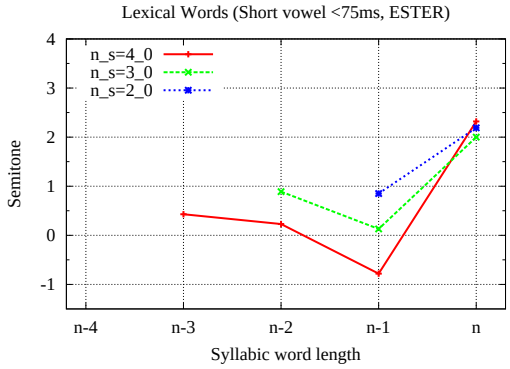
Fast/slow rate words: The selection of fast and slow rate words was carried out by filtering vocalic durations on **all vowels but the final vowel** of a word (as final vowels tend to be longer even in fast speech). Words, for which all non-final vocalic durations remained below 75 ms, were considered as (locally) fast rate words. The remaining ones correspond to the (locally) slow rate speech. For example, if a trisyllabic word has: first vowel 80 ms; second vowel 70 ms; final vowel 100 ms; then this word is in the slow rate group because the first vowel duration is above 75 ms. The empirically fixed threshold to separate between slow and fast is not critical. Our aim was to have more or less balanced subsets for polysyllabic words (occurring less often), which entails more words in the fast subset for bisyllabic words. This way, distinct subsets of lexical words were defined and average f_0 profiles were calculated for fast and slow rate words.

Table 5.3: Proportions of **fast** and **slow** rate for lexical words and all vowels in each corpus, ESTER and PFC.

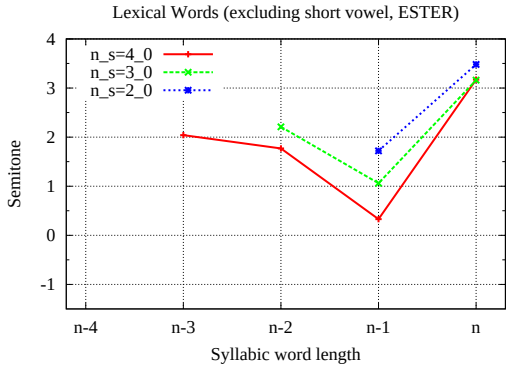
	ESTER		PFC	
Lexical words	Fast <75ms	Slow	Fast <75ms	Slow
bisyllabic	68%	32%	72%	28%
trisyllabic	56%	44%	56%	44%
4-syllabic	52%	48%	51%	49%
All	<75ms	>75ms	<75ms	>75ms
Vowels	60%	40%	63.5%	36.5%
Oral	67%	33%	67%	33%
Nasal	26%	74%	45%	55%

Proportions of investigated lexical words between fast rate words and slow rate words for the ESTER corpus and the PFC corpus are given in Table 5.3. Table 5.3 also presents the proportions of all vowels in the corpora to compare with. It is interesting to highlight that with the same threshold quite similar proportions of fast/slow rate categories are achieved for both corpora. The

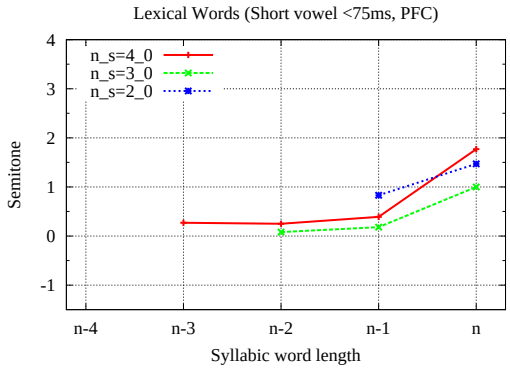
biggest difference is for bisyllabic lexical words for which the PFC corpus has 4% more in the fast rate group as compared to the ESTER corpus. We are aware that our criterion of 75 ms is not best tuned to handle any vowel types and that for example nasal vowels would require a different threshold than oral vowels. If we look into all vowels of the two corpora (not only lexical word ones), 60% of vowels belong to lower than 75ms and 40% of vowels to slow rate speech for the ESTER corpus, 63.5% to fast speech rate and 36.5% to slow rate for the PFC corpus. As mentioned above, duration of nasal vowels is generally longer than that of oral vowels and consequently, higher proportions of nasal vowels are measured for slow rate speech. However, for spontaneous speech, there is an increase of almost 20% of nasal vowels for spontaneous speech. This important difference might be due to the difference in speaking style, people speaking less carefully in conversational speech.



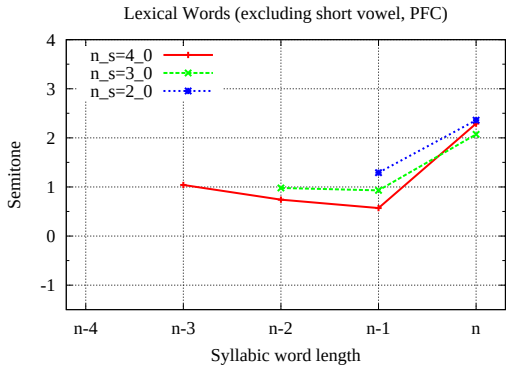
(a) Lexical words in the fast rate set for ESTER.



(b) Lexical words in the slow rate set for ESTER.



(c) Lexical words in the fast rate set for PFC.



(d) Lexical words in the slow rate set for PFC.

Figure 5.10: Mean f_0 profiles of n-syllabic lexical words (n=2–4) without final-schwa as a function of duration. **Left:** Fast rate set (<75ms, except for final vowels) **Right:** Slow rate (all except the fast rate set). **Top**: ESTER corpus. **Bottom**: PFC corpus.

Figure 5.10 shows the corresponding results. Left figures (Figures 5.10(a) and 5.10(c)) show average f_0 profiles of fast rate words (all non-final vowels are lower than 75ms). Right figures (Figures 5.10(b) and 5.10(d)) present average f_0 profiles for slow rate (long duration) words. Monosyllabic

lexical words were excluded as the fast/slow splitting criterion was applied on all but the final syllables.

Concerning the **ESTER** corpus results (top), fast rate (short duration) words (left figure) display lower f_0 profiles as compared to longer lasting words (right figure). This result corresponds to our earlier stated hypothesis: shorter words (less well articulated with shorter and even missing segments) have lower average f_0. However, despite the differences in duration, the f_0 profiles of fast (left) and slow (right) speaking rates keep the same shapes (the same also as the mean f_0 profiles, cf. Figure 5.4(a)). When moving from fast to slow rate words, the profiles are just transposed by about 1.5 semitones (higher). This result is not in line with our earlier stated second hypothesis that the overall Δf_0 might be reduced for shorter durations. Is this particular for journalistic speech with a large proportion of professional speakers?

The **PFC** corpus (bottom) with spontaneous speech features rather similar results. Profiles are raised by about 1 semitone for slow rate words. However, we can observe that fast spontaneous speech actually reduces the overall Δf_0. The characteristic minimum f_0 values on the penultimate syllables of lexical words are no longer observed here. Even though profiles changed only slightly, the observed changes are in favor of our second hypothesis. This hypothesis of a global Δf_0 reduction is thus validated for the PFC spontaneous speech corpus, but not for journalistic speech.

The results suggest that speech rate deceleration correlates with a global upward f_0 tendency of the corresponding words, especially in the case of broadcast news type.

Three prosodic parameters (f_0, duration, intensity) were investigated using average n-syllabic word profiles. Using this methodology, the impact of syllabic word length was studied with the underlying idea that word-internal syllables might be more prone to temporal reduction phenomena and pronunciation variants. The proposed profiles enable us to give a synthetic overview of what happens to f_0, duration and intensity for different syllabic positions of French words. The data was split in various subsets: lexical words, grammatical words, presence or absence of word final-schwa and different speaking styles. First of all, we presented f_0 profiles. Higher f_0 values could be measured at the final syllables of lexical words in the both corpora. The presence of word-final schwa correlates with a global f_0 increase. The spontaneous speech of the PFC corpus displayed flatter profiles than the prepared ESTER corpus. Grammatical words globally featured lower f_0 values than lexical words. Second, we conducted duration analyses. Longer final syllable durations were observed in lexical words of the two corpora. The duration variation range of final syllables was smaller with final-schwa than for words without final schwa. Spontaneous speaking style showed greater duration range variation of final syllables than prepared speech. Concerning grammatical words, longer final syllable durations were not observed in both corpora. Third, we investigated intensity. Contrary to the two preceding parameters (f_0 and duration), remarkable final intensity values could not be measured. For lexical words, most of the time final syllable intensity values were at best as high as first syllables. Much lower intensity values were noticed at final-schwa. As for grammatical words, the intensity values of final syllable `n` syllables were almost the same values as lexical word ones. This result was opposite to those from f_0 and duration where grammatical words had lower f_0 and shorter duration for final syllables as compared to lexical words. Final accentuation is thus best correlated with f_0 and duration. The study of duration impact on f_0 was investigated for lexical words without final-schwa. The data was divided into two categories (slow and fast). Slow rate categories showed higher f_0 values for

both speaking styles. Spontaneous speech further showed less variation in f_0 profiles as compared to slower spontaneous speech. Results confirm that word duration variation may entail prosodic profile variation, which in turn may be correlated with pronunciation variation.

5.3 Noun *versus* noun phrase

The previous section 5.2 presented prosodic parameter comparisons on a lexical basis. The achieved results provide typical average profiles for French words of different syllabic lengths for different speaking styles. However, prosodic parameters are particularly interesting to be studied on larger units such as phrases or prosodic words. The delimitation of larger units is not straightforward. As a step in this direction, we propose to study simple noun phrases limited to `determiner - noun` bigrams and to compare the corresponding profiles to single word noun profiles.

In the following, we focus on the comparison between **nouns** and **noun phrases** limited to the `determiner - noun` bigram to address the question of prosodic parameter profiles across word boundaries. The measurements pertain to the question whether the mean profile of an *n* length noun phrase is different from the profile of an *n* length noun. The investigated determiners are *le*, *la*, and *les* that correspond to "the" in English. The corresponding canonical pronunciations are /lə/, /la/, and /le/ or /lez/ with its liaison. This study is limited to noun words without final-schwa. Table 5.4 shows the noun and noun phrase tokens of both ESTER and PFC corpora.

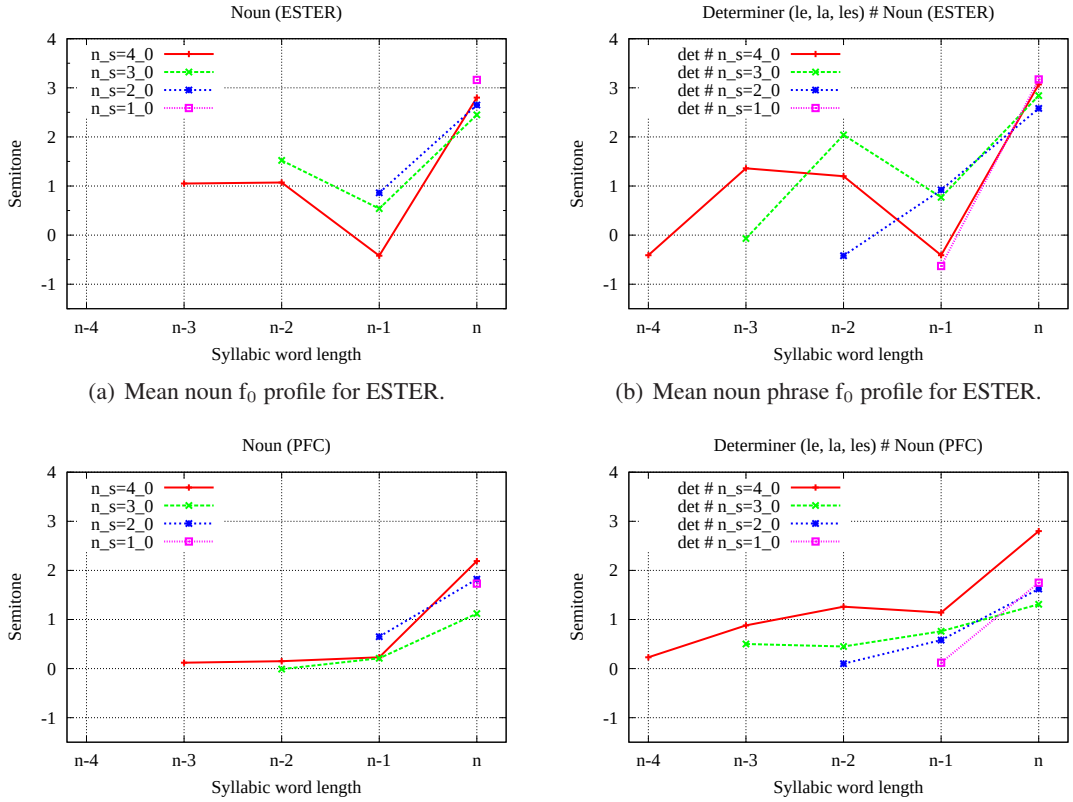
Table 5.4: Quantitative ESTER and PFC corpus description of nouns and noun phrases with regard to word tokens of word syllable length. `n` from 1 to 5. *Syll.class* n_s states n: the number of full syllables; s: absence (0)/presence (1) of final schwa.

Total # syll. n	*Syll. class* n_s	*#Noun* ESTER	PFC	*Syll. class* n_s	*#Det-Noun* ESTER	PFC
1	1_0	8222	5060	–	–	–
2	2_0	11791	4990	det#1_0	2243	969
3	3_0	5119	1330	det#2_0	2610	975
4	4_0	2923	641	det#3_0	1403	267
5	–	–	–	det#4_0	862	147

The comparison between noun and noun phrase with two speaking styles (prepared/spontaneous) will be showed according to f_0, duration and intensity respectively.

5.3.1 f_0 profiles

First of all, mean f_0 profiles were measured for noun and noun phrases, limited to the `determiner - noun` bigram. Figure 5.11 shows the mean f_0 profiles of noun words (left figures) and noun phrases (right figures). The results for the prepared ESTER corpus are illustrated at two top sub-figures and at bottom for the PFC corpus.



(a) Mean noun f_0 profile for ESTER.

(b) Mean noun phrase f_0 profile for ESTER.

(c) Mean noun f_0 profile for PFC.

(d) Mean noun phrase f_0 profile for PFC.

Figure 5.11: Mean f_0 profiles for `n`-syllabic length. **Left:** Nouns (without final schwa) **Right:** Noun phrase (`determiner-noun`).

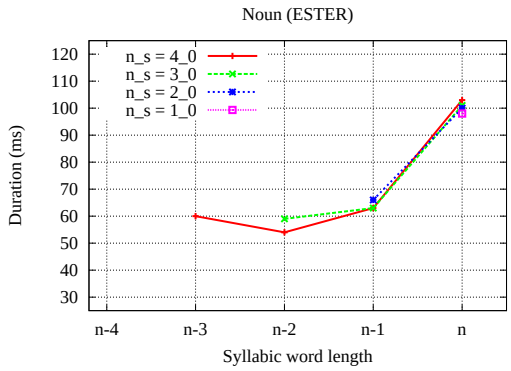
For the ESTER corpus, the left sub-figure shows the profiles for nouns whereas the right sub-figure provides the profiles of noun phrases. First, it can be seen that the overall profiles for nouns are quite similar to those of the lexical words (see Figure 5.4(a)). When comparing noun profiles to noun phrase profiles, we can observe that for a given *n* syllable length, profiles exhibit very different patterns for nouns than for noun phrases. For polysyllabic nouns, the profiles drop from the initial starting value to a minimal value on the penultimate syllable (`n-1`), to rise to an absolute maximum on the final syllable (`n`). However, polysyllabic noun phrase profiles start with a low f_0 value on the determiner (i.e. f_0 values below 0 semitones at the first point of each line). An average rise of about 2 semitones can be observed between determiner and initial noun syllable f_0 values for phrases of 4 and 5 syllables. This difference is reduced to 1.5 semitones for 3-syllabic phrases, as the profile is monotonically rising from the beginning to the end of the phrase. For monosyllabic words (phrases of 2 syllables) there is an average f_0 difference of nearly 4 ST between `determiner` and `noun`. The provided profiles show that journalistic speech contains important prosodic cues to distinguish between n-syllabic nouns and noun phrases. Let us now turn to the spontaneous speech of PFC.

The average f_0 profiles for spontaneous speech of the PFC corpus are shown in the two bottom figures): `noun` profiles (left Figure 5.11(c)) and `noun` phrase profiles (right Figure 5.11(d)). As for the ESTER corpus, the profiles of the PFC nouns are very similar to those observed globally

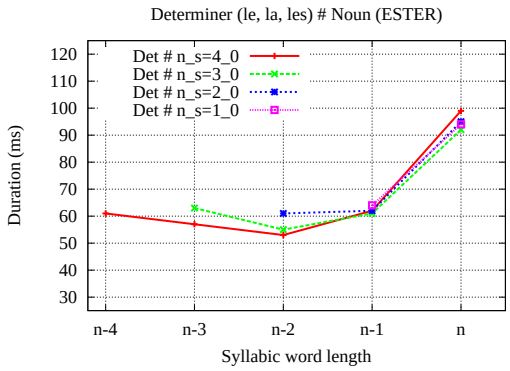
on PFC's lexical words: f_0 is rather stable on all word-initial and internal syllable positions and a rise can be observed for the final syllable. Sound comparisons between noun and noun phrases for PFC can be done only for syllabic lengths of 2 and 3. There are too few occurrences for lengths 4 and 5 to comment on reasonably. Unlike prepared speech, the comparisons don't evidence important differences between nouns and noun phrases in PFC's spontaneous speech material. It is not clear yet whether this result holds for any spontaneous speech corpus. It may be due to the small amount of occurrences in the noun phrase condition. However, it seems clear that prosodic cues to word boundaries are less marked in spontaneous face-to-face speech, where speakers and interlocutors may interrupt their conversation at any point to clarify the subject, if ever some unsolvable ambiguity arose.

Speaking style may cause differences in f_0 profiles between determiners and nouns in which f_0 values are lower for spontaneous speech (PFC corpus) than for prepared speech (ESTER corpus). However, for both speaking styles, it can be asserted that in noun phrases, the f_0 values start with relatively low values that rise as soon as the first syllable of the following noun is produced by the speaker. This f_0 rise information between determiner and first noun syllable may be of help to locate word boundaries and to disambiguate homophones such as *déblocage* ('unblocking') and *des blocages* ('blockings') which phonemes are /deblɔkaʒ/.

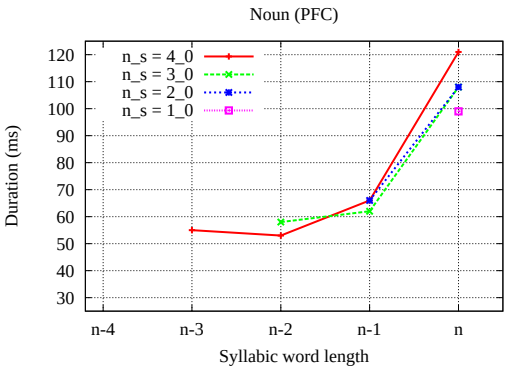
5.3.2 Duration profiles



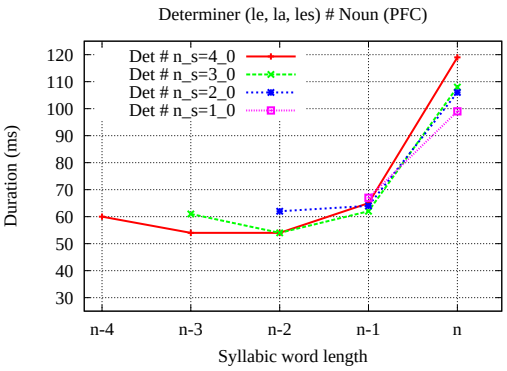
(a) Mean noun vocalic duration profile for ESTER.



(b) Mean noun phrase vocalic duration profile for ESTER.



(c) Mean noun vocalic duration profile for PFC.



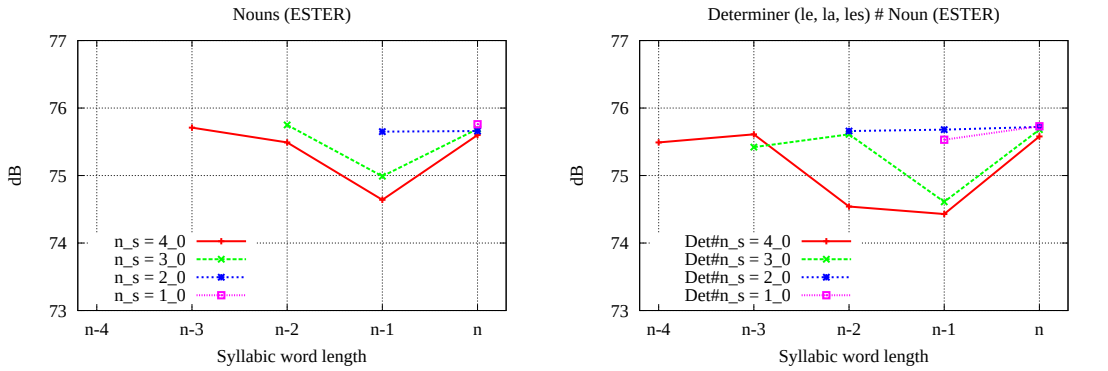
(d) Mean noun phrase vocalic duration profile for PFC.

Figure 5.12: Mean duration profiles for `n`-syllabic length. **Left:** Nouns (without final schwa) **Right:** Noun phrase (`determiner-noun`).

Figure 5.12 illustrates the mean duration profiles of noun words (left figures) and noun phrases (right figures). The results derived from the ESTER corpus are illustrated in the two top subfigures and from the PFC corpus at the bottom. No major profile differences can be noted between noun and noun phrase with the same `n` syllabic length.

Similar results are found for the PFC corpus (Figure 5.12 bottom). Duration profiles do not seem to provide cues to distinguish between items like "*lézard*" and "*les arts*" neither in journalistic speech nor in spontaneous speech.

5.3.3 Intensity profiles



(a) Mean noun vocalic intensity profile for ESTER. (b) Mean noun phrase vocalic intensity profile for ESTER.

Figure 5.13: Mean intensity profiles of `n`-syllabic length for ESTER journalistic speech. **Left:** Nouns (without final schwa) **Right:** Noun phrase (`determiner-noun`).

For the sake of completeness, we will produce intensity profiles of the ESTER corpus to investigate potential differences according to syntax. The results are displayed in Figure 5.13. Intensity profiles show that first and final syllables of a noun (Figure 5.13(a) left) have almost the same intensity values. Within noun phrases (right), the values of the determiner syllable, the first and the final noun syllables are also very close. It is interesting to note that the intensity is slightly lower on the determiner than on the first syllable, which might contribute as a cue for word segmentation. Profiles for PFC were computed but did not reveal interesting cues between nouns and noun phrases. Figures are not included.

5.3.4 Intervocalic measurements

In the previous sections, we have shown averaged fundamental frequency, duration and intensity profile tendencies comparing `nouns` with `noun phrases`. Average values are informative, but they do not give an idea of the underlying distributions. The average profiles display rises and drops in successive average values. In the following, we want to measure rises and drops which are actually observed in the individual speech samples. The current section focuses on intervocalic measurements for fundamental frequency (f_0) and duration. Intervocalic measure-

ments [Woehrling, 2009, p.93] are calculated between two consecutive vowels. Figure 5.14 illustrates the computing strategy using the example sequence: *lundi {breath} le jour* ("Monday {breath} the day").

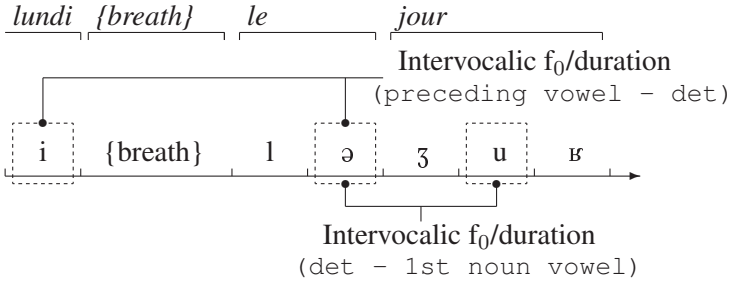


Figure 5.14: Intervocalic measurements. The time axis shows the aligned phone segments of the word sequence *lundi {breath} le jour* ("Monday {breath} the day"). Intervocalic time spans are illustrated between the /ə/ of determiner *le* and the preceding vowel and the following vowel.

5.3.4.1 Intervocalic f_0 distributions

The mean f_0 profiles of `noun` and `noun phrase` gave general tendencies with mean values in section 5.3.1. To verify to what extend the effectively realized f_0 evolution supports these tendencies, we calculated the difference of f_0 (denoted as Δf_0) between two consecutive vowels. The variation of f_0 between two consecutive vowels was calculated as in Equation 5.1:

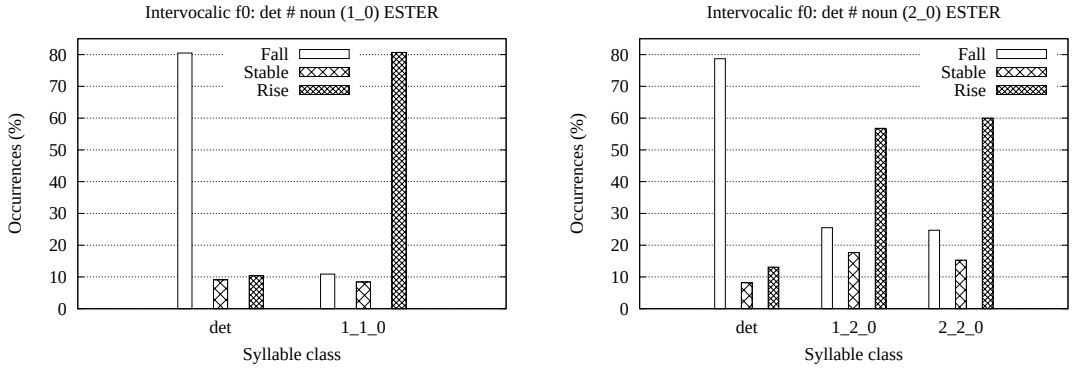
$$\Delta f_0(k) = f_0(V_k) - f_0(V_{k-1}) \tag{5.1}$$

where V_k is a target vowel and V_{k-1} is its predecessor vowel. $f_0(V_k)$ is a mean f_0 value of the target vowel [3] and $f_0(V_{k-1})$ is a mean f_0 value of the preceding vowel. If the f_0 value of the determiner vowel corresponds to a minimum f_0 value compared to the preceding vowel, then the Δf_0 of the determiner must be negative. If an initial accent is realized on the beginning of a `noun` word, then a high proportion of samples are expected to have a positive Δf_0 for the first vowel of `noun` and a negative one for the internal vowels of polysyllabic `nouns`.

To represent our intervocalic measurement distributions synthetically, they were categorized into three classes. The Δf_0 values in semitones (ST) were divided into: **Fall**, **Stable** and **Rise** classes. The **Fall** class contains the proportion of vowels in which Δf_0 was defined to be less than or equal to -1 ST (with respect to the preceding vowel). The **Stable** category consists of Δf_0 comprised between -1 ST and $+$ 1 ST. If Δf_0 is greater than or equal to 1 ST, then this value is in the **Rise** group.

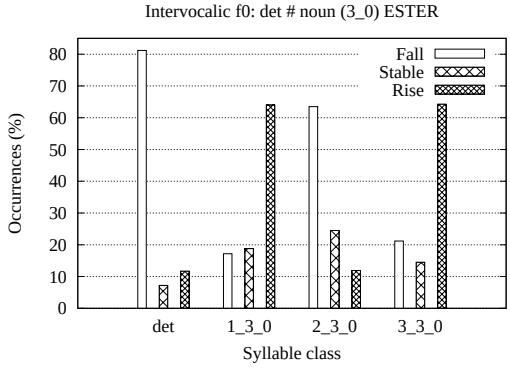
Fall	**Stable**	**Rise**
$\Delta\ f_0 \leq -1$ (ST)	$\Delta\ f_0 \in]-1\ \ +1[$ (ST)	1 (ST) $\leq \Delta\ f_0$

[3] Different ways of computing mean values per segment were experimented with: mean of 3 central frames; mean over all voiced frames of the segment. Results correspond to the latter option. Both methods gave quite similar results.

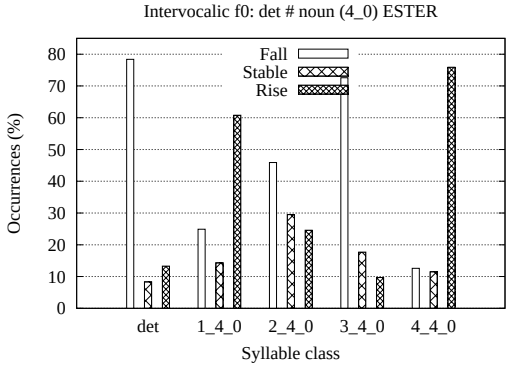


(a) Determiner # monosyllabic.

(b) Determiner # bisyllabic.



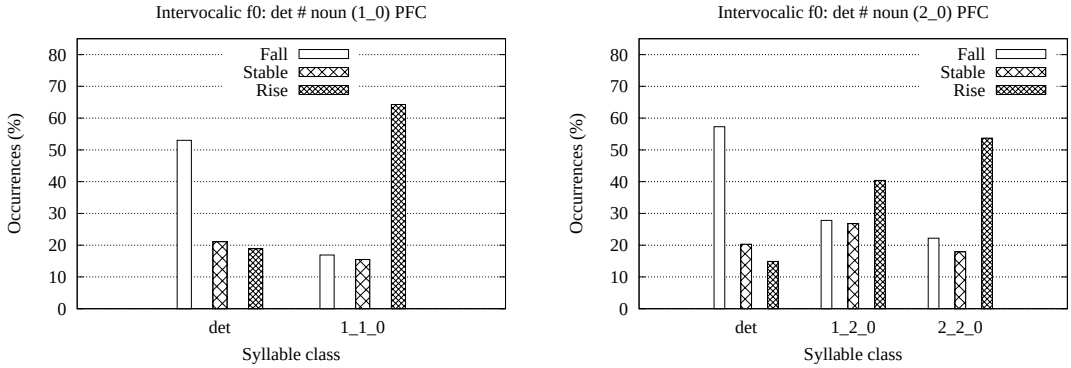
(c) Determiner # trisyllabic.



(d) Determiner # 4-syllabic.

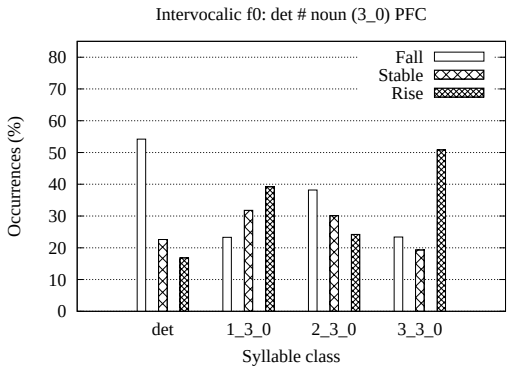
Figure 5.15: Intervocalic f_0 distributions of `noun phrase` for the ESTER corpus. **Top left:** determiner#monosyllic. **Top right:** determiner#bisyllabic. **Bottom left:** determiner#trisyllabic. **Bottom right:** determiner#4-syllabic.

Figure 5.15 illustrates the Δf_0 distributions of the ESTER corpus for the `noun phrases`. The four sub-figures display distributions for determiners followed by monosyllabic (top left), bisyllabic (top right), trisyllabic (bottom left) and 4-syllabic (bottom right) nouns. For all 4 noun lengths, about 80% of determiner vowels mark an f_0 drop of more than or equal to 1 ST compared to the preceding word vowel. This rate, in all sub-figures, does not seem to be dependent on the noun syllabic length. For `determiner - monosyllabic` noun in top left Figure 5.15(a), we can observe 80% of f_0 rise (more than or equal to 1 ST) on first (and last) vowel of a noun after determiner. Concerning Figure 5.15(b) derived from `determiner - bisyllabic noun`, important proportions of f_0 rises (about 60%) are distributed on the two vowels: the initial and the final vowel. This is consistent with the mean f_0 profiles as described in section 5.3.1. For the `determiner` and `trisyllabic` or `4-syllabic` nouns in Figures 5.15(c) and 5.15(d), we can find at least 60% of a f_0 rise on the initial vowel of a noun, that confirms the initial accent tendency, and 60% (respectively 70%) of f_0 fall on the penultimate (2_3_0) (respectively 3_4_0), preparing the final accent realization. Experiments described in [Welby, 2007] revealed that f_0 inflection called "elbow" between a function word and content word and early f_0 rise on a content word are cues to content word beginnings. In future experiments we plan to refine our

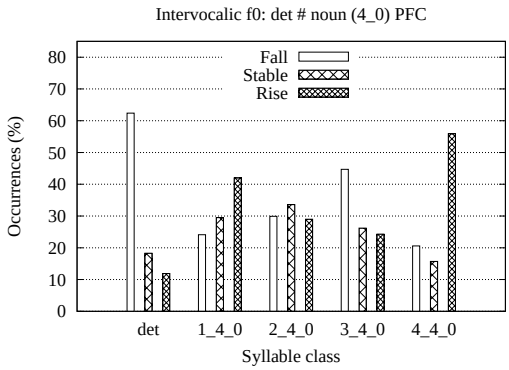


(a) Determiner # monosyllabic.

(b) Determiner # bisyllabic.



(c) Determiner # trisyllabic.



(d) Determiner # 4-syllabic.

Figure 5.16: Intervocalic f_0 distributions of `noun phrase` for the PFC corpus. **Top left:** determiner#monosyllabic. **Top right:** determiner#bisyllabic. **Bottom left:** determiner#trisyllabic. **Bottom right:** determiner#4-syllabic.

methodology to produce more detailed insight concerning these cues.

As for the PFC corpus (Figure 5.16), the observed tendencies are quite similar to those of the ESTER corpus, although with weakened proportions. The proportions of f_0 drops for the determiner are lowered by 20 – 30% and the f_0 rise rate at final vowels is also lowered by about 5–15%. As for the mean f_0 profiles of the PFC corpus which were flattened as compared to the ESTER corpus, intervocalic f_0 measurements also result in more flattened intervocalic Δ f_0 distributions than for the ESTER corpus. This result indirectly confirms the validity of the average f_0 profiles as computed earlier.

5.3.4.2 Intervocalic duration distributions

To complement the average duration profiles of section 5.3.2, we will now measure intervocalic durations which are actually observed in the individual speech samples. We will present these intervocalic duration statistics with the motivation to examine these durations with regard to the previous duration profiles.

For a target vowel, its intervocalic duration is measured by the time span between the centers of the target vowel and its preceding vowel. For word-initial vowels, the preceding vowel corresponds to the last vowel of the preceding word. The intervocalic duration can be seen as an approximation of syllabic duration. We may hypothesize frequent long intervocalic durations on word boundaries and for word final lexical syllables, whereas short durations may be expected on word-internal syllables. For example, if a determiner is closely connected to the preceding words in a prosodic phrase internal position such as "*j'ai vu* **le train**" (I saw the train), then its intervocalic duration may be short. However, on a prosodic word start such as "*quand je suis arrivé,* **le train** *était parti*" (when I arrived, the train was gone) the intervocalic duration on the determiner is expected to be long. The proposed measure may then give an indication of proportions of determiners in the different prosodic word positions.

Figures 5.17 (ESTER) and 5.18 (PFC) show the statistics for `noun phrase`, using the following duration classes:

Short	Middle	Long
$[30ms\ -\ 155ms]$	$[160ms\ -\ 220ms]$	$[225ms\ -\ [$

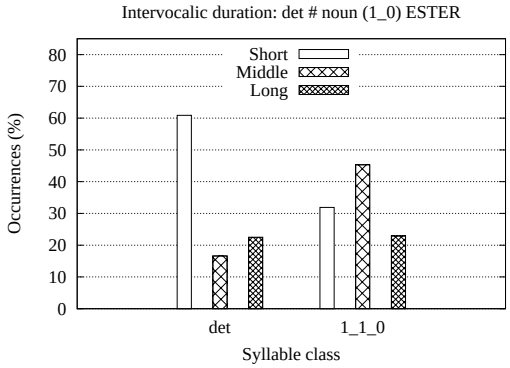
The boundaries of the duration classes were fixed with the idea of accepting a large proportion of items in the short class, as the maximum duration of 155 ms corresponds to a segment duration below 80 ms for a basic CV [4] syllable structure (half a vowel starting at the middle of V followed by 1 C followed by half a vowel to the middle of the next vowel). Middle and long duration classes then give an idea of the proportions of syllables with longer durations, middle being medium long and the long class comprising items for which segment durations are above 110 ms on average for a simple CV structure. We will not comment on the absolute rates observed for the different classes in the different investigated conditions. However, we will focus on the differences observed across conditions.

The results obtained from the ESTER corpus are displayed in Figure 5.17. We will first comment on the first sub-figure (top left) containing all noun phrases composed of a determiner (*le*, *la*, *les*) and a monosyllabic noun, before addressing the remaining ones giving the same type of information but for polysyllabic words of progressively increasing length. For determiners, sub-figure 5.17(a) shows a high proportion (over 60%) of short intervocalic durations connected to their predecessor words without duration lengthening. However, a significant proportion (20%) of tokens belongs to the long class which may correspond to a prosodic word initial position (cf. "*le train*" example above). For monosyllabic nouns, the same proportion of tokens belongs to the long class. However the rates of the short class are relatively low (30%) as can be expected on noun final syllables. The majority of tokens belong to the middle class. Looking at the polysyllabic noun cases ($n > 1$), one can observe:

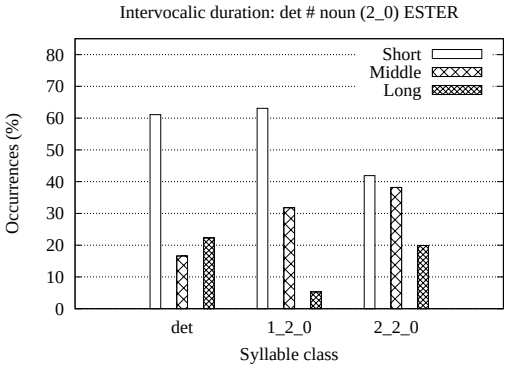
(i) All **initial noun syllables 1_n_0** have a close to zero rates for the Long duration class, which characterizes the connection between determiner and noun;

(ii) Short class rates are highest on **word-internal noun syllables m_n_0** ($m > 1$) and rates increase with syllabic word length;

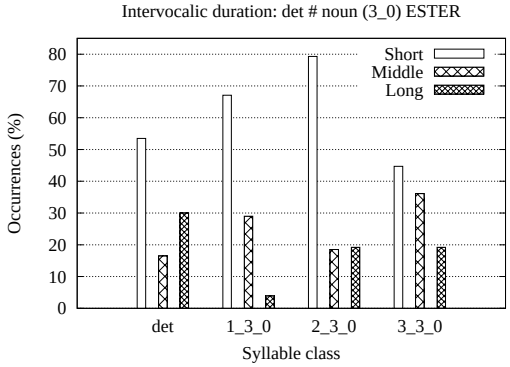
[4] For French, the CV syllable structure is the most frequent one with rates between 60 and 70% depending on syllabification rules and corpora [Adda-Decker *et al.*, 2005].



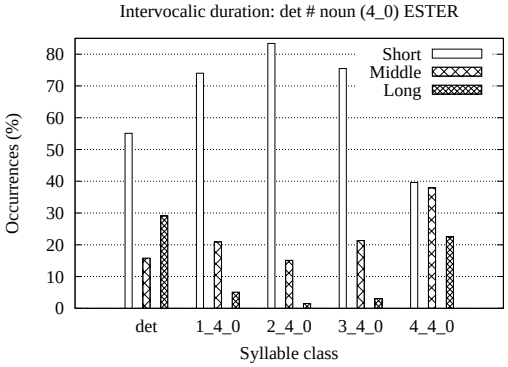
(a) determiner # monosyllabic noun.



(b) determiner # bisyllabic noun.



(c) determiner # trisyllabic noun.



(d) determiner # 4-syllabic noun.

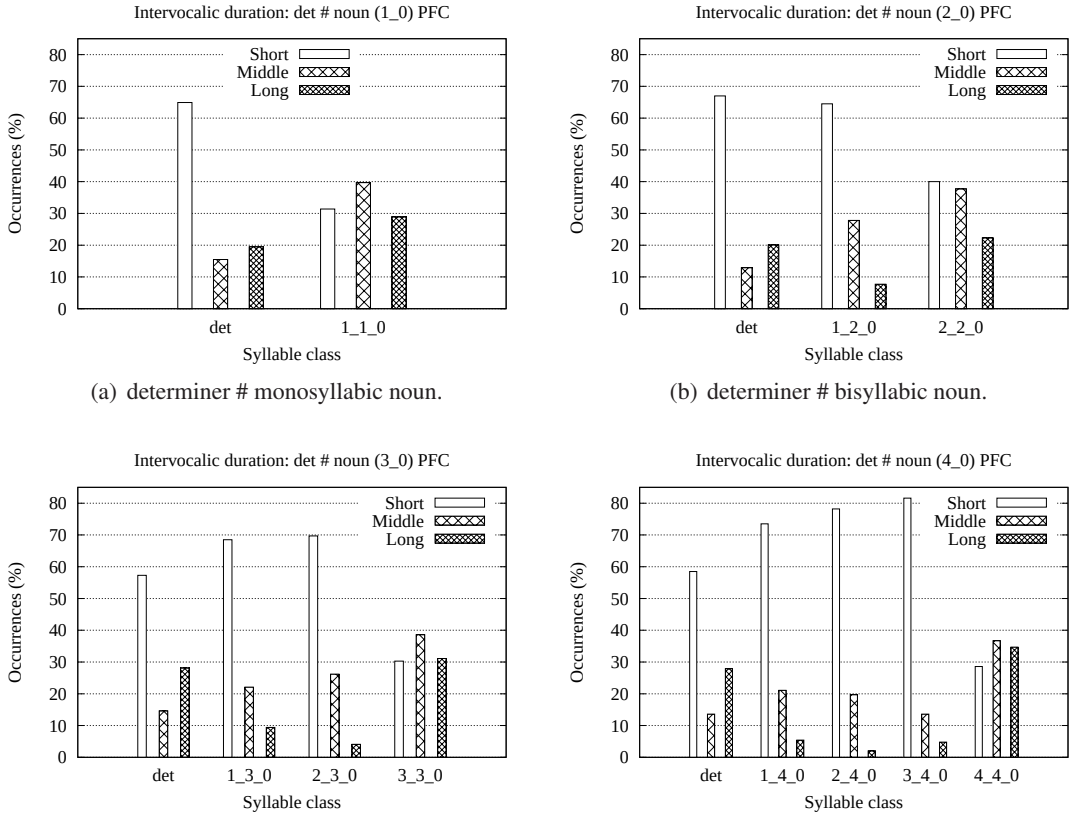
Figure 5.17: Intervocalic duration distributions of `noun phrase` for the ESTER corpus. **Top left:** determiner # monosyllic. **Top right:** determiner # bisyllabic. **Bottom left:** determiner # trisyllabic. **Bottom right:** determiner # 4-syllabic.

(iii) It can be observed that the proportion of small duration tokens increases globally with syllabic word length for all syllabic positions (initial, internal, final). However, as expected, the final vowel position yields much lower rates of Short class for all configurations.

Results confirm that longer syllabic words tend to be uttered more rapidly with the internal vowels particularly prone to temporal shortening. These phenomena are to be considered in future pronunciation dictionary design for ASR systems.

The proportion of tokens in the Short class of `determiners` is also high, but we cannot neglect the rates corresponding to the Long class which are at least as high as or even higher than for the final vowel of `noun`. Long intervocalic durations on determiners may be a cue to find a boundary between sentences or phrases.

Results for spontaneous speech of the PFC corpus are displayed in Figure 5.18. It is worthwhile to note that the spontaneous speech distributions are almost identical to those of prepared speech. However, some differences may be highlighted, which are in line with previous observa-



(a) determiner # monosyllabic noun.

(b) determiner # bisyllabic noun.

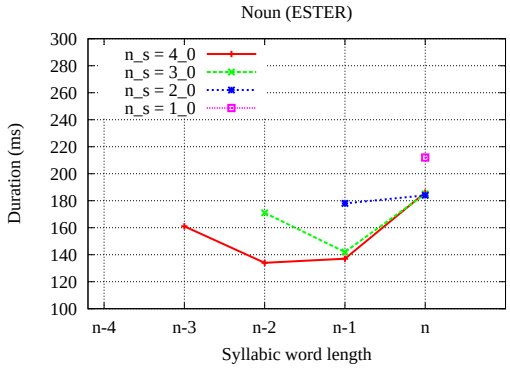
(c) determiner # trisyllabic noun.

(d) determiner # 4-syllabic noun.

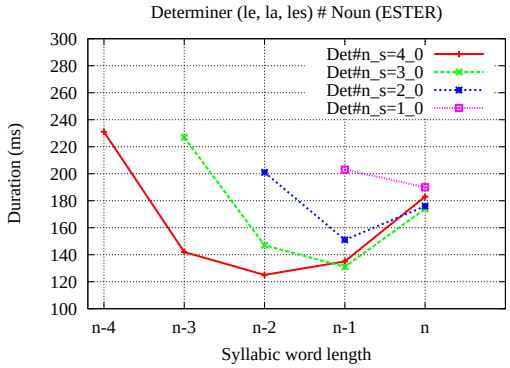
Figure 5.18: Intervocalic duration distributions of `noun phrase` for the PFC corpus. **Top left:** determiner # monosyllic. **Top right:** determiner # bisyllabic. **Bottom left:** determiner # trisyllabic. **Bottom right:** determiner # 4-syllabic.

tions (larger range of duration variation in long final syllables, cf. Figure 5.12(c)). For example the rate of Long class items on word-final syllables is about 10% higher for spontaneous speech than for prepared speech. One may note, when comparing ESTER vs. PFC polysyllabic data on noun-final syllables (right most distribution of sub-figures (b),(c) and (d)) a "falling" distribution for ESTER (the rate of short tokens remains highest with around 40%, a bit less for Middle class and even less for the Long class). For the PFC data, this "falling" tendency is not observed for the trisyllabic and 4-syllabic nouns: here the rate of Middle class items has the highest rates, and the Long class rates are above those of the Short classes. This observation translates the fact that the journalistic speech can be considered as steady-state dense speech, whereas spontaneous speech is more prone to rhythm or local speech rate variations.

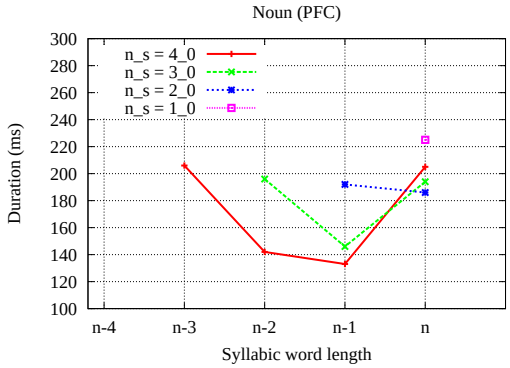
Coming back to the duration profiles of section 5.3.2, we noticed that average durations remained almost identical for all but noun-final syllables, in particular for the vowels around the determiner-noun boundary. We hypothesize that intervocalic duration may contribute to provide word boundary cues for the determiner. Intervocalic duration is measured as illustrated in Figure 5.14, with one minor difference: for `determiner` vowels (for which the preceding vowels correspond to the last vowel of the preceding word event though there might be a breath, silence, hesitation or



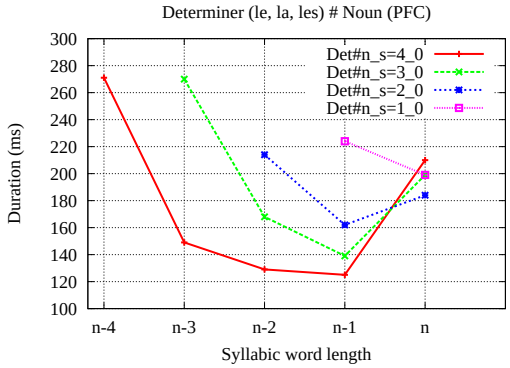
(a) Mean noun intervocalic vocalic duration profile for ESTER.



(b) Mean noun phrase intervocalic duration profile for ESTER.



(c) Mean noun intervocalic duration profile for PFC.



(d) Mean noun phrase intervocalic duration profile for PFC.

Figure 5.19: Mean intervocalic duration profiles for `n`-syllabic length. **Top left:** Noun for ESTER, **Top right:** Noun phrase for ESTER, **Bottom left:** Noun for PFC, **Bottom right:** Noun phrase for PFC. **Bottom:** Noun phrase (`determiner-noun`).

any other break) intervocalic durations above 3 seconds were excluded. The results for this measurement are illustrated in Figure 5.19. The results derived from the ESTER corpus are illustrated on the top and on the bottom for the PFC corpus. The mean intervocalic duration profiles of `nouns` are displayed on left panels and those of `noun phrases` are on right panels. For the ESTER corpus (top), we can observe for nouns (left) that intervocalic durations are highest for the last position `n` and lowest for word-internal positions, as expected. The noun phrase sub-figure (right) clearly illustrates that the longest intervocalic duration is achieved on the determiner position (duration between preceding vowel - `determiner` vowel) for each `n` syllabic word class. The PFC corpus results are similar to those of the ESTER corpus and in particular the high intervocalic durations on determiners provide a strong cue for word boundary location. However, some differences are noteworthy even though the low number of observations in each sample population urges us to remain cautious. The PFC corpus (as compared to ESTER) produces longer intervocalic durations on first and last syllables of nouns and also for the determiner of the noun phrase sub-figure (right). These results from `noun phrases` are expectable because `noun`

phrases can be uttered after breath, silence, or some hesitation event. These factors produce longer intervocalic duration and tend to be more frequent in spontaneous speech than in prepared speech.

5.3.5 Homophone noun phrases: fine phonetic detail?

Table 5.5: Quantitative ESTER corpus description of noun phrases (*la* noun vs. *l' a-* noun) w.r.t. word tokens of word syllabic length n. Counts are separated for nouns preceded by determiners *la* (top)/*l'* (bottom). *Syll.class* n_s states n: the number of full syllables; s: absence (0)/presence (1) of final schwa.

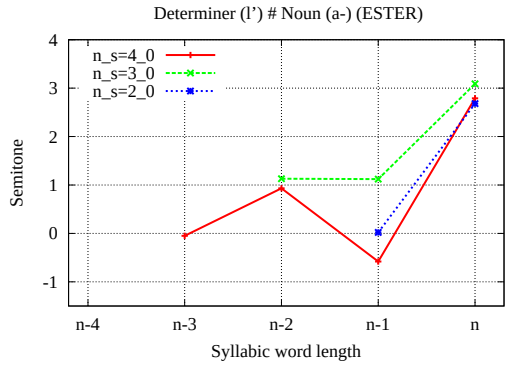
n	n_s	*#la # noun*	*Examples*
2	det#1_0	1073	la vie (the life)
3	det#2_0	1054	la tension (the tension)
4	det#3_0	528	la situation (the situation)
5	det#4_0	409	la reconstruction (the reconstruction)
n	n_s	*#l' # a- noun*	*Examples*
1	1_0	13	l'âge (the age)
2	2_0	265	l'avis (the opinion)
3	3_0	131	l'attention (the attention)
4	4_0	140	l'actualité (the actuality)

The comparison between *n*-length nouns and noun phrases (determiner–noun) highlighted major differences between their corresponding profiles, in particular for the f_0 profiles (see Figures 5.11(a) and 5.11(b)). Before closing this chapter, we would like to come back to the homophone problems earlier addressed in chapter 4. In the following, we propose to apply our previous methodology and findings to investigate the question of "fine phonetic detail" or more appropriately "fine prosodic detail" to discriminate between **homophone noun phrases** (a special type of multiword homophones as introduced at the beginning of this chapter) such as *la tension* (the tension) and *l'attention* (the attention) /latɑ̃sjɔ̃/. For those less familiar with French, we explain that determiner *la* is used before feminine singular nouns, whereas determiner *l'*, elided from *le* or *la*, is employed before masculine or feminine singular nouns which begin with a vowel or mute h (*h muet* in French) [5]. In our example, both sequences have the same canonical pronunciations, the same number of vowels and syllables. However, in the first case, the first vowel belongs to the determiner, whereas in the second case it corresponds to the noun-initial vowel. This latter noun phrase case tends to be closer to a simple noun case, as all the produced vowels belong to the noun. According to our comparative profiles of nouns and noun phrases, f_0 on determiners tends to be low before rising to the initial syllable of the following noun, whereas on lexical word initial syllables, f_0 tends to be relatively high before starting to drop towards the penultimate syllable minimum. Psycholinguistic studies by Spinelli et al. [2007; 2010] demonstrated that humans are able to discriminate ambiguous phrases of a definite article followed by a noun like *la fiche* and *l'affiche* [lafiʃ]. The authors revealed that in their controlled material the f_0 of the determiner *la* is lower than the f_0 of the first noun syllable *l'a–*. This is perfectly in line with what we might expect,

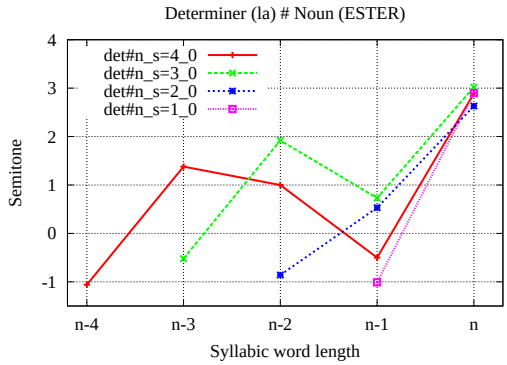
[5] Most *h* of French words is mute: *h* is not pronounced and the word is considered as if it begins with a vowel. For example, the word *habit* /abi/ (dress), masculine singular noun, becomes *l'habit* /labi/ ('the dress') instead of *le habit*.

given our earlier f_0 profiles. The question is now whether this tendency still holds for uncontrolled material with formal similarities (*la NOUN* vs. *l' a-NOUN*) extracted from large corpora.

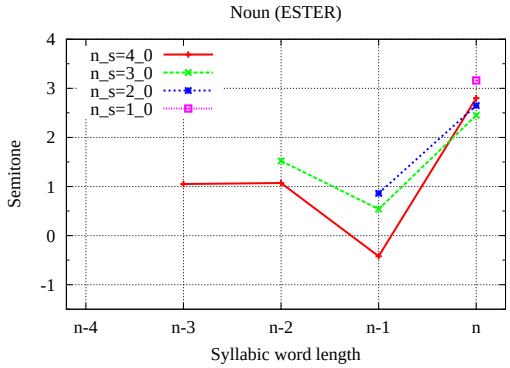
Because of the limited number of tokens, we limited our study to the ESTER corpus. Furthermore, in order to collect enough samples we limited the homophone constraint to the first syllable (ambiguity between *la* and *l'a*) and not to the whole noun phrase (*la fiche* and *l'affiche*) and put the word `homophone` between quotation marks. Exact counts are reported in Table 5.5. Monosyllabic *l' a-* noun phrases were excluded from the study because of both few tokens (13) and the lack of a monosyllabic *l'a-* noun contrast condition.



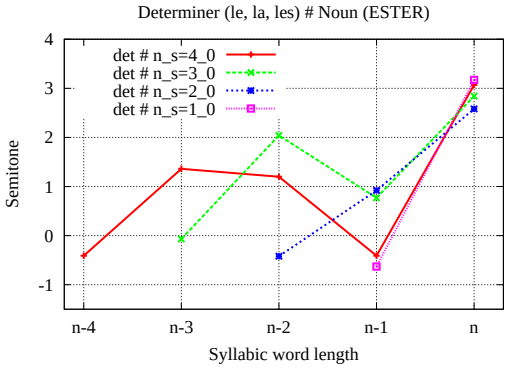
(a) Mean noun f_0 profile for ESTER: *l' # a-*noun.



(b) Mean noun phrase f_0 profile for ESTER: *la* # noun).



(c) Mean noun f_0 profile for ESTER.



(d) Mean noun phrase f_0 profile for ESTER.

Figure 5.20: Mean f_0 profiles for `n`-syllabic length in comparison with ambiguous noun phrase for ESTER corpus. **Top left:** *la* # noun, **Top right:** *l' # a-*noun, **Bottom left:** Noun phrase, **Bottom right:** Noun.

The f_0 profiles of our "homophone" noun phrases of the ESTER corpus are illustrated in Figure 5.20: *l' a-* noun phrase (Figure 5.20(a) top left) and *la* noun phrase (Figure 5.20(b) top right). To compare these "homophone" noun phrase profiles with average f_0 noun and noun phrase profiles of the ESTER corpus in section 5.3.1, the corresponding two sub-figures Figure 5.20(c) (mean noun f_0) and Figure 5.20(d) (mean noun phrase f_0) are added at the bottom.

For the comparison between *l'#a-* noun (top left) and *la*#noun (top right), we are thus mainly interested in the comparison of the average f_0 values of the first syllables, either belonging to the

vowel-initial noun (left) or to the determiner (right). As a preliminary remark, we may highlight that the average profiles of the (small) *la*#noun subset (top right) is quite similar to the full set of determiner-noun phrases (top bottom). Coming back to our focus, we may next observe that the top left figure of the n *l'*#*a-* noun profiles features initial vowel f_0 values which are about 1 ST higher than those of the contrast situation *la*#noun (top right). One may notice that *la* determiner f_0 values (top right) are lower than the average of all determiners (bottom right). Finally, we would like to address the differences between average *a*-nouns (top left) and all nouns (bottom left). Lower f_0 values on first vowel for *a*-nouns of about 0.5–1 ST can be observed. The mean f_0 values of *a-* beginning noun after *l'* determiner indicates a transition from determiner to noun. A more elaborate methodology may be designed to provide more in-depth investigations of the highlighted fine prosodic details in future studies.

5.4 Conclusion

A first aim of the presented corpus-based study was to establish an automated methodology to investigate overall prosodic properties of French on a lexical level. The idea was to test whether the proposed methodology managed to provide already known results. If so, it might also contribute to produce more original results, or at least worthfull hypotheses to be passed over to linguists for further in-depth investigations as well as to speech scientists for appropriate model adaptations and/or specific post-processing steps. Large annotated corpora allow us to envision a broad range of contextual investigations. A crucial issue is then to focus on the most relevant contexts. Relevance may be motivated either by linguistic criteria or criteria arising from automatic speech processing (e.g. ASR word errors). The proposed methodology combined different automatic processing steps (f_0 and intensity measurements using Praat, duration measurements from lexical and phonemic alignments using the LIMSI speech recognizer, POS tags with the LIMSI *Wmatch* tool including a French TreeTagger) to produce different levels of prosodic annotations together with sound word classes.

A second major aim was to compare prosodic properties of French across speaking styles. Spontaneous speech entails much higher word error rates than prepared (journalistic) broadcast speech. Our hypothesis was that beyond a higher number of pronunciation variants proper, there might be important measurable differences in prosodic cues and that both phenomena may be related. Our belief is that pronunciation variants and prosody are tightly intertwined. ASR and more particularly ASR errors provide researchers the opportunity to put focus on this complex problem and may open new pathways to a related pronunciation-prosody research domain.

More traditional prosodic investigations focus on larger than word units, which may be called prosodic words, intonation phrases, accent phrases, chunks... depending on authors and analysis levels. Our work started with lexical units (as these are the units of pronunciation variants) and we proposed to measure average prosodic (f_0, duration, intensity) profiles of French words or more precisely, of word classes. The introduction of **word classes** is a simple, nonetheless a major contribution, as a purely lexical basis would have limited our study to a small subset of word types due to the uneven word distribution in languages (Zipf's law). The introduced **lexical classes** are motivated by prosodic (syllabic length) and/or syntactic criteria. The proposed study thus examined word classes of syllabic length 1 (monosyllabic words), of syllabic length 2 (bisyllabic words)... Each class of n-syllabic words thus included a large number of tokens providing

a statistically interesting basis for further investigations. As different prosodic properties are expected for grammatical words as compared to lexical words, our data were split according to their status lexical vs. grammatical. Furthermore, the word-final schwa was expected to influence the overall prosodic realizations.

Two types of measurements were used for the proposed comparisons: average profiles (for f_0, duration and intensity) of syllable-based measurements and distributions of intervocalic measurements. The intervocalic distributions allowed us to check that the acoustic realizations were in line with the average profiles.

The measured f_0 and duration profiles confirmed the intonational patterns for French already known in the literature (i.e. tendency of f_0 rise and duration lengthening on final syllables in lexical word) and hence validated the potential of the proposed methodology. Regularities with increasing word syllable length could also be highlighted for lexical words: Δf_0 (rise) between penultimate and final syllables tends to grow with syllabic word length; word-final syllable duration tends to increase with syllabic word length (concomitantly with a slight average duration decrease for all but word-final syllables). Interestingly, our duration profiles show long durations on the final word syllable, however almost identical average durations on the preceding (initial and word-internal) syllables. This result of stable average durations goes in favor of the isochronous syllable duration theory [Pike, 1945] for French: when measuring average syllable (vowel) durations on non word-final syllables we find very close mean values across successive syllables (vowels). Furthermore, other detailed and less (or not yet) described findings deal with the correlation between word final schwa and a slight but global prosodic (f_0, duration, intensity) profile increase. We consider this as an interesting observation to be further checked in the future on larger data sets. This result may be of relevance for spotting named entities, or other words with semantic focus. A contrastive study between fast and slow speech rate words (cf. definitions used in section 5.2.4) showed that average f_0 profiles tend to be higher on slowly articulated words.

Concerning differences between **speech styles**, measured f_0 profiles for spontaneous face-to-face speech (PFC corpus) had much flatter shapes than those of the prepared speech ESTER corpus. This indicates that less acoustic-prosodic cues can be found in spontaneous speech. This might then be at least a partial explanation of increased ASR difficulties on spontaneous speech. Future studies require investigations of correlations between flat f_0 profiles and pronunciation variants (reduction, hypo-articulated speech). Speaking style does not impact average duration profiles as much as f_0 profiles. **Grammatical word** f_0 profiles showed relatively low f_0 values on word-final syllables. In particular, they don't exhibit f_0 rises on final syllables.

Prosodic parameters are particularly interesting to be studied on larger units such as phrases or prosodic words. The delimitation of larger units is not straightforward. As a step in this direction, we proposed to study simple noun phrases limited to `determiner - noun` bigrams and to compare the corresponding profiles to single word noun profiles. The measured profiles show that journalistic speech contains important prosodic cues to distinguish between n-syllabic nouns and noun phrases. For spontaneous speech, the results show only slight differences between nouns and noun phrases. It is not clear whether this result holds for any spontaneous speech corpus. However, it seems clear that prosodic cues to word boundaries are less marked in spontaneous face-to-face speech. Our results thus demonstrate that speaking style may cause important differences in f_0 profiles, in particular between determiners and nouns. However, for both speaking styles, noun phrases start with relatively low f_0 values that tend to rise to the first syllable of the following noun.

This f_0 rise information between determiner and first noun syllable may be of help to locate word boundaries and to disambiguate homophones such as *déblocage* ('unblocking') and *des blocages* ('blockings') which phonemes are /deblɔkaʒ/.

A final investigation was limited to noun phrases and in particular homophone noun phrases with the initial vowel belonging either to the determinant or to the noun (*la vie* "the life" vs *l'avis* "the opinion"). This allowed us to address the issue of "fine prosodic detail" to discriminate between **homophone noun phrases**. The related f_0 profiles highlighted a clear difference in average f_0 values with respect to the first syllable: low if the first vowel belonged to the determiner; some intermediary value (between low determiner and high first noun syllable) in the special case where the determiner had no vowel and the first vowel belonged already to the noun. This result on data extracted from large corpora is in line with results from a psycholinguistic studies [Spinelli *et al.*, 2007; Spinelli *et al.*, 2010] on a small set of controlled material.

We hope that the overall results of this chapter, dealing both with the proposed methodology and the prosodic measurements across speaking styles contribute to open new perspectives for pronunciation modeling in ASR and beyond, for corpus-based linguistic studies, with a focus on the inter-relation between pronunciations and prosody.

Conclusions

This thesis has investigated acoustic and prosodic characteristics of French using large-scale audio corpora of different speaking styles: prepared broadcast news speech and spontaneous face-to-face speech. The ESTER (*Évaluation des Systèmes de Transcription d'Émissions Radiophoniques*) corpus is mainly composed of broadcast news in which most of the speech is uttered in a prepared speaking style. Spontaneous speech stems from the PFC (*Phonologie du Français Contemporain*) corpus and was collected during interview sessions between two or three acquainted persons or between unknown persons. A long-term objective of the proposed investigations concerns improved automatic speech recognition (ASR) systems by improving pronunciation modeling. On a more short-term perspective, our goal was to increase our knowledge of pronunciation variation across speaking styles with a focus on acoustic-prosodic features. These features attracted our attention with the objective of discriminating between homophone words and word sequences. As a matter of fact, a major reason to ASR errors is homophones and near homophones. The French language includes a large proportion of homophones and multiword homophones. Almost any phoneme corresponds to a single written word and many phonemes may be written in different ways (/a/: *a*, *as*, *à*; /o/: *au*, *aux*, *eau*, *eaux*, *haut*, *hauts*, *oh*; /s/: *s'*, *c'*;...). This entails that a word of two phonemes (e.g. /ma/ *ma*) is decomposable in a homophonic word sequence of shorter words (/m#a/: *m'a*; *m'as*) and the question of acoustic-prosodic correlates to word boundaries becomes particularly relevant in such situations. The proportion of homophones in a given speech corpus is first related to language characteristics. Beyond these, speaking style may contribute to increase the proportion of homophonic sequences due to reduced pronunciations and hypo-articulated speech. We are then especially interested in particularities of segmental phonetics and prosody that may characterize pronunciations in terms of position within a word and across words as well as of grammatical categories. The term of prosody is broadly used to indicate accent, tone, stress... at a lexical level and intonation in more wider range at a postlexical, non-lexical, or supralexical level [Hirst and Di Cristo, 1998; Lacheret-Dujour, 2000]. Measurable acoustic correlates of prosody are mainly composed of fundamental frequency (f_0) related to voice height, duration in connection with rhythm and speech rate, and intensity to express voice strength or power.

During this thesis, a new methodology based on automatic analysis of large-scale spoken data was developed to investigate overall prosodic properties of French at a lexical level. Automatic processing tools facilitated our studies: the automatic speech recognition (ASR) and automatic speech alignment systems of LIMSI for phone/phonemic segments and word segments, the PRAAT software to extract f_0, first three formants, and intensity. This methodology was first developed for the acoustic-prosodic investigations of the frequent homophone word pairs (chapter 4), and then extended in chapter 5 to investigate and describe the whole French corpus.

As mentioned above, in French, many errors caused by ASR systems arise from frequent (near) homophone words, for which ASR systems principally depend on language model (LM) weights. In chapter 4, homophone words were selected based on their frequency in ASR errors to verify whether these homophones could be discriminated by their mere acoustic and prosodic properties depending on their part-of-speech (POS) dependency or their position within prosodic words/phrases. The acoustic analyses included segment duration, voicing ratio, f_0, and neighboring pauses. For the first selected homophone pair, "*et*" ("and")/"*est*" ("to be"), their duration and voicing ratio distributions showed clear differences to distinguish these two words. However, observed measurements were less convincing for the second pair: "*à*" ("to, at")/"*a*" ("to have"). Overall, the duration distribution of the conjunction *et* tended to have a flatter distribution than the verb *est*, while the *à*/*a* comparison did not show a significant difference between the two items. Function words (*et*, *à*) had weaker voicing ratios than the verbs (*est*, *a*). Co-occurrence of left and right pauses with the target words was in favor of conjunction *et* and preposition *à* due to their position within prosodic words/phrases.

At least some differences between the selected homophone pairs were observed thanks to the acoustic measurements. Thus some tests were carried out to verify whether these measured acoustic differences might be useful to automatically discriminate between the two homophone word pairs. To this end, an automatic classification task was designed with the help of the WEKA platform and a set of 62 acoustic and prosodic attributes including static and dynamic (denoted as Δ) parameters was defined. The automatic classification made use of the major relevant classification algorithms (Bayesian classifiers, Decision Trees, Support Vector Machine, etc.) of data mining techniques. The results highlighted the role of the prosodic (f_0, duration, voicing, and intensity) and contextual information (co-occurrence of pauses) in distinguishing between the target words. As for the acoustic measurements, automatic classification rates (*et*/*est* pair: 71.3% (mean), 79.8% (best) for the ESTER corpus and 76.3% (mean) and 83.1% (best) for the PFC corpus; *à*/*a* pair: 66.3% (mean), 72.9% (best) for the ESTER corpus and 61.6% (mean), 69.4% (best) for the PFC corpus) performed better for *et*/*est* than for the *à*/*a* pair. This must be related to the fact that 30% of the data kept a vowel timbre difference between *est* /ɛ/ and *et* /e/. Promising results were achieved for the PFC corpus since spontaneous speech generally presents more errors than prepared speech during automatic speech transcription processing.

A further question was whether all the 62 acoustic-prosodic attributes effectively contributed to the classification result. To identify the most informative features, some feature selection experiments were implemented. These experiments showed that inter-segmental or prosodic (22 inter-segmental attributes and 32 prosodic attributes) attributes alone achieved almost as good performances as the full set of attributes (average of the *et*/*est* pair: around 70% for the ESTER corpus and 77% for the PFC corpus. *à*/*a* pair: about 66% for the ESTER corpus and 61% for the PFC corpus). In the limited set of 15 most effective attributes, there were more prosodic attributes than formant related parameters, and more intra-phonemic than inter-phonemic attributes. The ratio of static and dynamic attributes was almost equal. Most important attributes were left pause, Δintensity, F2, Δf_0 , voicing ratio, and duration. These selected attributes also revealed as good results as the results from all 62 attributes.

As a last step, human performances to discriminate between these two specific homophone pairs were checked. In order to verify if humans mainly rely on acoustic-prosodic parameters to discriminate homophone words or if they tend to use context information similar to n-gram language models (LMs) for ASR systems, two types of perceptual tests were carried out: acoustic+language

model (AM+LM) condition test and language model (LM) condition test. The *et*/*est* (near) homophone pair of the ESTER corpus was considered for the perceptual test. The 7-gram chunks with as much information around the target word as used by a 4-gram LM-based transcription system were adapted to the perception test. Perceptual results were measured in terms of erroneous transcription of the target words compared to the reference transcriptions. Human error rates were then compared to ASR word error rates. Human transcriptions' analysis showed that no error was found in distractor stimuli. A very few error rates were shown on the perfectly decoded stimuli by the ASR system. An important increase in the human error rate was observed on the stimuli subset corresponding to *et*/*est* symmetric confusions.

The comparison between the system transcriptions and the human ones revealed that humans achieved better performances for both tests although stimuli which were hard to recognize for ASR systems because of the local ambiguity were also problematic for humans. Compared to the language model (LM) condition test, the AM+LM condition test generated slightly less errors. This result suggested that acoustic and prosodic information might help in the right selection of the target word in similar ambiguous syntactic structures.

As a step further, we examined overall prosodic properties of French both at lexical and phrase levels (cf. chapter 5). This work is based on the hypothesis that pronunciation variants are due to varying prosodic constraints. What's more, the proposed investigation raises the question of the link between a high number of the pronunciation variants and the measurable differences in prosodic cues. Chapter 5 focuses on this link via measures of average prosodic (f_0, duration, intensity) profiles of French words or more precisely, of word classes with n-syllabic word length. To sum up the premises of this work, the experiments described in chapter 5 have the following aims: 1) establishing an automated methodology to investigate overall prosodic properties of French words; 2) comparing prosodic properties of French across speaking styles (prepared and spontaneous).

Pronunciation variants are often due to shorter or longer pronunciations, to added or deleted segments or even syllables, which then entail different prosodic characteristics. Male speakers of the ESTER corpus and the PFC corpus were used to globally investigate prosodic (fundamental frequency (f_0), duration, and intensity) realizations via average prosodic profiles comparing between lexical and grammatical words or noun and noun phrase. The presented methodology took advantage of time-aligned phonemic and lexical transcriptions, as well as of prosodic and Part-Of-Speech (POS) annotations to compare average prosodic profiles according to word classes of given syllabic length, word final-schwa, duration, and phrases.

Three prosodic parameters (f_0, duration, intensity) were investigated using average n-syllabic word profiles. Using this methodology, the impact of syllabic word length was studied with the underlying idea that word-internal syllables might be more prone to temporal reduction phenomena and pronunciation variants. The proposed profiles enable us to give a synthetic overview of what happens to f_0, duration and intensity for different syllabic positions of French words. The data was split in various subsets: lexical words, grammatical words, presence or absence of word final-schwa, and different speaking styles.

First of all, we presented f_0 profiles. Higher f_0 values could be measured at the final syllables of lexical words in the both corpora. The presence of word-final schwa correlates with a global f_0 increase. The spontaneous speech of the PFC corpus displayed flatter profiles than the prepared ESTER corpus. Grammatical words globally featured lower f_0 values than lexical words. Second,

we conducted duration analyses. Longer final syllable durations were observed in lexical words of the two corpora. The duration variation range of final syllables was smaller with final-schwa than for words without final schwa. Spontaneous speaking style showed greater duration range variation of final syllables than prepared speech. Concerning grammatical words, longer final syllable durations were not observed in both corpora. Third, we investigated intensity. Contrary to two preceding parameters (f_0 and duration), remarkable final intensity values could not be measured. For lexical words, most of the time, final syllable intensity values were at best as high as first syllables. Much lower intensity values were noticed at final-schwa. As for grammatical words, the intensity values of final n syllables were almost the same values as lexical word ones. This result was opposite to those from f_0 and duration where grammatical words had lower f_0 and shorter duration for final syllables as compared to lexical words. Final accentuation is thus best correlated with f_0 and duration. The study of duration impact on f_0 was investigated for lexical words without final-schwa. The data was divided into two categories (slow and fast). Slow rate categories showed higher f_0 values for both speaking styles. Spontaneous speech further showed less variation in f_0 profiles as compared to slower spontaneous speech. Results confirm that word duration variation may entail prosodic profile variation, which in turn may be correlated with pronunciation variation.

Then we aimed at studying prosodic parameters on larger units such as phrases or prosodic words. In this purpose we proposed to study simple noun phrases limited to `determiner - noun` bigrams and to compare the corresponding profiles to single word noun profiles. The profiles have been investigated to address the question of prosodic parameter profiles across word boundaries. The measurements pertain to the question whether the mean profile of an n length noun phrase can be different from the profile of an n length noun. The investigated determiners are *le*, *la*, and *les* that correspond to "the" in English. The corresponding canonical pronunciations are /lə/, /la/, and /le/ or /lez/ with its liaison. This study is limited to noun words without final-schwa. The comparison between noun and noun phrase with two speaking styles (prepared/spontaneous) has been showed according to f_0, duration and intensity respectively.

Results pointed out that prosodic cues to word boundaries are less marked in spontaneous face-to-face speech, where speakers and interlocutors may interrupt their conversation at any point to clarify the subject, if ever some unsolvable ambiguity arose. Speaking style may causes differences in f_0 profiles between determiners and nouns in which f_0 values are lower for spontaneous speech (PFC corpus) than for prepared speech (ESTER corpus). However, for both speaking styles, it can be asserted that in noun phrases, the f_0 values start with relatively low values that rise as soon as the first syllable of the following noun which is produced by the speaker. This f_0 rise information between determiner and first noun syllable may be of help to locate word boundaries and to disambiguate homophones such as *déblocage* ('unblocking') and *des blocages* ('blockings') which phonemes are /deblɔkaʒ/. As for duration profiles, similar results are found for the two corpora. Duration profiles do not seem to provide cues to distinguish between items like "*lézard*" and "*les arts*" neither in journalistic speech nor in spontaneous speech. Finally, intensity profiles show that first and final syllables of a noun have almost the same intensity values. Within noun phrases, the values of the determiner syllable, the first and the final noun syllables are also very close. It is interesting to note that the intensity is slightly lower on the determiner than on the first syllable, which might contribute as a cue for word segmentation. Finally, intervocalic f_0 analysis shows that most of `determiners` have a drop f_0 value compared to the preceding vowel and first and final vowels of `nouns` have f_0 rise compared to the preceding vowel. Intervocalic duration profile

results show long intervocalic duration between `determiner` vowel and preceding word vowel highlighting a phrase boundary. These average results indicate that measurable cues contributing to word boundary location can be found in large speech corpora.

Finally, an investigation limited to ambiguous homophone noun phrases was conducted with the initial vowel belonging either to the determinant or to the noun (e.g. *la tante* "the aunt" vs. *l'attente* "the wait" for the phonemes /latãt/). This study allowed us to address the issue of "fine prosodic detail" to discriminate between homophone noun phrases. The related f_0 profiles highlighted a clear difference in average f_0 values with respect to the first syllable: lower if the first vowel belonged to the determiner (*la*); some intermediary value (between low determiner and high first noun syllable) in the special case where the determiner (*l'*) had no vowel and the first vowel belonged already to the noun. This result on data extracted from large corpora was in line with results from psycholinguistic studies [Spinelli *et al.*, 2007; Spinelli *et al.*, 2010] on a small set of controlled material.

Further work

The knowledge of overall prosodic properties may be considered as a first step to the elaboration of word-class specific rules for pronunciation variants. The long term objective is to improve acoustic modeling using acoustic and prosodic parameters to reduce automatic transcription errors. One of the aims was to highlight the prominent role of prosodic features in French. The proposed work may then contribute to demonstrate the effectiveness of large corpus-based studies using ASR tools for extracting and describing prosodic features, in order to obtain knowledge which may contribute to lexical segmentation.

These efforts should ultimately lead to improvements of ASR performances, and its multiple applications (e.g., named-entity, information retrieval, speech understanding, event detection and tracking, and automatic speech translation, etc.). For the time being, a collaboration with colleagues (Sophie Rosset, Marco Dinarelli) is underway to test the usefulness of the proposed acoustic and prosodic features to the localization of focus or/and named-entity within the framework of discriminative classifiers such as conditional random fields (CRF).

A future step would be to implement the current findings in an ASR post-processing approach to improve word boundary locations which ultimately should lead to reduced recognition error rates. On a larger scale, future studies should include more extensive POS sequences, and more detailed analyses of f_0 patterns within syllables. Extending the methodology to other speaking styles and other languages will be taken into consideration, so as to pursue cross-linguistic comparisons of the relevant prosodic parameters on the basis of the findings from the above-mentioned empirical studies. More corpus-based studies are vital to obtain a thorough quantitative and exhaustive linguistic analysis of acoustic and prosodic features involved. We now dispose of a formidable set of tools that allows us to generate new predictions and provide answers to scientific bottlenecks given the available speech corpora. The issues that can be addressed through large-scale corpus-based phonetics studies are of interest to speech engineers, linguists, and cognitive scientists alike.

Part III

Appendix

Appendix A

62 selected attributes

This appendix presents 62 selected attributes for automatic classification of homophone words concerning chapter 4.

A.1 Intra-phonemic attributes: 40 attributes

duration (1): segmental duration;
voicing ratio (4): mean, begin, middle, end;
f_0 (7): mean, begin, middle, end, Δbegin-middle, Δmiddle-end, Δbegin-end;
F1 (7): mean, begin, middle, end, Δbegin-middle, Δmiddle-end, Δbegin-end;
F2 (7): mean, begin, middle, end, Δbegin-middle, Δmiddle-end, Δbegin-end;
F3 (7): mean, begin, middle, end, Δbegin-middle, Δmiddle-end, Δbegin-end;
intensity (7): mean, begin, middle, end, Δbegin-middle, Δmiddle-end, Δbegin-end.

A.2 Inter-phonemic attributes: 22 attributes

duration (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
f_0 (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
F1 (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
F2 (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
F3 (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
Intensity (3): Δpreceding-target, Δtarget-following, Δpreceding-following;
Pauses[1] (4): left pause of which a target word starting with a vowel, left pause of which a target word starting with a consonant then followed by a vowel, right pause of which a target word finishing with a vowel, right pause of which a target word finishing with a consonant preceded by a vowel.

[1] In fact, our target words starting with a vowel, just one left pause is considered in this study. And one right pause is also taken because our target words are not finished by a consonant. Even the word *est* finished by a consonant in the case of liaison, there is no pause between the phoneme /t/ and a follwing vowel of a follwoing word.

Appendix B

Homophone classification results

Table B.1: Homophone (*et*/*est* pair) classification results (in %) of ESTER corpus in terms of algorithms and attribute types. The employed attribute number for each category is demonstrated in parentheses.

ESTER corpus		*et* vs. *est*						
Algorithms \ Attributes		15 best (15)	15 all best (15)	all (62)	formants (30)	prosody (32)	intra- (40)	inter- (22)
bayes	BayesNet	69.2459	68.2995	67.5198	63.4635	66.005	62.8504	68.9483
	NaiveBayes	65.2818	63.8831	60.8208	54.1724	63.0349	56.949	64.4426
	NaiveBayesUpdateable	65.2818	63.8831	60.8208	54.1724	63.0349	56.949	64.4426
functions	Logistic	73.2189	71.0374	76.3615	64.5497	75.2664	68.0942	71.0523
	MultilayerPerceptron	77.1293	74.689	77.6739	67.4513	77.3049	72.6891	72.7963
	RBFNetwork	67.246	68.4334	65.9901	61.3654	65.5586	62.2642	66.4216
	SMO	69.1239	67.374	75.8348	64.7104	71.6088	68.0198	67.2519
	SimpleLogistic	72.9808	71.0196	76.3437	64.4604	75.1562	67.8977	71.0315
	VotedPerceptron	57.8448	57.5323	63.5647	62.9873	57.1514	61.6213	60.919
rules	ConjunctiveRule	67.246	67.246	67.246	57.1722	67.246	58.7286	67.246
	DecisionTable	73.8141	73.0343	74.436	64.7015	74.3795	69.3709	73.4004
	JRip	76.9359	75.8645	78.6233	66.2401	78.5846	71.7338	75.058
	NNge	71.6088	70.2637	70.4035	60.8238	71.7606	64.8503	67.8531
	OneR	56.2228	56.2734	56.3389	56.1246	54.8539	55.9461	56.2288
	PART	76.8972	75.6175	77.4299	66.133	77.9329	71.4005	74.32
	Ridor	72.9004	70.9928	75.683	62.4665	74.7961	68.4245	71.1267
	ZeroR	56.943	56.943	56.943	56.943	56.943	56.943	56.943
trees	ADTree	75.3199	74.3408	76.2841	65.4277	76.2812	69.4244	74.2902
	DecisionStump	67.246	67.246	67.246	56.943	67.246	58.3715	67.246
	J48	76.5877	75.4598	76.4091	66.011	77.1353	70.4244	74.7188
	LMT	77.5847	76.0907	79.7512	67.2043	79.2066	73.2427	75.6622
	NBTree	76.805	74.8944	78.3257	64.1271	77.4894	70.5375	73.9986
	REPTree	76.1502	75.2991	78.1739	64.5735	78.1055	71.2934	74.4152
	RandomForest	76.924	75.0967	79.3494	66.755	79.4923	72.5939	75.7038
	RandomTree	66.9871	66.1092	64.7253	59.5232	66.8799	62.8474	65.1062

Table B.2: Homophone (*et*/*est* pair) classification results (in %) of PFC corpus in terms of algorithms and attribute types. The employed attribute number for each category is demonstrated in parentheses.

PFC corpus		*et* vs. *est*						
Algorithms \ Attributes		15 best (15)	15 all best (15)	all (62)	formants (30)	prosody (32)	intra- (40)	inter- (22)
bayes	BayesNet	76.6759	76.8628	72.1001	64.1325	76.5067	66.4382	74.4414
	NaiveBayes	72.4829	73.9607	69.0822	61.3549	75.3761	64.907	77.299
	NaiveBayesUpdateable	72.4829	73.9607	69.0822	61.3549	75.3761	64.907	77.299
functions	Logistic	80.4148	80.13	80.7798	63.2689	80.6374	69.5896	80.1122
	MultilayerPerceptron	80.5395	80.2457	80.1656	66.3759	79.7828	70.2662	79.2932
	RBFNetwork	72.6164	72.474	68.0851	61.1413	72.1802	64.631	74.5215
	SMO	79.9786	79.9786	80.2368	61.0166	80.2368	69.8834	80.2368
	SimpleLogistic	80.4505	80.3703	80.6463	63.1532	80.6374	70.4887	80.1923
	VotedPerceptron	59.7347	60.3045	62.9485	62.8861	60.9988	61.622	56.8147
rules	ConjunctiveRule	79.9786	79.9786	79.9786	61.1502	79.9786	66.7052	79.9786
	DecisionTable	80.5573	80.2902	80.6552	64.8358	80.7887	71.2721	80.3614
	JRip	80.2546	79.943	81.225	65.6103	80.8155	70.8893	80.7442
	NNge	76.4177	76.302	75.7233	61.0255	76.4711	65.4678	77.5394
	OneR	79.9786	79.9786	79.9786	54.696	79.9786	58.0789	79.9786
	PART	80.3169	80.3614	80.9223	64.8981	80.8155	69.8211	80.6463
	Ridor	79.1774	78.9994	79.1329	62.4054	79.2753	68.3967	78.7501
	ZeroR	55.5239	55.5239	55.5239	55.5239	55.5239	55.5239	55.5239
trees	ADTree	79.8896	80.673	80.4416	65.6014	80.2635	70.0792	80.5039
	DecisionStump	79.9786	79.9786	79.9786	61.1769	79.9786	66.5628	79.9786
	J48	80.6196	80.2991	79.5513	65.3877	80.2724	69.6163	79.8095
	LMT	81.3496	80.3703	83.0499	66.5717	82.3467	71.6995	81.2161
	NBTree	79.8451	78.5365	80.2279	64.1414	80.86	68.6816	80.8956
	REPTree	80.2368	79.7472	81.2517	64.1681	81.314	69.6786	79.9875
	RandomForest	79.9697	78.4652	81.314	65.4767	81.4386	69.7142	80.3614
	RandomTree	71.7262	71.0763	65.7082	59.3964	69.4561	60.9988	67.9516

Table B.3: Homophone (*à/a* pair) classification results (in %) of ESTER corpus in terms of algorithms and attribute types. The employed attribute number for each category is demonstrated in parentheses.

ESTER corpus		*à* vs. *a*						
Algorithms \ Attributes		15 best (15)	15 all best (15)	all (62)	formants (30)	prosody (32)	intra- (40)	inter- (22)
bayes	BayesNet	59.2926	59.8595	54.5148	58.5021	56.2187	52.37	68.2057
	NaiveBayes	48.7827	49.0063	48.7764	48.2819	49.1355	48.3386	45.0128
	NaiveBayesUpdateable	48.7827	49.0063	48.7764	48.2819	49.1355	48.3386	45.0128
functions	Logistic	66.8514	65.7743	69.5411	67.0499	67.0908	66.7381	66.5081
	MultilayerPerceptron	67.1097	67.4624	71.3615	68.949	68.6656	68.5585	68.127
	RBFNetwork	64.7381	66.0483	64.4295	66.3853	64.3791	64.3791	64.3791
	SMO	64.3791	64.3791	64.4452	64.3791	64.4106	64.3791	64.4106
	SimpleLogistic	67.0058	65.768	69.5317	67.0467	67.1254	66.8073	66.4955
	VotedPerceptron	65.8152	64.1649	66.9396	66.9585	63.5098	66.3759	60.7382
rules	ConjunctiveRule	67.2861	64.3791	68.2687	64.3791	68.2309	64.3791	68.1931
	DecisionTable	69.5096	67.2073	69.8781	67.053	69.7206	67.3585	69.327
	JRip	70.0482	67.9947	72.4198	68.3254	71.7017	68.6845	70.7159
	NNge	62.2626	62.0642	63.1697	62.6279	61.7177	61.3052	63.3303
	OneR	61.765	61.8784	61.7429	61.7681	58.732	61.8091	59.4879
	PART	70.3568	68.5868	71.3048	67.6262	71.1348	68.0262	70.5017
	Ridor	68.1049	65.831	69.9852	66.7412	69.4403	66.7412	69.1285
	ZeroR	64.3791	64.3791	64.3791	64.3791	64.3791	64.3791	64.3791
trees	ADTree	69.7364	67.916	70.9458	67.2073	69.9285	67.7396	69.9569
	DecisionStump	68.2498	64.3791	68.2498	64.3791	68.2498	64.3791	68.2498
	J48	70.0135	68.6593	68.9837	67.2546	70.3222	67.1884	70.3568
	LMT	70.8986	68.9238	72.9237	68.4892	72.2875	68.8514	71.0592
	NBTree	68.7065	67.7711	70.8891	66.527	70.1868	67.1569	69.8466
	REPTree	69.8246	67.4183	71.4245	67.4939	70.5868	67.9916	69.8466
	RandomForest	69.4655	66.2845	72.4324	67.4656	71.1285	68.2309	70.2435
	RandomTree	62.8264	60.4674	62.1177	61.0028	62.461	60.5304	63.0374

Table B.4: Homophone (*à/a* pair) classification results (in %) of PFC corpus in terms of algorithms and attribute types. The employed attribute number for each category is demonstrated in parentheses.

PFC corpus		*à* vs. *a*						
Algorithms \ Attributes		15 best (15)	15 all best (15)	all (62)	formants (30)	prosody (32)	intra- (40)	inter- (22)
bayes	BayesNet	63.2502	58.2324	61.9387	58.9736	61.0834	57.4911	64.8182
	NaiveBayes	53.9701	56.1511	55.2245	53.3571	54.3977	53.7135	51.2188
	NaiveBayesUpdateable	53.9701	56.1511	55.2245	53.3571	54.3977	53.7135	51.2188
functions	Logistic	63.1361	60.0428	66.1012	62.6515	61.7819	59.8147	61.2972
	MultilayerPerceptron	62.0242	58.7028	63.6493	61.7249	59.9002	58.6743	61.397
	RBFNetwork	55.5524	60.0428	54.9394	59.4155	55.2245	55.3243	54.7541
	SMO	62.794	58.3892	66.2295	62.181	61.1689	60.0285	61.1262
	SimpleLogistic	63.1932	59.9287	65.7306	62.2381	61.5823	59.4726	60.8268
	VotedPerceptron	60	54.7398	61.1404	61.0549	55.082	58.4462	54.0413
rules	ConjunctiveRule	56.0086	56.0656	56.0371	56.7498	56.0371	52.5731	56.5502
	DecisionTable	65.8161	59.8574	65.0036	59.273	65.6165	58.3179	65.4882
	JRip	66.6429	61.9102	66.6714	59.7149	67.4127	59.2017	65.2887
	NNge	60.0713	57.4911	59.216	58.3892	58.9451	56.1796	58.8738
	OneR	53.799	54.0841	51.7177	52.0029	52.7014	54.7113	52.0314
	PART	65.0178	61.283	63.9629	58.3607	64.9038	58.4462	64.4619
	Ridor	63.0221	57.149	63.5353	57.4198	62.851	56.2651	60.7128
	ZeroR	51.9886	51.9886	51.9886	51.9886	51.9886	51.9886	51.9886
trees	ADTree	65.6736	61.9244	67.2559	60.3706	66.928	59.7719	65.1746
	DecisionStump	56.5502	56.5502	56.5502	57.6764	56.5502	54.0413	56.5502
	J48	65.6023	61.6821	63.9344	58.8311	65.3885	57.2915	63.8061
	LMT	67.4982	62.01	69.3799	62.3378	67.6978	60.0143	65.8874
	NBTree	64.9323	58.66	66.5859	58.9166	66.0584	57.9473	64.8182
	REPTree	64.02	59.7719	66.0727	59.4013	66.2153	58.3321	65.2031
	RandomForest	65.8589	59.886	65.2174	60.1996	65.5167	59.0164	65.5167
	RandomTree	58.3036	56.4932	56.9351	54.5545	57.5624	54.3122	58.2038

Appendix C

Average prosodic parameters

This appendix presents average prosodic parameters (f_0, intensity, and duration) of the audio data concerning chapter 5. Hence only male speakers' data were computed.

C.1 Fundamental frequency and intensity

The averages of f_0 (in green line) and intensity (in black line) are illustrated in Figure C.1 (ESTER corpus) and Figure C.2 (PFC corpus).

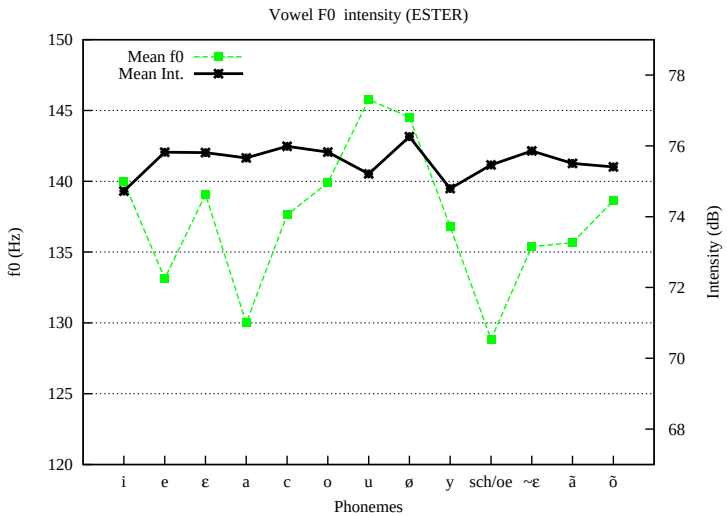


Figure C.1: Mean f_0 and intensity of male speakers in the ESTER corpus. /c/ means opend o /ɔ/, sch/œ means /ə/ and /œ/ phonemes, ∼ɛ means a nasal vowel $\tilde{\varepsilon}$.

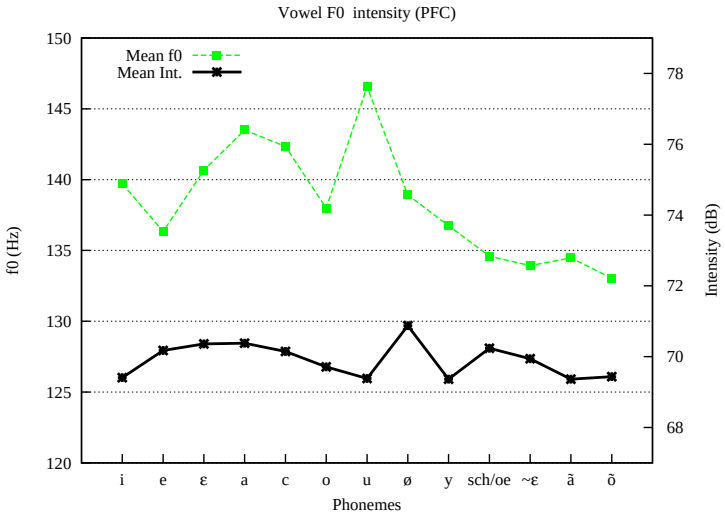


Figure C.2: Mean f_0 and intensity of male speakers in the PFC corpus. /c/ means opend o /ɔ/, sch/œ means /ə/ and /œ/ phonemes, ∼ɛ means a nasal vowel ɛ̃.

C.2 Duration

Duration average and distributions in four categories (short 30-40 ms, medium 50-60 ms, long 70-90 ms, extra-long more than 100 ms) are illustrated: vocalic duration in Figure C.3 (ESTER corpus) and Figure C.4 (PFC corpus), and consonant duration in Figure C.5 (ESTER corpus) and in Figure C.6 (PFC corpus).

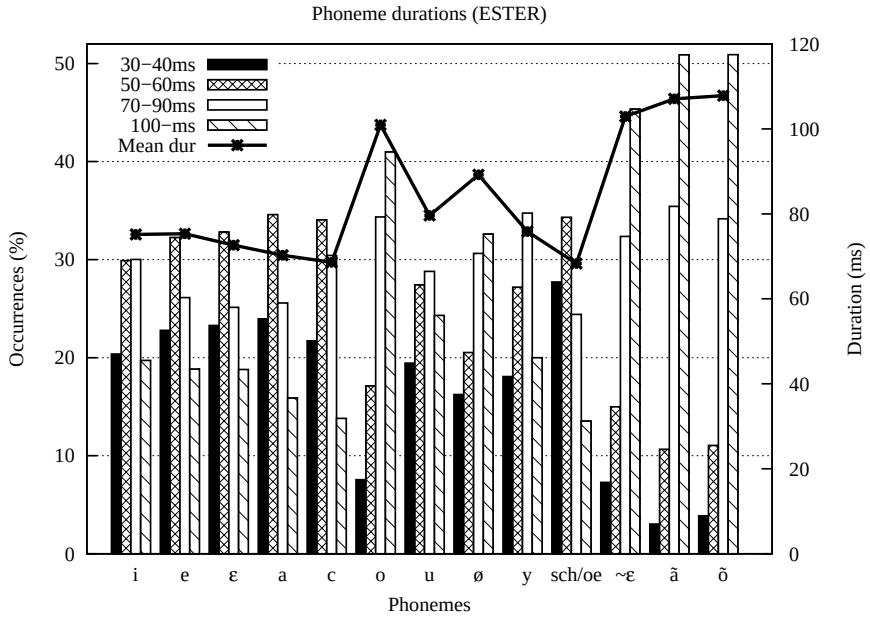


Figure C.3: French vowel durations of male speakers in the ESTER corpus. /c/ means opend o /ɔ/, sch/œ means /ə/ and /œ/ phonemes, ∼ɛ means a nasal vowel ɛ̃.

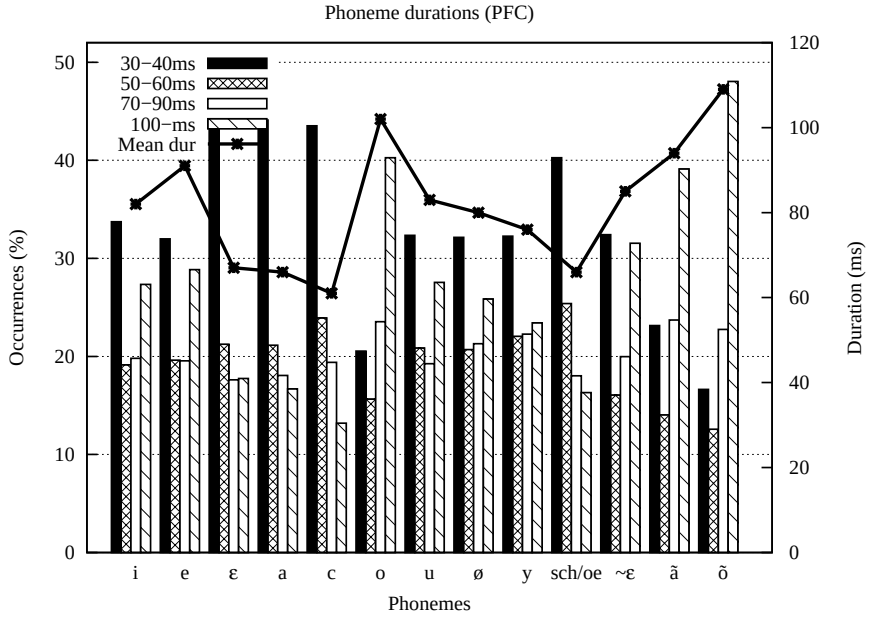


Figure C.4: French vowel durations of male speakers in the PFC corpus. /c/ means opend o /ɔ/, sch/œ means /ə/ and /œ/ phonemes, ∼ɛ means a nasal vowel ɛ̃.

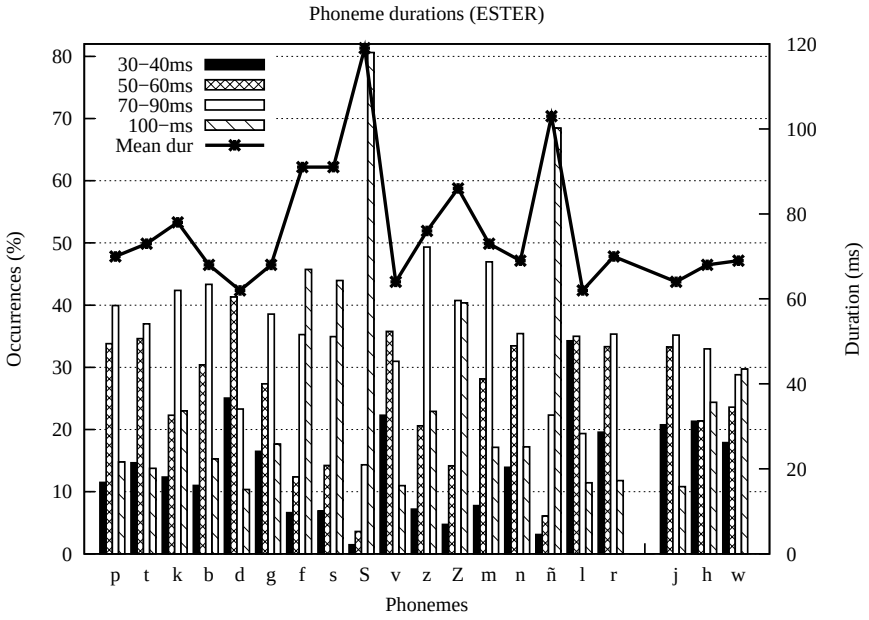


Figure C.5: French consonant and semi-vowel durations of male speakers in the ESTER corpus. /S/ means /ʃ/, /Z/ is /ʒ/, /ñ/ signifies /ɲ/, and /h/ is a semi-vowel /ɥ/.

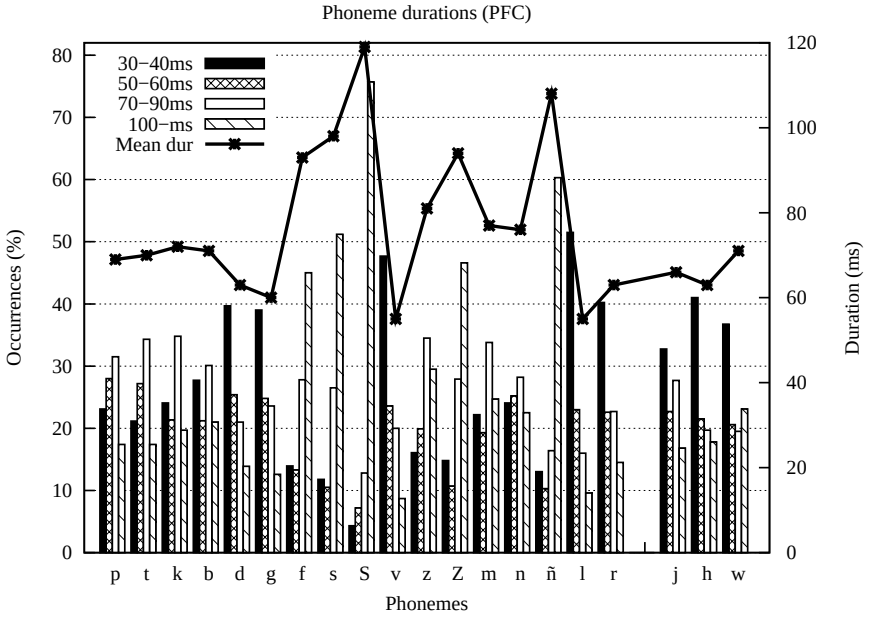


Figure C.6: French consonant and semi-vowel durations of male speakers in the PFC corpus. /S/ means /ʃ/, /Z/ is /ʒ/, /ñ/ signifies /ɲ/, and /h/ is a semi-vowel /ɥ/.

Appendix D

f_0 Profiles in Terms of POS

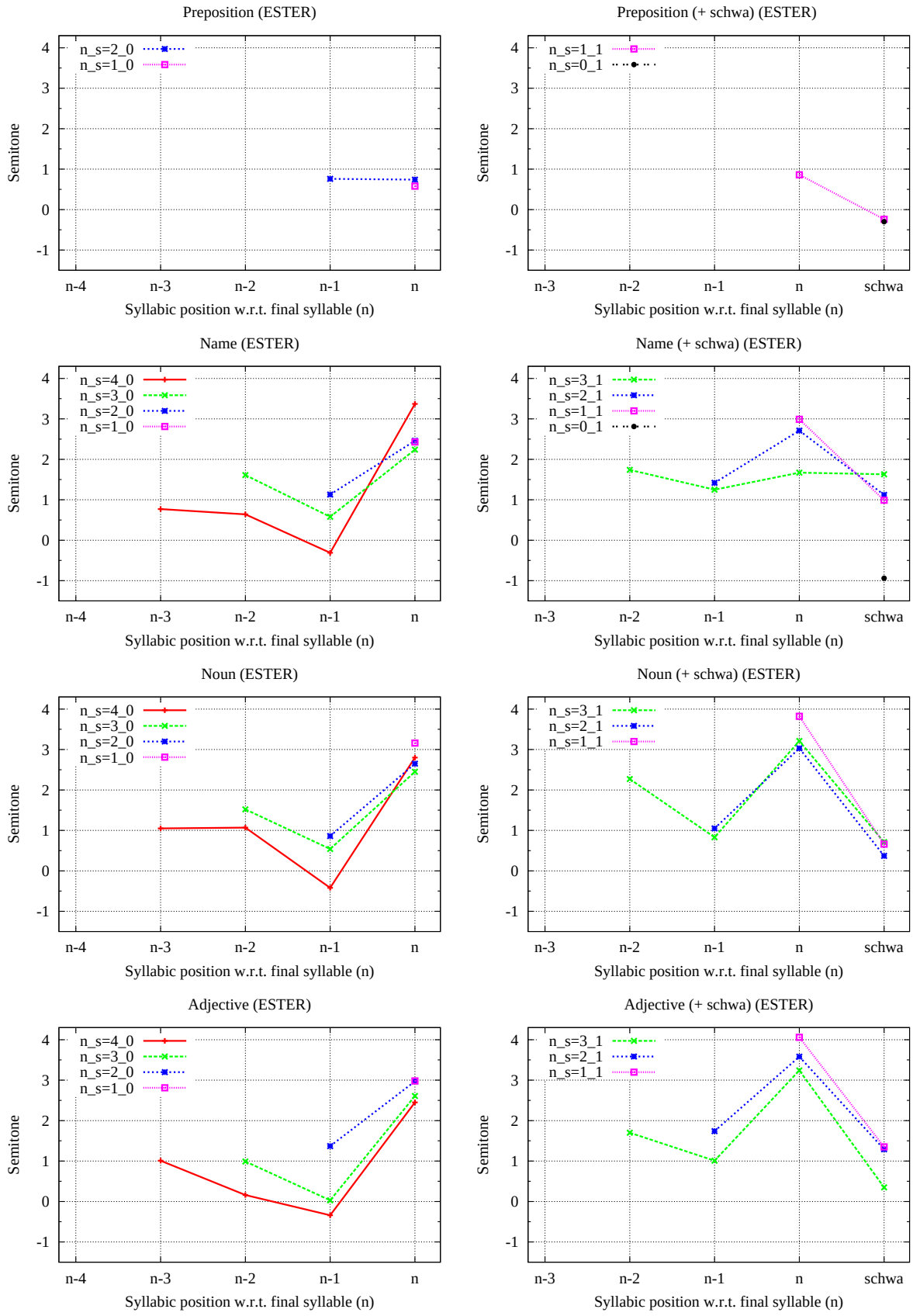


Figure D.1: Profils of average f_0 in terms of POS.

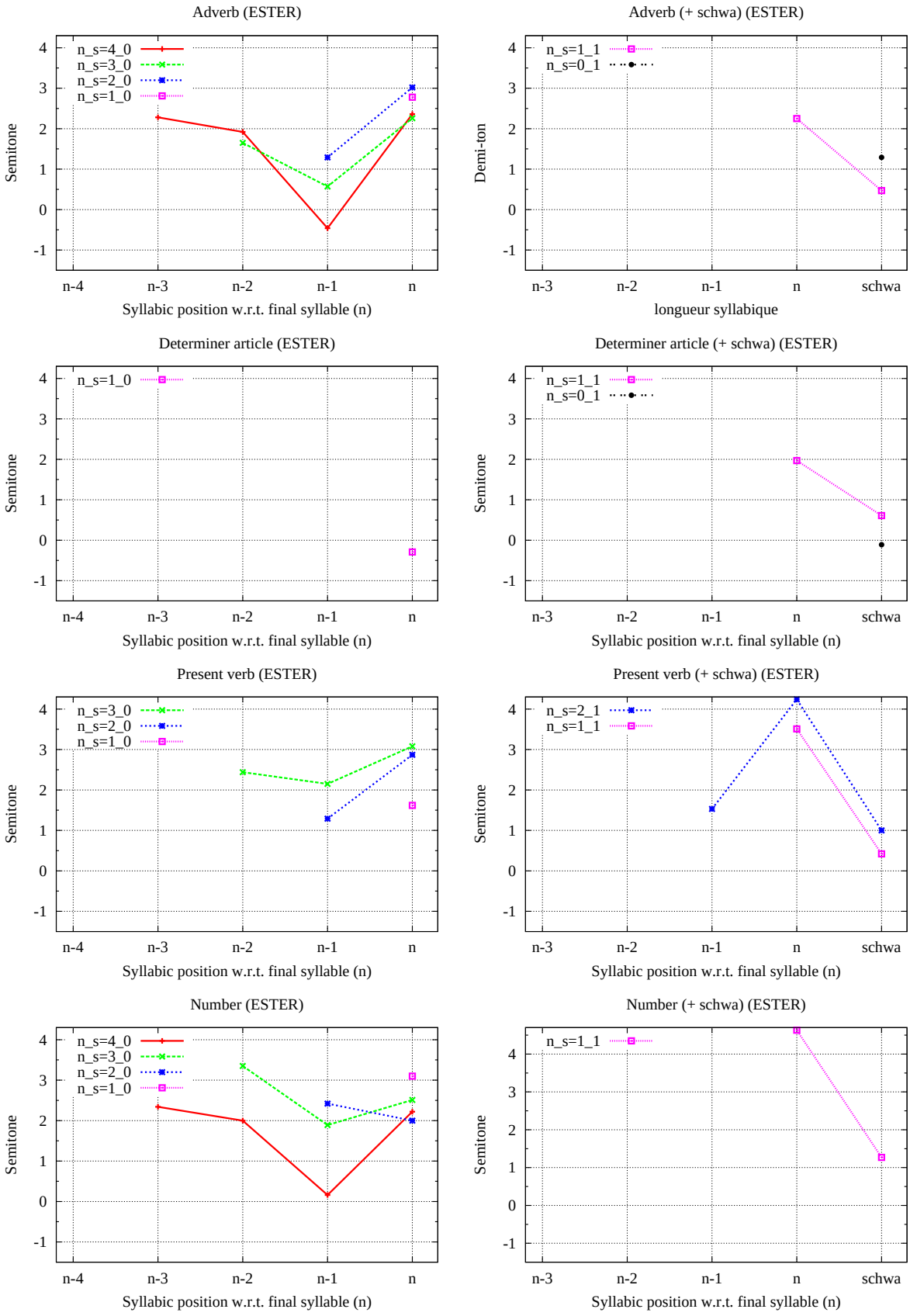


Figure D.2: Profils of average f_0 in terms of POS. Part2

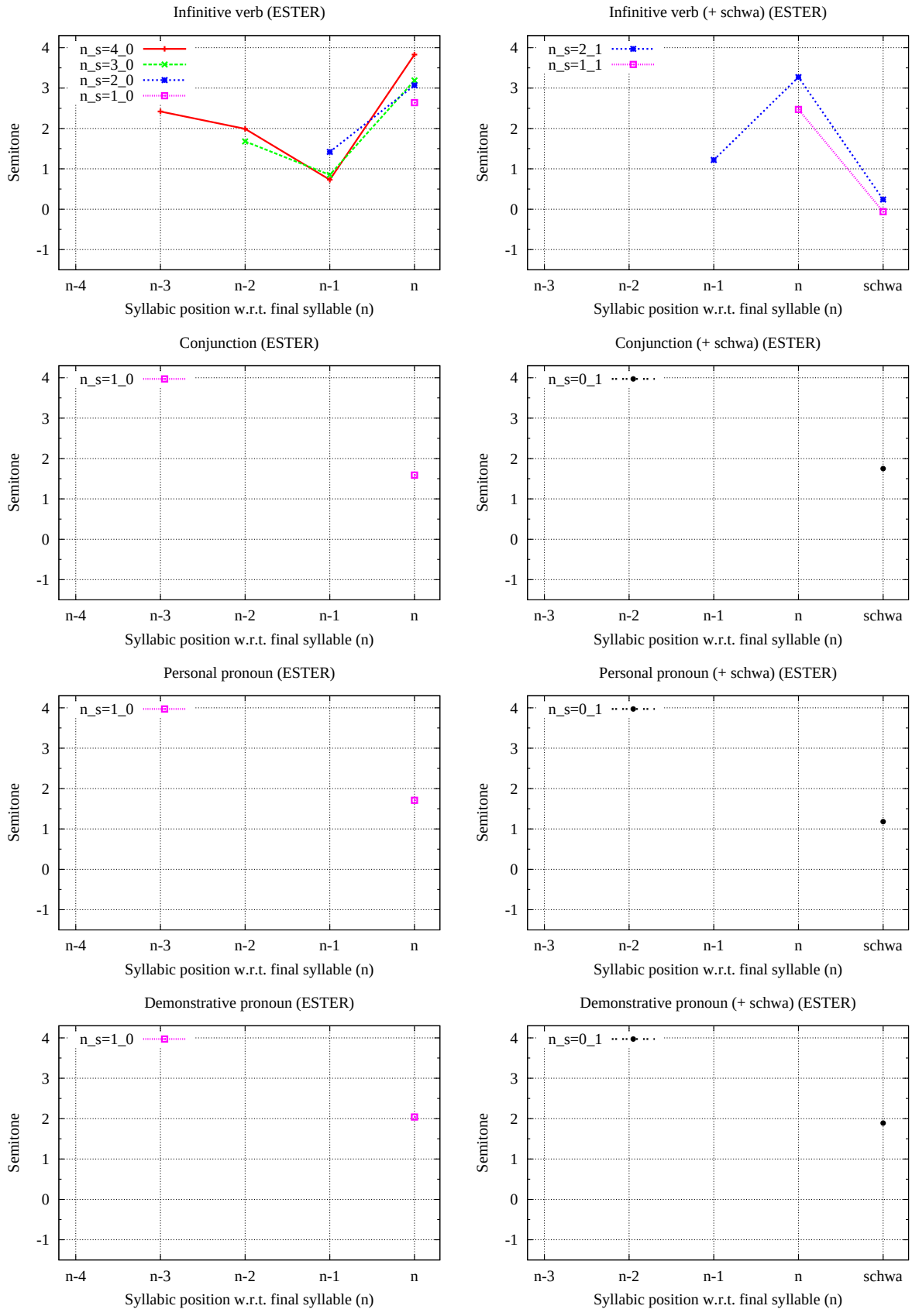


Figure D.3: Profils of average f_0 in terms of POS. Part3

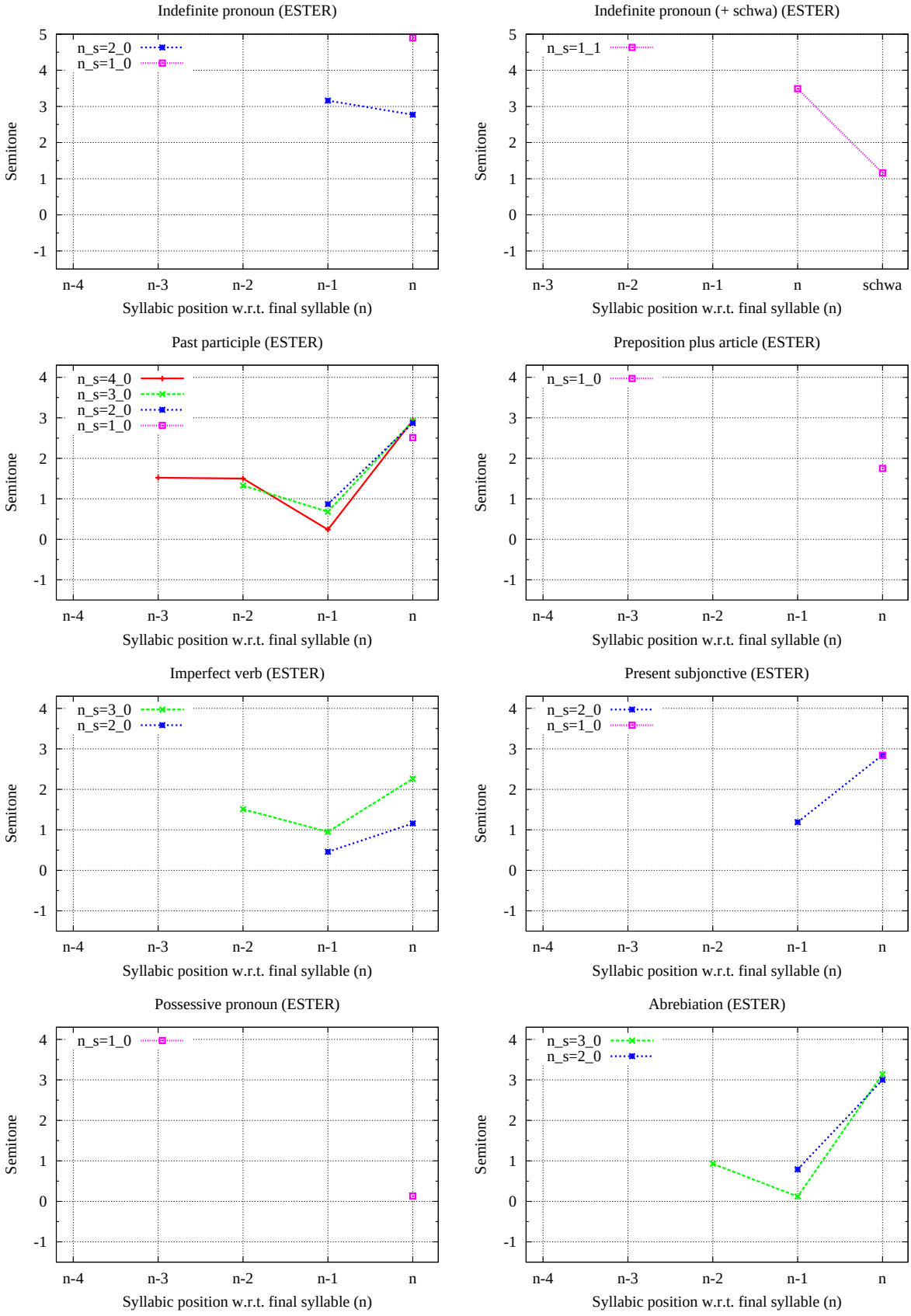


Figure D.4: Profils of average f_0 in terms of POS. Part4

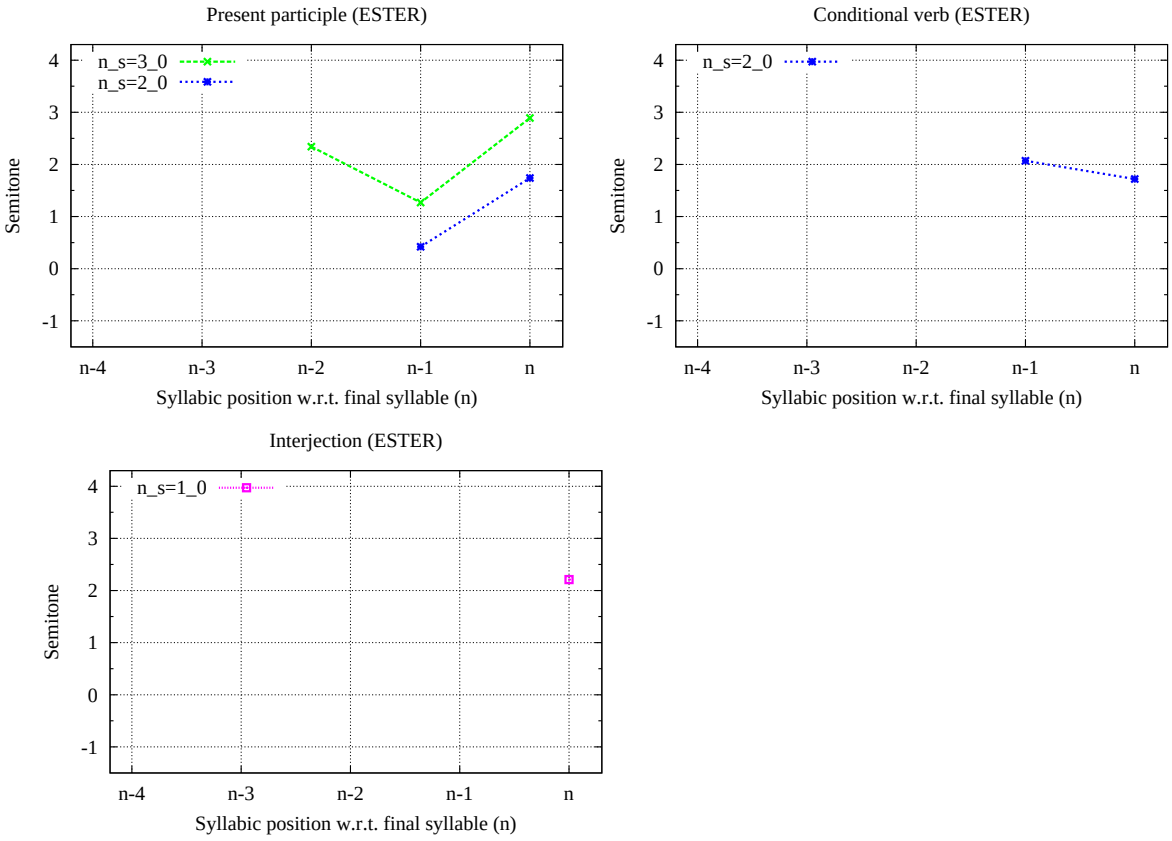


Figure D.5: Profils of average f_0 in terms of POS. Part5

Appendix E

f_0 Profiles: PFC text reading

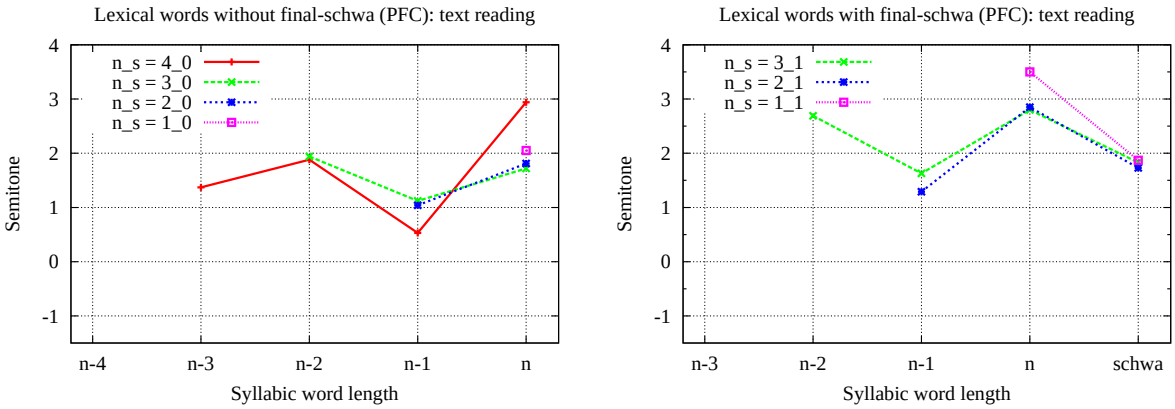


Figure E.1: Lexical word profiles of average f_0 for the PFC corpus (text reading).

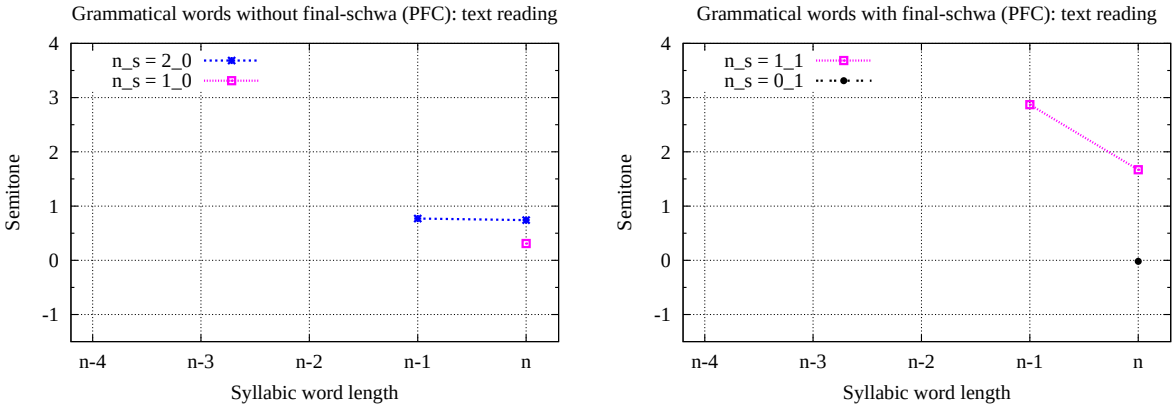


Figure E.2: Grammatical word profiles of average f_0 for the PFC corpus (text reading).

Author's publications

[1] I. Vasilescu, R. Nemoto and M. Adda-Decker. Vocalic hesitations vs vocalic systems: a cross-language comparison. In *16th International Congress of Phonetic Science*, Saarbrücken, Germany, August 6–10, 2007.

[2] R. Nemoto, M. Adda-Decker and I. Vasilescu. Fouille de données audio pour la classification automatique de mots homophones. In *Extraction et gestion des connaissances* (EGC'2008), pages 445–456, Sophia-Antipolis, France, 2008.

[3] R. Nemoto, M. Adda-Decker and I. Vasilescu. Mots fréquents homophones en français : analyse acoustique et classification automatique par fouille de données. In *Journées d'Étude sur la Parole*, pages 337–340, Avignon, France, 2008.

[4] R. Nemoto, I. Vasilescu and M. Adda-Decker. Speech errors on frequently observed homophones in French: Perceptual evaluation vs automatic classification. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation* (LREC'08), Marrakech, Morocco, May 28–30, 2008.

[5] I. Vasilescu, M. Adda-Decker and R. Nemoto. Caractéristiques acoustiques et prosodiques des hésitations vocaliques dans trois langues. *Traitement Automatique des Langues*, 49(3):199–228, 2008.

[6] M. Adda-Decker, J. Durand and R. Nemoto. Stratégies de démarcation du mot en français : une étude expériementale sur grand corpus. In *Proceedings of the Journées d'Études Linguistiques de Nantes*, Nantes, France, 18–19 juin, 2009.

[7] R. Nemoto, M. Adda-Decker and J. Durand. Investigation of lexical f_0 and duration patterns in French using large broadcast news speech corpora. In *Proceedings of Speech Prosody*, Chicago, USA, May 11–14, 2010.

[8] R. Nemoto, M. Adda-Decker and J. Durand. Word boundaries in French: Evidence from large speech corpora. In European Language Resources Association (ELRA), editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10), Valletta, Malta, May 19–21, 2010.

[9] R. Nemoto, M. Adda-Decker and J. Durand. Démarcation lexicale en français : profils prosodiques sur grand corpus. In *Journées d'Étude sur la Parole*, Mons, Belgique, 25–28 mai, 2010.

References

[Abercrombie, 1967] D. Abercrombie. *Elements of general phonetics*. Edinburgh University Press, 1967.

[Adda *et al.*, 1997] G. Adda, M. Adda-Decker, J.-L. Gauvain, and L. Lamel. Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 35–40, Avignon, France, 1997.

[Adda-Decker and Lamel, 1998] M. Adda-Decker and L. Lamel. Pronunciation variants across systems, languages and speaking style. In H. Strik, J.M. Kessens, and M. Wester, editors, *Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 1–6, Rolduc, The Netherlands, 1998.

[Adda-Decker and Lamel, 1999] M. Adda-Decker and L. Lamel. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29:83–98, 1999.

[Adda-Decker and Lamel, 2005] M. Adda-Decker and L. Lamel. Do speech recognizers prefer female speakers? In *Proceedings of Interspeech*, pages 2205–2208, Lisbon, Portugal, 2005.

[Adda-Decker and Snoeren, 2011] M. Adda-Decker and N. Snoeren. Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39(3):261–270, 2011.

[Adda-Decker *et al.*, 1999a] M. Adda-Decker, G. Adda, J.-L. Gauvain, and L. Lamel. Large vocabulary speech recognition in French. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1999.

[Adda-Decker *et al.*, 1999b] M. Adda-Decker, Ph. Boula de Mareüil, and L. Lamel. Pronunciation variants in French: schwa & liaisons. In *14th International Congress of Phonetic Sciences*, pages 2239–2242, San Francisco, USA, 1999.

[Adda-Decker *et al.*, 2002] M. Adda-Decker, Ph. Boula de Mareüil, G. Adda, and L. Lamel. Investigating syllabic structure and its variation in speech from French radio interviews. In *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pages 89–94, Aspen Lodge, USA, 2002.

[Adda-Decker *et al.*, 2003] M. Adda-Decker, B. Habert, C. Barras, G. Adda, Ph. Boula de Mareüil, and P. Paroubek. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In *DiSS'03*, pages 67–70, Göteborg, Sweden, 2003.

[Adda-Decker *et al.*, 2005] M. Adda-Decker, Ph. Boula de Mareüil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, 46(2):119–139, 2005.

[Adda-Decker *et al.*, 2008] M. Adda-Decker, C. Gendrot, and N. Nguyen. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues (TAL)*, 49(3):13–46, 2008.

[Adda-Decker *et al.*, 2011] M. Adda-Decker, I. Vasilescu, N. Snoeren, D. Yahia, and L. Lamel. Towards exploring linguistic variation in ASR errors: paradigm and tool for perceptual experiments. In *New Tools and Methods for very-large-scale phonetics research*, Pennsylvania, USA, 2011.

[Adda-Decker, 2006] M. Adda-Decker. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Journées d'étude sur la Parole*, pages 389–400, Dinard, France, 2006.

[Adda-Decker, 2007] M. Adda-Decker. Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole. In *Actes des 5èmes Journées Linguistiques de Nantes*, pages 211–216, Nantes, France, 2007.

[Atal and Hanauer, 1971] B. S Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, 1971.

[Aubergé and Bailly, 1995] V. Aubergé and G. Bailly. Generation of intonation: a global approach. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2065–2068, Madrid, Spain, 1995.

[Averbuch *et al.*, 1987] A. Averbuch, L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jelinek, B. Lewis, R. Mercer, J. Moorhead, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, P. Spinelli, D. Van Compernolle, and H. Wilkens. Experiments with the Tangora 20,000 word speech recognizer. pages 701–704, 1987.

[Bagou *et al.*, 2002] O. Bagou, C. Fougeron, and U. H. Frauenfelder. Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. In B. Bel and I. Marlien, editors, *Proceedings of Speech Prosody*, pages 59–62, Aix-en-Provence, France, 2002.

[Bailly, 1983] G. Bailly. *Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique : établissement d'un modèle de génération*. PhD thesis, Institut national polytechnique de Grenoble, 1983.

[Banel and Bacri, 1994] M.-H. Banel and N. Bacri. On metrical patterns and lexical parsing in French. *Speech Communication*, 15:115–126, 1994.

[Barker and Cooke, 2007] Jon Barker and Martin Cooke. Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417, 2007.

[Bates and Ostendorf, 2002] R. Bates and M. Ostendorf. Modeling pronunciation variation in conversational speech using prosody. In *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pages 42–47, Aspen Lodge, USA, 2002.

[Béchet *et al.*, 1999] F. Béchet, A. Nasr, T. Spriet, and R. de Mori. Large span statistical language models: application to homophone disambiguation for large vocabulary speech recognition in French. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999.

[Beckman and Edwards, 1990] M. E. Beckman and J. Edwards. Lengthening and shortening and the nature of prosodic constituency. In J. Kingston and M. E. Beckman, editors, *Laboratory Phonology I*, pages 152–178. Cambridge University Press, Cambridge, 1990.

[Beckman and Pierrehumbert, 1986] M. E. Beckman and J. Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook*, 1986.

[Bell *et al.*, 2002] A. Bell, M. L. Gregory, J. M. Brenier, D. Jurafsky, A. Ikeno, and C. Girand. Which predictability measures affect content word durations? In *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pages 1–5, Aspen Lodge, USA, 2002.

[Bell *et al.*, 2009] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60:92–111, 2009.

[Binnenpoorte *et al.*, 2005] D. Binnenpoorte, C. Cucchiarini, L. Boves, and H. Strik. Multiword expressions in spoken language: an exploratory study on pronunciation variation. *Computer Speech and Language*, 19(4):433–449, 2005.

[Bloch, 1950] B. Bloch. Studies in colloquial Japanese IV: Phonemics. *Language*, 26:86–125, 1950.

[Boersma and Weenink, 2008] P. Boersma and D. Weenink. *Praat: doing phonetics by computer*, 2008. www.praat.org.

[Boula de Mareüil *et al.*, 2001] Ph. Boula de Mareüil, C. d'Alessandro, F. Beaugendre, and A. Lacheret-Dujour. Une grammaire en tronçons appliquée à la génération de la prosodie. *Traitement Automatique des Langues*, 42(1):223–252, 2001.

[Boula de Mareüil *et al.*, 2003] Ph. Boula de Mareüil, M. Adda-Decker, and V. Gendner. Liaisons in French: a corpus-based study using morpho-syntactic information. In *15th International Congress of Phonetic Sciences*, Barcelone, 2003.

[Bürki *et al.*, 2007] A. Bürki, C. Fougeron, and C. Gendrot. On the categorical nature of the process involved in schwa elision in French. In *Interspeech*, 2007.

[Campione, 2001] E. Campione. *Étiquetage semi-automatique de la prosodie dans les corpus oraux: algorithmes et méthodologie*. PhD thesis, Université Aix-Marseille I, 2001.

[Candea, 2000] M. Candea. *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'«hésitation» en français oral spontané. Étude sur un corpus de récits en classe de français*. PhD thesis, Université Paris III, Paris, France, 2000.

[Carré and Hombert, 2002] R. Carré and J. M. Hombert. Variabilité phonétique en production et perception de parole : stratégies individuelles. In J. Lautrey, B. Mazoyer, and P. van Geert, editors, *Invariants et Variabilité dans les Sciences Cognitives*. Presses de la Maison des Sciences de l'Homme, Paris, 2002.

[Chao, 1930] Y.-R. Chao. A System of tone letters. *La Maître phonétique*, 45:24–27, 1930.

[Coy and Barker, 2007] A. Coy and J. Barker. An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Communication*, 49(5):384–401, 2007.

[Crystal, 1997] D. Crystal. *A dictionary of linguistics and phonetics*. Blackwell, 1997.

[Cutler and Carter, 1987] A. Cutler and D. M. Carter. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2:133–142, 1987.

[Cutler and Norris, 1988] A. Cutler and D. Norris. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14:113–121, 1988.

[Cutler *et al.*, 1997] A. Cutler, D. Dahan, and W. Van Donselaar. Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, 40(2):141–201, 1997.

[Cutler, 1991] A. Cutler. Linguistic rhythm and speech segmentation. In J. Sundberg, L. Nord, and R. Carlson, editors, *Music, Language, Speech and Brain*, pages 157–1166, 1991.

[Cutler, 1997] A. Cutler. The syllable's role in the segmentation of stress languages. *Language & Cognitive Processes*, 12(5/6):839–845, 1997.

[Cutler, 2005] A. Cutler. The lexical statistics of word recognition problems caused by L2 phonetic confusion. In *Interspeech*, Lisbon, Portugal, 2005.

[Dahan *et al.*, 2001] D. Dahan, J. Magnuson, and M. Tanenhaus. Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, 42:317–367, 2001.

[Das and Picheny, 1996] S. K. Das and M. A. Picheny. Issues in practical large vocabulary isolated word recognition: The IBM Tangora system. *Automatic Speech and Speaker Recognition Advanced Topics*, pages 457–479, 1996.

[Davis *et al.*, 1952] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24(6):627–642, 1952.

[Delais-Roussarie and Durand, 2003] E. Delais-Roussarie and J. Durand. *Corpus et variation en phonologie du français : méthodes et analyses*. Presses Universitaires du Mirail, Toulouse, 2003.

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

[Delattre, 1966a] P. Delattre. L'accent final en français : accent d'intensité, accent de hauteur, accent de durée. In *Studies in French and Comparative Phonetics*, pages 65–68. The Hague, Mouton & co., 1966.

[Delattre, 1966b] P. Delattre. Les dix intonations de base du français. *French review*, 40(1):1–14, 1966.

[Denes, 1959] P. Denes. The design and operation of the mechanical speech recognizer at University College London. *Journal of the British Institution of Radio Engineers*, 19(4):219–234, 1959.

[Deshmukh *et al.*, 1996] N. Deshmukh, R.J. Dunca, A. Ganapathiraju, and J. Picone. Benchmarking human performance for continous speech recognition. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1996.

[Di Cristo and Hirst, 1986] A. Di Cristo and D. J. Hirst. Modelling french micromelody : analysis and synthesis. *Phonetica*, 43:11–30, 1986.

[Di Cristo, 1998] A. Di Cristo. Intonation in French. In D. Hirst and A. Di Cristo, editors, *Intonation Systems: A Survey of Twenty Languages*, pages 195–218. Cambridge University Press, Cambridge, 1998.

[Doddington and Schalk, 1981] G. Doddington and T. Schalk. Speech recognition: turning theory to practice. In *IEEE spectrum 18*, pages 26–32, 1981.

[Dolmazon *et al.*, 1997] J.M. Dolmazon, F. Bimbot, G. Adda, M. El Béze, J. C. Caërou, J. Zeiliger, and M. Adda-Decker. Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 13–18, Avignon, France, April 1997.

[Duez, 1982] D. Duez. Pauses silencieuses et pauses non silencieuses. *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA)*, 8:85–114, 1982.

[Duez, 1991] D. Duez. *La pause dans la parole de l'homme politique*. Paris, 1991.

[Dufour, 2010] R. Dufour. *Transcription automatique de la parole spontanée*. PhD thesis, Université du Maine, 2010.

[Durand *et al.*, 2002] J. Durand, B. Laks, and C. Lyche. La phonologie du français contemporain : usages, variétés et structure. In C. Pusch and W. Raible, editors, *Romanistische Korpuslinguistik - Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, pages 93–106. Gunter Narr Verlag, Tübingen, 2002.

[Durand *et al.*, 2003] J. Durand, B. Laks, and C. Lyche. Le projet "Phonologie du français contemporain (PFC)". volume 33, pages 3–9. La Tribune International des Langues Vivantes, 2003.

[Durand *et al.*, 2005] J. Durand, B. Laks, and C. Lyche. Un corpus numérisé pour la phonologie du français. In G. Williams, editor, *La linguistique de corpus*, pages 205–217. Presses Universitaires de Rennes, Rennes, 2005.

[Estève *et al.*, 2010] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas. The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In *Proceedings of the International Conference on Language Resources and Evaluation*, Malta, 2010.

[Evermann *et al.*, 2005] G. Evermann, H. Y. Chan, M. J. F. Gales, B. Jia, D. Mrva, P. C. Woodland, and K. Yu. Training LVCSR systems on thousands of hours of data. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 209–212, Philadelphia, USA, 2005.

[Ferguson, 1980] J. D. Ferguson. Hidden markov analysis: An introduction. In *Hidden Markov Models for Speech*. Institute for Defense Analyses, Princeton, NJ, 1980.

[Forgie and Forgie, 1956] J. W. Forgie and C. D. Forgie. Results obtained from a vowel recognition computer program. *Journal of the Acoustical Society of America*, 31(11):1480–1489, 1956.

[Fosler-Lussier and Morgan, 1999] E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 1999.

[Fouché, 1969] P. Fouché. *Traité de prononciation français moderne et contemporain*. Éditions Klincksieck, Paris, 1969.

[Fougeron and Steriade, 1999] C. Fougeron and D. Steriade. Au delà de la syllabe : le rôle des informations articulatoires stockées dans le lexique pour l'analyse de la chute de schwa. In *Journée d'Etudes Linguistiques*, pages 122–127, Nantes, France, 1999.

[Fry, 1959] D. B. Fry. Theoretical aspects of mechanical speech recognition. *Journal of the British Institution of Radio Engineers*, 19(4):211–218, 1959.

[Furui, 1986] S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:52–59, 1986.

[Furui, 2005] S. Furui. 50 years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology*, 1(2):64–74, 2005.

[Galibert, 2009] O. Galibert. *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris-Sud 11, 2009.

[Galliano *et al.*, 2005] S. Galliano, É. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of Interspeech*, pages 1149–1152, Lisbon, Portugal, 2005.

[Galliano *et al.*, 2006] S. Galliano, É. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 139–142, Genoa, Italy, 2006.

[Galliano *et al.*, 2009] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 Evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech*, pages 2583–2586, Brighton, UK, 2009.

[Gao *et al.*, 2006] Y. Gao, B. Zhou, R. Sarikaya, M. Afify, H. Kuo, W. Zhu, Y. Deng, C. Prosser, W. Zhang, and L. Besacier. IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. In *Proceedings of the Workshop on Medical Speech Translation at HLT-NAACL*, pages 57–60, New York, USA, 2006.

[Gauvain and Lamel, 2003] J.-L. Gauvain and L. Lamel. Large vocabulary speech recognition based on statistical methods. In W. Chou and B.-H. Juang, editors, *Pattern Recognition in Speech and Language Processing*, chapter 5, pages 149–189. CRC Press, 2003.

[Gauvain *et al.*, 2005] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. Where are we in transcribing French broadcast news? In *9th European Conference on Speech Communication and Technology*, pages 1665–1668, Lisbonne, 2005.

[Gendrot and Adda-Decker, 2005] C. Gendrot and M. Adda-Decker. Impact of duration on F_1/F_2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *9th European Conference on Speech Communication and Technology*, pages 2453–2456, Lisbonne, 2005.

[Gendrot *et al.*, 2008] C. Gendrot, M. Adda-Decker, and J. Vaissière. Les voyelles /i/ et /y/ du français : focalisation et variations formantiques. In *XXVIIèmes Journée d'Etude de la Parole*, pages 205–208, Avignon, 2008.

[Ghio, 2007] A. Ghio. L'onde sonore : réalités physiques et perception. In P. Auzou, V. Rolland, S. Pinto, and C. Ozsancak, editors, *Les Dysarthries*, pages 81–90. Solal, 2007.

[Godfrey *et al.*, 1992] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520, 1992.

[Goldwater *et al.*, 2010] S. Goldwater, D. Jurafsky, and C. D. Manning. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52:181–200, 2010.

[Good, 2004] P. I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 3rd edition, 2004.

[Graff, 2002] D. Graff. An overview of Broadcast News corpora. *Speech Communication*, 37(1–2):15–26, 2002.

[Gravier *et al.*, 2004a] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri. The ESTER evaluation campaign of rich transcription of French broadcast news. In *Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC 2004)*, pages 885–888, Lisboa, Portugal, 2004.

[Gravier *et al.*, 2004b] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Journées d'études sur la Parole*, Fez, Morocco, 2004.

[Guyon and Elisseeff, 2003] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[Hallé and Adda-Decker, 2007] P.A. Hallé and M. Adda-Decker. Voicing assimilation in journalistic speech. In *16th International Congress of Phonetic Sciences*, pages 493–496, Saarbrücken, Germany, 2007.

[Hawkins and Local, 2007] S. Hawkins and J. Local. Sound to sense: introduction to the special session. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 181–184, 2007.

[Hirose *et al.*, 2005] K. Hirose, K. Sato, Y. Asano, and N. Minematsu. Synthesis of F contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis. *Speech Communication*, pages 385–404, 2005.

[Hirschberg *et al.*, 2004] J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43:155–175, 2004.

[Hirst and Di Cristo, 1984] D. Hirst and A. Di Cristo. French intonation: a parametric approach. *Die Neueren Sprachen*, 83(5):554–569, 1984.

[Hirst and Di Cristo, 1998] D. Hirst and A. Di Cristo, editors. *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, Cambridge, 1998.

[Hogden *et al.*, 2007] J. Hogden, Ph. Rubin, E. McDermott, S. Katagiri, and L. Goldstein. Inverting mappings from smooth paths through r^n to paths through r^m: A technique applied to recovering articulation from acoustics. *Speech Communication*, 49(5):361–383, 2007.

[Huet *et al.*, 2010] S. Huet, G. Gravier, and P. Sébillot. Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. *Computer Speech and Language*, 24(4):663–684, 2010.

[Imai *et al.*, 2000] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando. Progressive 2-pass decoder for real-time broadcast news captioning. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul , Turkey, 2000.

[Itakura and Saito, 1970] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan*, 53A:36–43., 1970.

[Jelinek, 1976] F. Jelinek. Continuous speech recognition by statistical methods. *IEEE*, 64(4):532–556, 1976.

[Jelinek, 1985] F. Jelinek. The development of an experimental discrete dictation recognizer. *IEEE*, 73(11):1616–1624, 1985.

[Jelinek, 1998] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[Juang and Rabiner, 2005] B.-H. Juang and L. R. Rabiner. *Automatic speech recognition - A brief history of the technology*. Second edition, 2005.

[Juang, 1985] B. H. Juang. Maximum likelihood estimation for mixture multivariate stochastic observations of markov chains. *AT&T Technical Journal*, 64(6):1235–1249, 1985.

[Jun and Fougeron, 2000] S.-A. Jun and C. Fougeron. A phonological model of French intonation. In A. Botinis, editor, *Intonation: Analysis, Modelling and Technology*, pages 209–242. Kluwer Academic Publishers, 2000.

[Jun and Fougeron, 2002] S.-A. Jun and C. Fougeron. The realizations of the accentual phrase in French intonatoin. 14:147–172, 2002.

[Jurafsky and Martin, 2008a] D. Jurafsky and J. H. Martin. Automatic speech recognition. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 9. Prentice Hall, second edition, 2008.

[Jurafsky and Martin, 2008b] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, second edition, 2008.

[Kawahara *et al.*, 2003] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pages 135–138, Tokyo, Japon, 2003.

[Kessens, 2002] J. M. Kessens. *Making a difference on automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition*. PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen, the Netherlands, 2002.

[Kim and Cho, 2009] S. Kim and T. Cho. The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in korean. *Journal of the Acoustical Society of America*, 125(5):3373–3386, 2009.

[Kim *et al.*, 2010] S. Kim, M. Broersma, and T. Cho. Native and non-native prosodic cues in segmentation and learning. In *33rd Generative Linguistics in the Old World Colloquium*, Wroclaw, Poland, 2010.

[Klatt, 1977] D. H. Klatt. Review of the arpa speech understanding project. *Journal of the Acoustical Society of America*, 62(6):1345–1366, 1977.

[Kolář *et al.*, 2010] J. Kolář, Y. Liu, and E. Shriberg. Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech. *Speech Communication*, 52:236–245, 2010. Available Online.

[Kratochvil, 1998] P. Kratochvil. Intonation in Beijing Chinese. In D. Hirst and A. Di Cristo, editors, *Intonation Systems: A Survey of Twenty Languages*, pages 417–431. Cambridge University Press, Cambridge, 1998.

[Labov, 1972] W. Labov. *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia, 1972.

[Lacheret-Dujour and Beaugendre, 1999] A. Lacheret-Dujour and F. Beaugendre. *La prosodie du français*. CNRS, Paris, 1999.

[Lacheret-Dujour and Victorri, 2002] A. Lacheret-Dujour and B. Victorri. La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum*, 24(1–2):55–72, 2002.

[Lacheret-Dujour, 2000] A. Lacheret-Dujour. La prosodie – Niveaux d'analyse et probl`emes de représentation. In J. L. Schwartz and P. Escudié, editors, *La parole : des modèles cognitifs aux machines communicantes*, pages 245–282. Hermes, Paris, 2000.

[Ladd, 1996] R. D. Ladd. *Intonational phonology*. Cambridge University Press, 1996.

[Lamel and Gauvain, 2003] L. Lamel and J.-L. Gauvain. Speech recognition. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter 16, pages 305–322. Oxford University Press, New York, 2003.

[Landwehr *et al.*, 2005] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.

[Lehiste, 1970] I. Lehiste. *Suprasegmentals*. MIT Press, Cambridge, MA, USA, 1970.

[Lehiste, 1977] I. Lehiste. Isochrony reconsidered. *Journal of Phonetics*, 5:253–263, 1977.

[Léon, 2007] P. R. Léon. *Phonétisme et prononciation du français*. Armand Colin, Paris, 5e edition, 2007.

[Leung *et al.*, 2008] C.-C. Leung, M. Ferras, C. Barras, and J.-L. Gauvain. Comparing prosodic models for speaker recognition. In *Proceedings of Interspeech*, pages 1945–1948, 2008.

[Lindblom, 1963] B. Lindblom. Spectrogaphic study of vowel reduction. *Journal of the Acoustical Society of America*, 35:1773–1781, 1963.

[Lindblom, 1990] B. Lindblom. Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modeling*, pages 403–439. Kluwer Academic Publishers, Dordrecht, 1990.

[Lippmann, 1987] R. P. Lippmann. An introduction to computing with neural nets. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 4(2):4–22, 1987.

[Lippmann, 1997] R. P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[Liscombe, 2007] J. Liscombe. *Prosody and speaker state: Paralinguistics, pragmatics, and proficiency*. PhD thesis, Columbia University, 2007.

[Luce and Pisoni, 1998] P. A. Luce and D. B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36, 1998.

[Maekawa, 2003] K. Maekawa. Corpus of spontaneous Japanese: Its design and evaluation. In *IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[Marslen-Wilson and Zwitserlood, 1989] W.D. Marslen-Wilson and P. Zwitserlood. Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15:576–585, 1989.

[Martin, 1975] Ph. Martin. Analyse phonologique de la phrase française. *Linguistics*, 146:35–68, 1975.

[Martin, 2010] Ph. Martin. Prosodic structure revisited: a cognitive approach. The example of French. In *Speech Prosody*, Chicago, IL, USA, 2010.

[Mattys *et al.*, 1999] S. L. Mattys, P. W. Jusczyk, P. A. Luce, and J. L. Morgan. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38(4):465–494, 1999.

[Mertens, 1987] P. Mertens. *L'intonation du français. De la description linguistique à la reconnaissance autimatique*. PhD thesis, Université de Leuven, 1987.

[Mertens, 2002] P. Mertens. Synthesizing elaborate intonation contours in text-to-speech for French. In *Speech Prosody*, pages 499–502, Aix-en-Provence, France, 2002.

[Mertens, 2004] P. Mertens. Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des langues*, 45(2):109–130, 2004.

[Mertens, 2006] P. Mertens. A predictive approach to the analysis of intonation in discourse in French. In Y. Kawaguchi, I. Fonágy, and T. Moriguchi, editors, *Prosody and Syntax: Cross-linguistic Perspectives*, volume 3 of *Usage-Based Linguistic Informatics*, pages 65–101. John Benjamins Publishing Company, 2006.

[Mertens, 2009] P. Mertens. Prosodie, syntaxe, discours : autour d'une approche prédictive. In H.-Y. Yoo and E. Delais-Roussarie, editors, *Interface Discours & Prosodie (IDP)*, pages 19–32, Paris, 2009.

[Metze, 2007] F. Metze. Discriminative speaker adaptation using articulatory features. *Speech Communication*, 49(5):348–360, 2007.

[Moore, 2007] Roger K. Moore. Spoken language processing: Piecing together the puzzle. *Speech Communication*, 49(5):418–435, 2007.

[Nespor and Vogel, 1986] M. Nespor and I. Vogel. *Prosodic Phonology*. Foris, Dordrecht, 1986.

[Nusbaum and Pisoni, 1987] H. Nusbaum and D. Pisoni. Automatic measurement of speech recognition performance: a comparison of six speaker-dependent recognition devices. 2:87–108, 1987.

[Olson and Belar, 1956] H. F. Olson and H. Belar. Phonetic typewriter. *Journal of the Acoustical Society of America*, 28(6):1072–1081, 1956.

[Oppenheim *et al.*, 1968] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham. Nonlinear filtering of multiplied and convolved signals. *IEEE*, 56(8):1264–1291, 1968.

[Ostendorf *et al.*, 2003] M. Ostendorf, I. Shafran, and R. Bates. Prosody models for conversational speech recognition. In *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pages 147–154, 2003.

[Pallett *et al.*, a] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, A. Martin, and M. A. Przybocki.

[Pallett *et al.*, b] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki.

[Pellegrino, 2009] F. Pellegrino. *De l'identification des langues à la complexité phonologique*. Habilitation à diriger des recherches, Université Lumière Lyon 2, Université de Lyon, 2009.

[Pierrehumbert, 1980] J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, 1980.

[Pike, 1945] K. L. Pike. *Intonation of American English*. Ann Arbor Press: University of Michigan, 1945.

[Pols, 1999] L. C. W. Pols. Flexible, robust, and efficient human speech processing versus present-day speech technology. In *XIVth International Congress of Phonetic Sciences ICPhS'99*, volume 1, pages 9–16, San Fransisco, CA, 1999.

[Post, 2000] B. Post. *Tonal and phrasal structures in French intonation*. The Hague: Holland Academic Graphics, 2000.

[Price *et al.*, 1988] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett. The darpa 1000-word resource management database for continuous speech recognition. In *Proceedings of ICASSP*, pages 651–654, Philadelphia, Pennsylvania, 1988.

[Prieto *et al.*, 2010] P. Prieto, E. Estebas-Vilaplana, and M. M. Vanrell. The relevance of prosodic structure in tonal articulation: Edge effects at the prosodic word level in Catalan and Spanish. *Journal of Phonetics*, 38(4):688–707, 2010.

[R Development Core Team, 2008] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0, www.R-project.org.

[Rabiner *et al.*, 1979] R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon. Speaker independent recognition of isolated words using clustering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27:336–349, 1979.

[Rabiner, 1989] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recogition. *IEEE*, 77(2):257–286, 1989.

[Ramus, 2002] F. Ramus. Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, 2:85–115, 2002.

[Reddy, 1966] D. Reddy. An approach to computer speech recognition by direct analysis of the speech wave. Technical report, C549, Computer Science Department, Stanford University, 1966.

[Rietveld, 1980] A. C. M. Rietveld. Word boundaries in the french language. *Language and Speech*, 23:289–296, 1980.

[Riley *et al.*, 1999] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraçlar, C. Wooters, and G. Zavaliagkos. 1999.

[Rossi, 1981] M. Rossi. Vers une théorie de l'intonation. Klincksieck, Paris, 1981.

[Rouas, 2007] J.-L. Rouas. Automatic prosodic variations modelling for language and dialect discrimination. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1904–1911, 2007.

[Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[Saraçlar and Khudanpur, 2004] M. Saraçlar and S. Khudanpur. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech and Language*, 18(4):375–395, 2004.

[Scharenborg, 2007] O. Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49:336–347, 2007.

[Schmid, 1994] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994.

[Selkirk, 1986] E. Selkirk. On derived domains in sentence phonology. In *Phonology Yearbook*, volume 3, pages 371–405. 1986.

[Selkirk, 1996] E. Selkirk. The prosodic structure of function words. pages 187–214, 1996.

[Sethy *et al.*, 2002] A. Sethy, S. Narayanan, and S Parthasarthy. A syllable based approach for improved recognition of spoken names. In *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pages 30–35, Aspen Lodge, USA, 2002.

[Shen *et al.*, 2008] W. Shen, J. Olive, and D. Jones. Two protocols comparing human & machine phonetic discrimination performance in conversational speech. In *Proceedings of Interspeech*, Brisbane, Australia, 2008.

[Shinozaki and Furui, 2001] T. Shinozaki and S. Furui. Error analysis using decision trees in spontaneous presentation speech recognition. In *Proceedings of IEEE ASRU*, 2001.

[Shinozaki and Furui, 2003] T. Shinozaki and S. Furui. An assessment of automatic recognition techniques for spontaneous speech in comparison with human performance. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[Shriberg, 2001] E. Shriberg. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journl Of the international phonetic association*, 31(1):153–169, 2001.

[Siegler and Stern, 1995] M. Siegler and R. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995.

[Silverman *et al.*, 1992] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling English prosody. In *ICSLP*, 1992.

[Sloboda and Waibel, 1996] T. Sloboda and A. Waibel. Dictionary learning for spontaneous speech recognition. In *4th international conference on spoken language processing*, Philadelphia, U.S.A., 1996.

[Spinelli *et al.*, 2003] E. Spinelli, J. McQueen, and A. Cutler. Processing resyllabified words in French. *Journal of Memory and Language*, 48:233–254, 2003.

[Spinelli *et al.*, 2007] E. Spinelli, P. Welby, and A.-L. Schaegis. Fine-grained access to targets and competitors in phonemically identical spoken sequences: the case of French elision. *Language and cognitive processes*, 22(6):828–859, 2007.

[Spinelli *et al.*, 2010] E. Spinelli, N. Grimaud, F. Meunier, and P. Welby. An intonational cue to segmentation in phonemically identical sequences. *Attention, Perception & Psychophysics*, 72(3):775–787, 2010.

[Stolcke *et al.*, 1998] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 5, pages 2247–2250, Sydney, Australia, 1998.

[Stolcke *et al.*, 2006] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 14, pages 1729–1744, 2006.

[Strik and Cucchiarini, 1998] H. Strik and C. Cucchiarini. Modeling pronunciation variation for ASR: overview and comparison of methods. In H. Strik, J.M. Kessens, and M. Wester, editors, *Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 137–144, Rolduc, The Netherlands, 1998.

[Strik and Cucchiarini, 1999] H. Strik and C. Cucchiarini. Modeling pronunciation variation for ASR: A servey of the literature. *Speech Communication*, 29:225–246, 1999.

[Traunmüller, 2005] H. Traunmüller. Auditory scales of frequency representation. http://www2.ling.su.se/staff/hartmut/bark.htm, 2005.

[Vaissière and Le Corre, 1976] J. Vaissière and C. Le Corre. Premiers essais de segmentation automatique de la parole continue en mots à partir des variations du fondamental dans la phrase française. In *8èmes Journées d'Etudes sur la Parole*, pages 345–352, 1976.

[Vaissière and Michaud, 2006] J. Vaissière and A Michaud. Prosodic constituents in French: a data-driven approach. In Y. Kawaguchi, I. Fonágy, and T. Moriguchi, editors, *Prosody and Syntax: Cross-linguistic Perspectives*, volume 3 of *Usage-Based Linguistic Informatics*, pages 47–64. John Benjamins Publishing Company, 2006.

[Vaissière, 1971] J. Vaissière. *Contribution à la synthèse par règles du français*. PhD thesis, Université de Grenoble, 1971.

[Vaissière, 1976] J. Vaissière. Utilisation, pour la reconnaissance de la parole continue, de marqueurs prosodiques extraits de la fréquence du fondamental. In *8èmes Journées d'Etudes sur la Parole*, 1976.

[Vaissière, 1980] J. Vaissière. La structuration acoustique de la phrase française. pages 530–560, 1980.

[Vaissière, 1991] J. Vaissière. Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord, and R. Carlson, editors, *Music, Language, Speech and Brain*, pages 108–121, 1991.

[Vaissière, 1997] J. Vaissière. Langues, prosodie et syntaxe. *Revue Traitement Automatique des Langues, ATALA*, 38(1):53–82, 1997.

[Vaissière, sous press] J. Vaissière. Les universaux de substance prosodiques. *Les universaux sonores*, sous press.

[van Dooren and van den Eynde, 1982] K. van Dooren and K. van den Eynde. A structure for the intonation of dutch. *Linguistics*, 20:203–235, 1982.

[Varga *et al.*, 2002] I. Varga, S. Aalburg, B. Andrassy, S. Astrov, J.G. Bauer, C. Beaugeant, C. Geissler, and H. Hoge. ASR in mobile phones–an industrial approach. *IEEE Transactions on Speech and Audio Processing*, 10(8):562–569, 2002.

[Vasilescu and Adda-Decker, 2006] I. Vasilescu and M. Adda-Decker. Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech. In *Interspeech*, pages 1850–1853, Pittsburgh, Pennsylvania, USA, 2006.

[Vasilescu *et al.*, 2009] I. Vasilescu, M. Adda-Decker, L. Lamel, and P. Hallé. A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English. In *Interspeech*, pages 144–147, Brighton, UK, 2009.

[Vasilescu *et al.*, 2011] I. Vasilescu, D. Yahia, N. Snoeren, M. Adda-Decker, and L. Lamel. Cross-lingual study of ASR errors: on the role of the context in human perception of near-homophones. In *Interspeech*, Florence, Italy, 2011.

[Vicsi and Szaszák, 2010] K. Vicsi and G. Szaszák. Using prosody to improve automatic speech recognition. *Speech Communication*, 52:413–426, 2010.

[Vidrascu and Devillers, 2007] L. Vidrascu and L. Devillers. Five emotion classes detection in real-world call center data : the use of various types of paralinguistic features. In *ParaLing 2007*, Saarbrücken, Germany, 2007.

[Vieru-Dimulescu *et al.*, 2007] B. Vieru-Dimulescu, Ph. Boula de Mareüil, and M. Adda-Decker. Identification of foreign-accented french using data mining techniques. In *ParaLing 2007*, Saarbrücken, Germany, 2007.

[Vintsyuk, 1968] T. K. Vintsyuk. Speech discrimination by dynamic programming. *Kibernetika*, 4:81–88, 1968.

[Vitevitch and Luce, 1998] M. S. Vitevitch and P. A. Luce. When words compete: Levels of processing in spoken word perception. *Psychological Science*, 9:325–329, 1998.

[Waibel *et al.*, 1989] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.

[Waibel *et al.*, 2003] A. Waibel, A. Badran, A. W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. M. Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang. Speechalator: Two-way speech-to-speech translation in your hand. In *Proceedings of HLT-NAACL*, pages 29–30, 2003.

[Warner and Arai, 2001] N. Warner and T. Arai. Japanese mora-timing: A review. *Phonetica*, 58:1–25, 2001.

[Warner *et al.*, 2010] N. Warner, T. Otake, and T. Arai. Intonational structure as a word boundary cue in japanese. *Language and Speech*, 53:107–131, 2010.

[Watanabe *et al.*, 2005] M. Watanabe, Y. Den, K. Hirose, and N. Minematsu. The effects of filled pauses on native and non-native listeners speech processing. In *DiSS*, pages 169–172, 2005.

[Welby, 2006] P. Welby. French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, 34:343–371, 2006.

[Welby, 2007] P. Welby. The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, 49:28–48, 2007.

[Wester and Fosler-Lussier, 2000] M. Wester and E. Fosler-Lussier. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000.

[Wester, 2002] M. Wester. *Pronunciation variation modeling for Dutch automatic speech recognition*. PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen, the Netherlands, 2002.

[Wiren and Stubbs, 1956] J. Wiren and H. Stubbs. Electronic binary selection system for phoneme classification. *Journal of the Acoustical Society of America*, 28:1082–1091, 1956.

[Witten and Frank, 2005] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.

[Woehrling, 2009] C. Woehrling. *Accents régionaux en français : Perception, analyse et modélisation à partir de grands corpus*. PhD thesis, Université Paris-Sud 11, 2009.

[Young, 1996] S. Young. Large vocabulary continuous speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–57, 1996.

[Zellner, 1998] B. Zellner. *Caractérisation et prédiction du débit de parole en français. Une étude de cas*. PhD thesis, Université de Lausanne, 1998.